REMIGIJUS LAPINSKAS

Lecture Notes

PRACTICAL ECONOMETRICS. I. REGRESSION MODELS

PRAKTINĖ EKONOMETRIJA. I. REGRESINIAI MODELIAI

Paskaitų konspektas

remigijus.lapinskas@mif.vu.lt

Vilnius 2013.12

Contents

1. INTRODUCTION

Econometrics is the application of mathematics and statistical methods to economic data and can be described as the branch of economics that aims to give empirical content to economic relations. Econometrics allows economists to sift through mountains of data to extract simple relationships.

There are two types of economic data: (a) macroeconomic data, representing quantities and variables related to a national economy as a whole, usually based on national census; and (b) microeconomic data, representing information about the economic behavior of individual persons, households, and firms. Macroeconomic data are usually given as a set of time series (we shall model these data in the PE.II course), while microeconomic data are obtained mainly through statistical surveys and are given as cross-sectional data. These two types of data, related to macroeconomic theory and microeconomic theory, respectively, require different approaches; and sometimes information obtained from both types of data has to be combined; obtaining macroeconomic information from microeconomic data is called aggregation.

1.1. Regression models

Economics is a system of interconnected components, that is, $Y = f(X_1, X_2, ...)$ where the list of X's may contain many variables. For example, let Y be a total production (the monetary value of all goods produced in a given country during one year); then Y depends on X_1 (= L = labor input (the total number of person-hours worked in a year)), X_2 (= K = capital input (the monetary worth of all machinery, equipment, and buildings)), and $X_3, X_4, X_5, ...$ – other quantities such as productivity, land or raw materials, inventory, economic and political situation in the country or internationally (described by many variables) etc. Some of the variables are observed or measured, others (such as political system in the country or unpredictable personal solutions or person‘s ability etc) are difficult or impossible to quantify. Thus $Y = f(X_1, X_2, U_3, U_4, U_5, ...)$ where U's stand for unobservable variables. To filter out the effects of unobservables U's, we assume that, once X's are known, the effect of U's is not very big, it is in some sense the same for any collection of X's. More specifically, we assume that $Y = f(X_1, X_2, \varepsilon)$ where ε is a random variable such that its average does not depend on X's: $E(\varepsilon \mid X_1, X_2) = const$. In the production function case, the Cobb-Douglas law says that $Y = \beta_0 L^{\beta_1} K^{\beta_2} \varepsilon$. The purpose of econometrics is to use the *data*

Y	X_1	X_2
Y_1	X_{11}	X_{21}
Y_2	X_{12}	X_{22}
....	....	....
Y_N	X_{1N}	X_{2N}

obtained through observing N countries or companies in order to approximately restore or *estimate* the unknown parameters β_0, β_1 and β_2.

The best developed is the additive case where $Y = f(X_1,..., X_k) + \varepsilon$ (our production function example may easily be transformed to this shape after taking logarithms: $\log Y = \log \beta_0 + \beta_1 \log L + \beta_2 \log K + \log \varepsilon$); more generally, in what follows, we shall usually analyze the linear[1] case $Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \varepsilon$ where we want to find „good" formulas to estimate β's, choose the right functional form (maybe not X_m but $\log(X_m)$?), to test whether our model matches the economic theory and so on.

The function $f(X_1,..., X_k; \beta_0, \beta_1,..., \beta_k)$ is called the *regression* function, the 'independent' variable X is usually called the regressor or predictor or *explanatory* variable (there may be one or more of these), the 'dependent' variable Y is the *response* variable. The random component ε (called *error* or disturbance or (economic) shock) is usually assumed (to simplify the analysis) normally distributed. Our aim is to reveal the regression function by removing the error – or as much of it as we can.

The class from which the regression functions are selected (or the *model*) is usually one of the following types:

1. a linear function of β_0 and β_1 (for example, $Y = \beta_0 + \beta_1 X + \varepsilon$; it is a *simple* (univariate linear) regression or $\log Y = \beta_0 + \beta_1 X$, this is called a *log-linear* regression),
2. a polynomial function of X (that is $Y = \beta_0 + \beta_1 X + ... + \beta_p X^p + \varepsilon$) (called a *polynomial* (linear) regression),
3. a linear function of β's (for example, $Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \varepsilon$, this is a multivariate linear regression, an example of multiple regression),
4. any other type of function, with one or more parameters (for example, $Y = \beta_0 \exp(\beta_1 X) + \varepsilon$, a nonlinear regression[2], or $P(Y = 1) = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2)$ where Y takes only two values, 0 or 1, and Φ is the standard normal distribution function (the model is called the probit regression) etc.

1.2. Statistical Data and their Models

This section introduces common types of economic data and describes some basic models associated with their use.

- **Cross-sectional data**

Some researchers often work with data that is characterized by individual *units*. These units might refer to companies, people or countries. For instance, a researcher investigating theories relating to portfolio allocation might collect data on the return earned on the stocks of many different companies at more or less the same time. With such *cross-sectional* data, the method of ordering the data (for example, by alphabet or size) usually does not matter.

[1] "Linear" means that β's enter the equation as multipliers of X.

[2] It is nonlinear because it is not of the form …+ $\beta_1 \cdot X$.

Typically, for the cross-sectional data the notations X_i, Y_i, and like are used to indicate an observation on variables X, Y etc for the ith individual. Observations in a cross-sectional data set run from unit $i = 1$ to N. By convention, N indicates the number of cross-sectional units (e.g., the number of companies surveyed). For instance, a researcher might collect data on the share price of $N = 100$ companies at a certain point in time. In this case, Y_1 will be equal to the share price of the first company, Y_2 the share price of the second company, and so on.

- **Time series data**

Financial researchers are often interested in phenomena such as stock prices, interest rates, exchange rates, etc. This data is collected at specific points in time. In all of these examples, the data are ordered by time and are referred to as *time series* data. The underlying phenomenon which we are measuring (e.g., stock prices, interest rates, etc.) is referred to as a variable. Time series data can be observed at many *frequencies*. Commonly used frequencies are: *annual* (i.e. a variable is observed every year), *quarterly* (i.e. four times a year), *monthly*, *weekly* or *daily*.

In this course, we will use the notation Y_t to indicate an observation on variable Y (e.g., an exchange rate) at time t. A series of data runs from period $t = 1$ to $t = T$. 'T' is used to indicate the total number of time periods covered in a data set. To give an example, if we were to use monthly time series data from January 1947 through October 1996 on the UK pound/US dollar exchange – a period of 598 months – then $t = 1$ would indicate January 1947, $t = 598$ would indicate October 1996 and $T = 598$ the total number of months. Hence, Y_1 would be the pound/dollar exchange rate in January 1947, Y_2 this exchange rate in February 1947, etc. Time series data are typically presented in chronological order.

One objective of analysing economic data is to predict or forecast the future values of economic variables. One approach to do this is to build a more or less structural (for example, regression) econometric model, describing the relationship between the variable of interest with other economic quantities, to estimate this model using a sample of data, and to use it as the basis for forecasting and inference. Although this approach has the advantage of giving economic content to one's predictions, it is not always very useful. For example, it may be possible to adequately model the contemporaneous relationship between unemployment and the inflation rate, but as long as we cannot predict future inflation rates we are also unable to forecast future unemployment. ◀◀

The most interesting results in econometrics during the last 20-30 years were obtained in the intersection of cross-sectional and time series methods. These lecture notes are basically devoted to cross-sectional data, however some sections will examine

- **regression models for time series**

Another possibility to combine the two above-mentioned methods is to deal with the so-called

- **panel data**

A data set containing observations on multiple phenomena observed over multiple time periods is called panel data. Panel data aggregates all the individuals, and analyzes them in a period of time. Whereas time series and cross-sectional data are both one-dimensional, panel data sets are two-dimensional.

person	year	income	age	sex
1	2003	1500	27	1
1	2004	1700	28	1
1	2005	2000	29	1
2	2003	2100	41	2
2	2004	2100	42	2
2	2005	2200	43	2

In the above example, a data set with panel structure is shown. Individual characteristics (income, age, sex) are collected for different persons and different years. Two persons (1 and 2) are observed over three years (2003, 2004, and 2005). Because each person is observed every year, the data set is called a panel.

1.3. Software

There are many statistical software programs. Broadly speaking, they can be divided into commercial (SAS, SPSS, EViews,...) and free (R, GRETL,...) software; on the other hand, according to the way the procedures are performed, they can be divided into menu-driven (GRETL, EViews,...) and programmable (R). The latter two groups nowadays have shown a tendency to unite – for example, EViews, GRETL and the commercial S-Plus, all allow to program your steps or perform them from the toolbar. This course is accompanied by computer labs where statistical procedures will be parallelly performed with GRETL (it is very good for teaching purposes) and R (the most powerfull and cutting-edge statistical program).

****

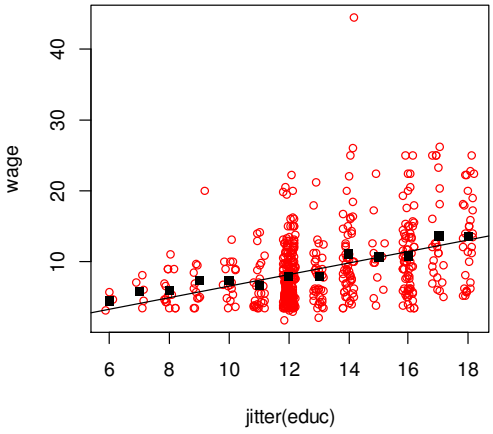
All the data necessary for these computer labs are placed in the ...\dataPE folder.

2. INTRODUCTORY EXAMPLES

When we consider the nature and form of a relationship between any two or more variables, the analysis is referred to as *regression analysis*.

2.1. Some Examples of Regression Models

2.1 example. Imagine that you are a university student, completing your sixteenth (=12+4) year of education and considering whether to go on to graduate studies for two-year master's program or to proceed directly to the job market. Many factors will weigh in your decision, one of which may be the financial payoff. You may well be interested in how earnings of persons with master's degrees compare with those of persons with bachelor's degrees.



In the figure on the right you can see cross-section wage data from CPS5_n.txt[1] consisting of a random sample taken from the national survey of persons in the labor force who have already completed their education. From the underlying data, it turns out that for the 70 persons with `educ`=16 (years), the mean `wage` was 10.84 ($/hour), while for the 24 persons with `educ`=17, the mean `wage` was 13.61 (this is a substantial difference).

At this point, your interest in the relation between wages and education may have been piqued enough to move from a personal decision-making context to a general scientific one[2]. It seems natural to focus on the means of those Y's (`wages`) for each of the thirteen distinct values of X (`educ`), much as we did previously for the two values $X=16$ and $X=17$. In our figure, those thirteen *sample conditional means* have been superimposed as black squares on the scatter diagram of 528 persons. Many people would take this display as a natural summary of the relation between Y and X in this data set, or sample, with $N=528$ paired observations (X_i,Y_i). At each distinct value of X, x_j, we can calculate the mean value of Y. At education level x_j, $j=6,7,...18$, this subsample mean of Y may be labeled as $\bar{Y}\,|\,X=x_j$ to be read as the sample mean of Y conditional on (or given that) $X=x_j$. Making this calculation for all j we can assemble the results into the conditional sample mean function $h(X)=\bar{Y}\,|\,X$. Here $h(X)$ denotes the function whose value at x_j, namely $h(x_j)$, is equal to $\bar{Y}\,|\,X=x_j$.

The variable `educ` has 13 distinct values and this discrete, sliced sample approximation to the true, however unknown, conditional expectation $E(Y\,|\,X)$ could be a reasonable approach in this example. The most interesting hypothesis in this context is $H_0: E(Y\,|\,X)\equiv const$ which

[1] All the data sets of this course are placed in the .../PEdata folder.
[2] Recall – modeling is the daily bread of econometricians!

is equivalent to $H_0 : E(Y \mid X = x_6) = ... = E(Y \mid X = x_{18})$ which is equivalent to $H_0 : Y$ *does not depend on* X. We can use the ANOVA procedure to compare these (conditional) means or some direct regressional methods to be discussed later – both reject our hypotheses at 5% significance level, consequently, `wage` indeed depends on `educ`.

We had to estimate eighteen parameters when using the above approach. However, **if** it happens that 1) the discrete variable X can be treated as a numeric (but not as a nominal or group variable `educ`[3]) and 2) $E(Y \mid X) = \beta_0 + \beta_1 X$, this new parametrization[4] of $E(Y \mid X)$ would be preferable because it is easier to estimate two coefficients, β_0 and β_1, rather than eighteen (the estimation will be more precise and, also, we can interpret β_1 in a reasonable way as a measure of influence of X on Y). To find the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ of the unknown coefficients, one can draw a straight line through the "middle of the cloud" (or, in other words, a straight line which is the most close to all the points in the scatter diagram). The "closest" line (or, in other words, the coefficients β_0 and β_1) can be estimated differently, for example, as such that $\sum_{i=1}^{N} |Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)|$ is minimum or $\sum_{i=1}^{N} (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2$ is minimum or in a similar manner (the second method, called the *method of ordinary least squares* or OLS, is the most popular). In the next chapter, we shall explain how to achieve this goal (note that in order to test $H_0 : Y$ *does not depend on* X now we have to test a simple hypothesis $H_0 : \beta_1 = 0$).

2.2 example. Open R and run the following script where the below used data set `oats` contains four variables, among them:

`N` nitrogen fertilizer, levels 0.0, 0.2, 0.4, and 0.6 cwt/acre (`N` is a factor, i.e., nominal or group variable!)
`NN` it is discussible whether the nominal variable `N` can be converted to a numeric variable `NN` taking values 1, 2, 3, 4!
`Y` yields in 1/4lbs per sub-plot, each of area 1/80 acre.

```
library(MASS)
data(oats)
attach(oats)
par(mfrow=c(1,2))
plot(N,Y)
NN=as.numeric(N) # convert N to a numeric variable NN
cor(Y,NN) # =0.6130266 - relationship is strong
plot(jitter(NN),Y)
```

To analize the impact of fertilizer on the yield, 72 adjacent lots were used by a researcher and dressed with four different levels of fertilizer. The fertilizer `N` takes four different values but these are not numbers:

```
> N
[1] 0.0cwt 0.2cwt 0.4cwt 0.6cwt 0.0cwt 0.2cwt 0.4cwt ...............
Levels: 0.0cwt 0.2cwt 0.4cwt 0.6cwt
```

[3] We can aggregate our data and introduce a new nominal variable `educ.n` taking four values: primary, secondary, bachelor, and master; now the assumption $E(wage \mid educ.n) = \beta_0 + \beta_1\, educ.n$ makes no sense.

[4] It means that all conditional expectations are exactly on a straight line.

N is a nominal variable called a factor in R and a discrete variable in GRETL and this means that we cannot claim that the increment of N between the levels 0.2cwt and 0.0cwt is the same as between 0.6cwt and 0.4cwt (the arithmetical operations are not defined between factors). Thus, we should estimate the conditional expectation $E(Y \mid N)$ separately for each level of N (for each slice of our scatter diagram) and this is done by

```
mod1=lm(Y~N)
summary(mod1)
points(NN,predict(mod1),pch=15,col=2,cex=2) # red squares in Fig.2.1
```

On the other hand, if we dare treat N as a numeric variable and call it NN, we can add a regression line with

```
mod2=lm(Y~NN)
abline(mod2)
summary(mod2)
Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)    67.139       6.215   10.802   < 2e-16 ***
NN             14.733       2.270    6.492  1.04e-08 ***
```

(thus the estimate of the regression line is $Y = 67.1 + 14.7\,NN$). The scatter diagram shows that the yield increases together with the fertilizer and it seems that the linear model $Y = \beta_0 + \beta_1 \mathrm{NN} + \varepsilon$ satisfactorily describe the relationship.

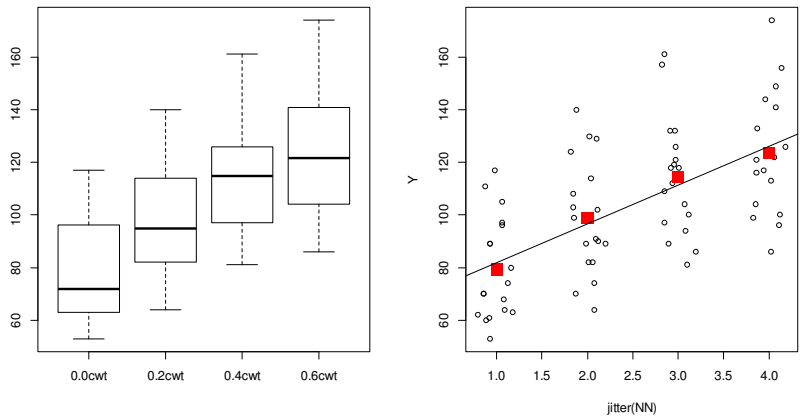


Figure 2.1. Two graphs obtained with `plot(N,Y)` where N is treated as a factor (left) or as a numeric variable (right; the red squares are conditional means of Y in groups)

The principal difference between 2.1 and 2.2 examples is that in the latter case the explanatory variables N and NN are not random, they are fixed in advance by the researcher. Note that in economics, as a rule, X or X's are random which brings some additional complications.

2.1 exercise. Analyze the impact of the variety V on Y. ◀◀

2.3 example. Open GRETL and go to File| Open data| Sample file...| POE 4th ed.| andy. The data set contains sales, price and advert (expenditure on advertising). We want to find the forecast of sales based on price and advert. In this three dimensional case, the ge-

ometric interpretation is of little help (how to plot points? how to draw the „best“ plane? how to retrieve the coefficients from the graph?) therefore the general, mathematical approach is

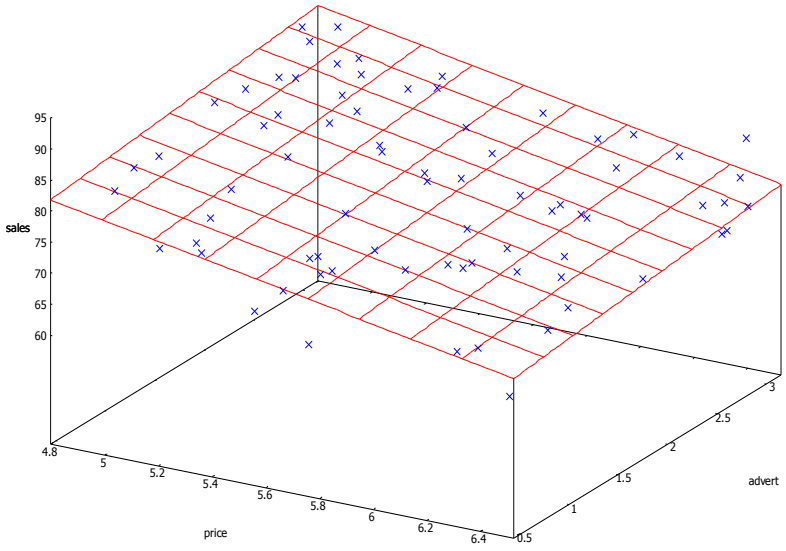


Figure 2.2. A 3D scatter diagram (`sales` vs `price` and `advert`)

the only recource. In GRETL, the OLS model is obtained through

```
ols sales 0 price advert

Dependent variable: sales

            coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       118.914       6.35164     18.72     2.21e-029 ***
  price        -7.90785     1.09599     -7.215    4.42e-010 ***
  advert        1.86258     0.683195     2.726    0.0080    ***
```

thus, the regression model is $sales = 118.91 - 7.91\,price + 1.86\,advert$. Note that in four- and higher-dimensional cases the geometric approach is altogether inapplicable. ◀

2.2. Concluding Remarks

Any economic variable Y, generally, depends on (random) observable X (or X's) and unobservable variables which we treat as random variable ε, $Y = f(X, \varepsilon)$. Regression modeling aims to filter out the effects of ε and estimate the “mean” dependence of Y on X which is denoted as $E(Y \mid X)$ (called the average of Y provided we know X or expected value of Y conditional on X). If X takes only finite number of values x_i, it is easy to define and estimate $E(Y \mid X = x_i)$ – just take a sample mean in a given slice containing one element (see 2.1 example; note that the dependence of $E(Y \mid X = x_i)$ on x_i can be quite irregular).

On the other hand, if X may take all the values from a certain interval, then the slices in sample will contain either none of X's or one element (sometimes, two or more) of X. To illustrate, open GRETL and go to File| Open Data| Sample file…| Ramanathan| data2-2, select `colgpa` (grade point average in college) and `hsgpa` (high school gpa), right-click on selection and choose XY scatterplot – you will see the scatterplot and the OLS regression line (see Fig. 2.2, left). This regression line is clearly one of the options to model `colgpa` dependence on `hsgpa`. However, if we want to follow our slicing technique, it is not quite clear how to estimate $E(Y \mid X)$ (= $E($`colgpa`$\mid$`hsgpa`$)$) for each slice (it will be either NA for some slices or an average `colgpa` in other slices). An alternative approach to estimate $E(Y \mid X)$, assuming $Y = g(X) + \varepsilon$ for some smooth g, is to take some wider slice around moving x and average Y's inside it or, still better, to take a weighted average, i.e., instead of $\frac{1}{\#\, j} \sum_{j:|x_j - x|<h} Y_j$ to calculate $\sum_{j:|x_j - x|<h} w_j Y_j$ where weights w_j are "big" for x_j close to x and smaller for remote x_j's. This idea is implemented in both GRETL and R with the `loess` (locally weighted regression) function (in GRETL, go to Model| Robust estimation| Loess… – you will see the graph depicted in Figure 2.2, right).

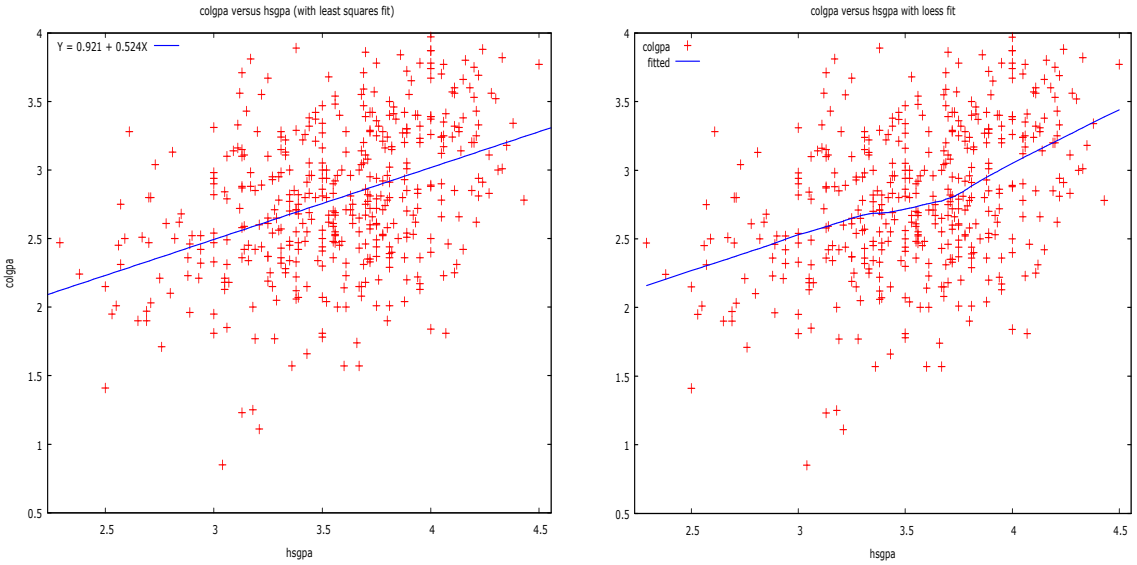


Figure 2.3. The OLS regression (left) and the *loess* regression of `colgpa` vs `hsgpa`; the dependence is close to linear

As in the case with finite number of values of X, it is difficult to interpret the right graph – we see that Y depends on X, but it is not quite clear how to quantify it. Below, we shall explain how to estimate $E(Y \mid X) = \beta_0 + \beta_1 X$ using OLS; the coefficient $\hat{\beta}_1$ then gives the change of Y when X increases by 1. However, in cases where $E(Y \mid X)$ is far from regular curve (i.e., straight line, parabola, logarithm and like), `loess` is a valuable alternative to the most popular model of the least squares studied in following chapters.

3. UNIVARIATE REGRESSION

3.1. Introduction

1. Let us assume that we observe an r.v. X with unknown distribution F or, what is the same in the continuous r.v. case, unknown density function f (this is sometimes called a *population*). To get a feeling about its shape, i.e., the „chances" or probabilities of different values, we analyse a finite *random sample* $(X_1, X_2, ..., X_N)$, where $X_i, i = 1, ..., N,$ are i.i.d.r.v.'s, all with the same distribution F. In fact, we usually collect only one realisation of this random sample, namely, a set of numbers $(x_1, x_2, ..., x_N)$, called a *concrete sample*. There are many ways to test hypothesis about the shape of F, but often it suffices to estimate its mean (plus, maybe, variance) of X. One of the methods to estimate the mean $\mu = EX$ is the method of *least squares* (what other methods do you know?): we look for a number $\hat{\mu}$ which in some sense (in squared differences sense) is central or closest to all the points $x_1, ..., x_n$ –

$$\sum_{i=1}^{N} (x_i - \hat{\mu})^2 = \min_{m \in R} \sum_{i=1}^{N} (x_i - m)^2 .$$

To minimize the expression $(f(m) =) \sum_{i=1}^{N} (x_i - m)^2$, we differentiate it with respect to m and equate the derivative to zero – the solution of the equation is the sample mean $\hat{\mu} = \hat{\mu}^{OLS} = (x_1 + ... + x_N)/N = \bar{x}$ (this number is called the OLS or *ordinary least square estimate* of μ). If we repeat the same procedure with another concrete sample, we shall get a slightly different estimate $\hat{\mu}^{(2)}$, then, if we repeat the same procedure for the third time, we shall get $\hat{\mu}^{(3)}$ etc; hopefully all these estimates (that is, the realizations of the random *estimator*[1] $\hat{\mu} = (X_1 + ... + X_N)/N = \bar{X}$) will fluctuate around μ. Indeed, the estimator $\hat{\mu}$ (= $\hat{\mu}^{OLS}$) is an *unbiased* estimator of μ because $E\hat{\mu} = E(X_1 + ... + X_N)/N = \mu$. Based on random $\hat{\mu}$, we can test different hypothesis about μ (to do this, if the sample is small, we also must know the distribution of X, but if N is "large", then according to the Central Limit Theorem (CLT), $\hat{\mu} \sim N(\mu, (\sigma/\sqrt{N})^2)$ and, for example, $\hat{\mu} - 2\sigma/\sqrt{N} \le \mu \le \hat{\mu} + 2\sigma/\sqrt{N}$ with approximately 95% confidence which implies that when the sample size increases, we get more and more precise estimates of μ). The estimator $\hat{\mu}$ is also a *consistent estimator* of μ, that is, $\hat{\mu}$ tends to (or collapses onto) μ in probability as $N \to \infty$ (this follows from the Chebyshoff inequality: $P(|\hat{\mu} - \mu| > \lambda) \le \text{var}\,\hat{\mu}/\lambda^2 = \sigma^2/N\lambda^2 \to 0$).

---

[1] Let the distribution F_θ of a r.v. X depends on unknown parameter θ. Definition. Any function f of the random sample $(X_1, X_2, ..., X_N)$ is called an *estimator* of θ and denoted $\hat{\theta}$, $\hat{\theta} = f(X_1, ..., X_N)$. ◀ Estimator is a random variable and we are interested only in "good" estimators, for example, the unbiased ones, i.e., such that $E\hat{\theta} = \theta$.

2. Much of applied econometric analysis begins with the following premise: Y and X are two variables and we are interested in "explaining Y in terms of X". Some examples are: Y is oats yield and X is amount of fertilizer or Y is wage and X is years of education etc. Thus, we are looking for a function f which describes the relationship $Y = f(X)$. On the other hand, any single X will never explain an economic variable Y, therefore all other non-observable factors influencing Y can be included into the random *disturbance term* ε: $Y = f(X, \varepsilon)$ (here X can be nonrandom (as in 2.2 example) or random (as in 2.1 or 2.4 examples); in what follows, we shall assume, as is common in economics, that X is random but, in any case, Y is a r.v. which depends on both X and ε). The simplest function f is linear: $Y = \beta_0 + \beta_1 X + \varepsilon$ and, when this function is appropriate, our purpose is to restore (estimate) its unknown coefficients β_0 and β_1 from all available data, i.e., the sample[2] $(X_1, Y_1), ..., (X_N, Y_N)$.

Thus, we expect that "on average" the relationship $Y = \beta_0 + \beta_1 X$ holds but this "on average" concept is rather subtle. The two-dimensional r.v. (X, Y) (more precisely, its distribution) is, in fact, described by the two-dimensional r.v. (X, ε) but, in order to get "good" estimates of β_0 and β_1, we have to impose certain restrictions on the properties of ε and, also, on interaction (if any) between X and ε. Let us denote by $E(Y \mid X)$ a conditional expectation of Y provided we know the value of r.v. X or, shorter, *conditional expectation of Y on X* (this conditional expectation is closely related to the conditional probability $P(B \mid A)$ or conditional density $f(y \mid x)$ or similar concepts). Here are some properties of $E(Y \mid X)$ (we assume that X is a r.v.):

CE1 $E(Y \mid X)$ depends on (is a function of) the r.v. X, i.e., it is a random variable itself; for instance, in 2.1 example, where X is a discrete r.v.,

$$E(Y \mid X) = \begin{cases} E(Y \mid X = x_6) \text{ with probability } P(X = x_6) \\ ..... \\ E(Y \mid X = x_{18}) \text{ with probability } P(X = x_{18}) \end{cases}$$

and $\text{var}(Y \mid X = x) = \sum \left(y_j - E(Y \mid X = x) \right)^2 P(Y = y_j \mid X = x)$,

or, in 2.2 example, where X is a continuous r.v.,

$$E(Y \mid X = x) = \int_{-\infty}^{\infty} y \, f_{Y|X}(y \mid x) \, dy$$

(here the conditional density $f_{Y|X}(y \mid x)$ is defined as $f_{(X,Y)}(x, y) / f_X(x)$).

[2] According to our convention, we should use lower case letters to denote concrete sample and upper case letters to denote random sample, but following tradition we will always use upper case letters (is this a concrete or random sample, will be clear from the context).

2.1 example. Let (X,Y) be a bivariate normal r.v. It can be shown that

$$f_{Y|X}(y\,|\,x)=\frac{1}{\sqrt{2\pi}\sigma_Y(1-\rho^2)}\exp\left\{-\frac{1}{2\sigma_y^2(1-\rho^2)}\left[y-\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)-\mu_Y\right]^2\right\}$$

which is equal to the density function of a one-dimensional normal r.v. with mean $\mu_Y+(\sigma_Y/\sigma_X)(x-\mu_X)$ and variance $(1-\rho^2)\sigma_y^2$. Thus we get $E(Y\,|\,X=x)=\mu_Y+\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)$.

CE2 $E(f(V)\,|\,V)=f(V)$ (once you know V, $f(V)$ is fully explained).

CE3 If the r.v.'s V and U are independent (i.e., $P(V\in B\,|\,U\in A)=P(V\in B)$ for any events A and B), then $E(V\,|\,U)=EV$.

CE4 $E(E(\varepsilon\,|\,X))=E\varepsilon$ (for example, $\int_{-\infty}^{\infty}E(\varepsilon\,|\,X=x)f_X(x)\,dx=E\varepsilon$).

CE5 $E(\beta_0+\beta_1X+\varepsilon\,|\,X)=\beta_0+\beta_1E(X\,|\,X)+E(\varepsilon\,|\,X)$ $(=\beta_0+\beta_1X+E(\varepsilon\,|\,X))$, i.e., conditional expectation is a <u>linear</u> operator.

CE6 The following two identities are called the *Laws of iterated expectations*: let Y, X_1, and X_2 be three (generally, dependent) r.v.'s; then

$$E(E(Y\,|\,(X_1,X_2))\,|\,X_1)=E(Y\,|\,X_1)$$
$$E(E(Y\,|\,X_1)\,|\,(X_1,X_2))=E(Y\,|\,X_1)$$

(the smaller information set[3] always dominates). For example, if $Y=\beta_0+\beta_1X_1+\beta_2X_2+\varepsilon$, then $E(E(\beta_0+\beta_1X_1+\beta_2X_2+\varepsilon\,|\,(X_1,X_2))\,|\,X_1)=E(\beta_0+\beta_1X_1+\beta_2X_2+\varepsilon\,|\,X_1)=\beta_0+\beta_1X_1+\beta_2E(X_2\,|\,X_1)+E(\varepsilon\,|\,X_1)$. Another example: let $Y_1,Y_2,...,Y_{t-2},Y_{t-1},Y_t$ be a time series and Ω_s, $1\le s\le t$, a set containing all the information about $Y_1,Y_2,...,Y_{s-1},Y_s$; then

$$E(E(Y_t\,|\,\Omega_{t-2})\,|\,\Omega_{t-1})=E(Y_t\,|\,\Omega_{t-2})$$
$$E(E(Y_t\,|\,\Omega_{t-1})\,|\,\Omega_{t-2})=E(Y_t\,|\,\Omega_{t-2})$$

◀◀

Note that all these properties can be generalized to the case where X is a vector-valued r.v. $\vec{X}$.

The CE5 property implies that, <u>if</u> $E(\varepsilon\,|\,X)\equiv 0$ for every given X, the r.v. Y is on average $\beta_0+\beta_1X$ (this line is called a *regression line*), i.e.,

$$E(\beta_0+\beta_1X+\varepsilon\,|\,X)=\beta_0+\beta_1X$$

To estimate β_0 and β_1, we can use the least squares method again: find $\hat{\beta}_0$ and $\hat{\beta}_1$ such that

[3] Clearly, two r.v.'s X_1 and X_2 contain more information about (can better explain) Y than only X_1.

$$\sum_{i=1}^{N}(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i)^2=\min_{b_0,b_1\in R}\sum_{i=1}^{N}(Y_i-b_0-b_1X_i)^2$$

(here X_i, Y_i are the sample values). To find the formulas for these estimators, one has to differentiate the rhs expression with respect to b_0 and b_1, equate the derivatives to zero and solve respective system. The procedure is described in Sec. 3.2 where also the conditions (in addition to $E(\varepsilon \mid X)=0$) for $\hat{\beta}_0$ and $\hat{\beta}_1$ to be "good" estimators of β_0 and β_1 are presented.

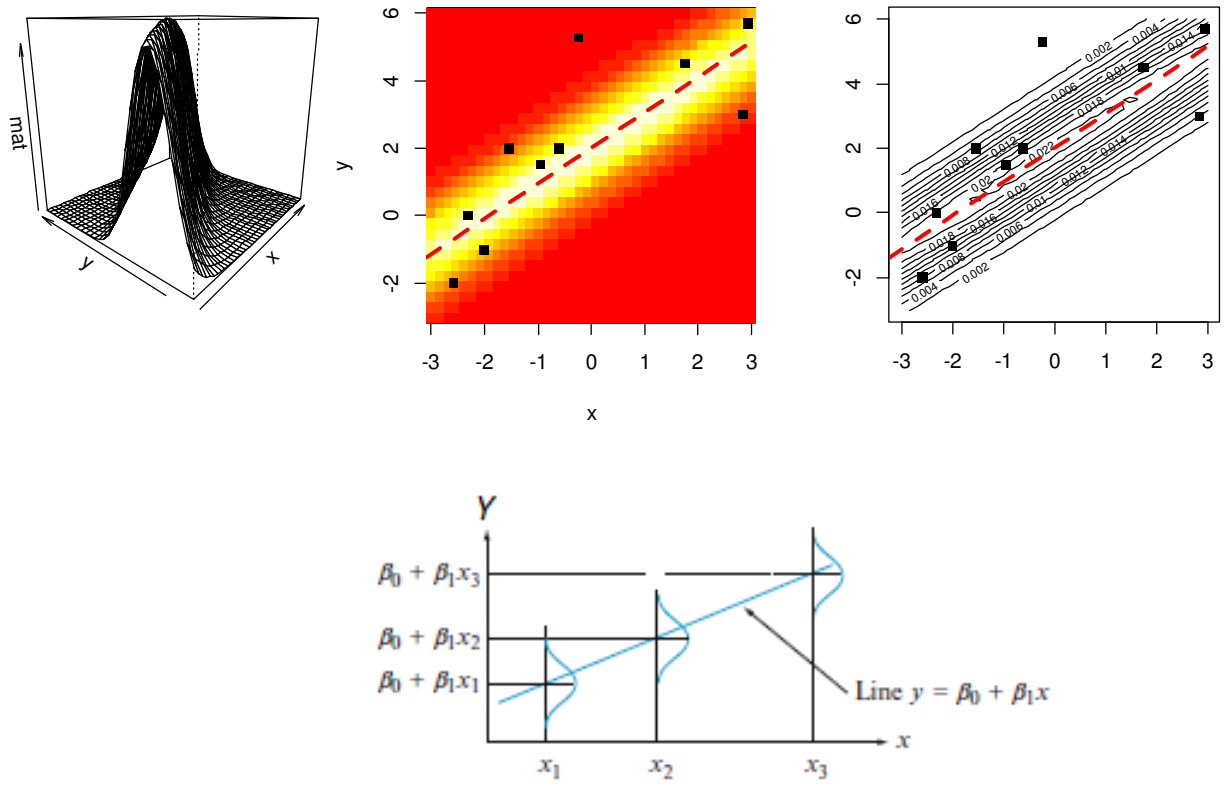


Figure 3.1. The graph of the joint density of (X,Y) (left); two variants of the same plot with sample points and regression line $\beta_0+\beta_1X$ (center and right) ; three conditional distributions of Y (bottom; note that $E(Y \mid X=x_2)=\beta_0+\beta_1x_2$)

****

In the previous chapter, we presented some examples of (X,Y) data and their regression models. However, even if our assumption that the data is generated by the linear[4] model

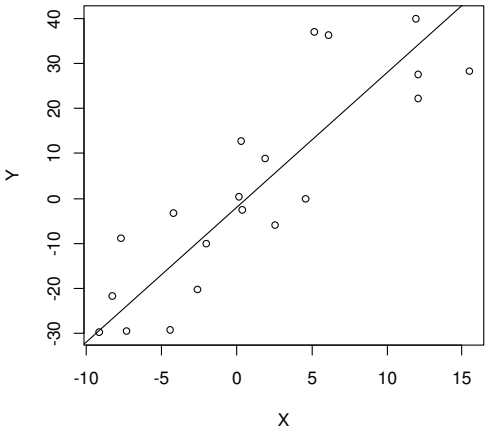
[4] Recall that the word 'linear' describes the way the coefficients enter the model. Thus the model $Y=\beta_0+\beta_1X^2+\varepsilon$ $(=\beta_0+\beta_1Z+\varepsilon)$ is also linear with respect to β_0 and β_1. More explicitly, it should be called linear quadratic model („linear" applies to coefficients and „quadratic" to X^2).

$Y = \beta_0 + \beta_1 X + \varepsilon$ is correct, we usually do not know the exact values of the coefficients β_0 and β_1.

In this chapter, we explain how to use the sample data to find the "best" (under certain conditions) estimators of the coeficients β_0 and β_1 in $Y = \beta_0 + \beta_1 X + \varepsilon$

To present some variants of the estimation procedures, we shall analyze the following example. Let us assume that our data is a sample from the (X,Y) population created by the data generating process (DGP) which is described by the equation $Y = -2 + 3X + \varepsilon$. We treat X as a normal random variable and use R to generate a sample of size N=20:

```
N=20
set.seed(6)
X=rnorm(N,sd=7)
eps=rnorm(N,sd=10)# X and ε are ind.
Y=-2+3*X+eps
plot(X,Y)
abline(-2,3)
```



The black regression line $Y = (\beta_0 + \beta_1 X =)\ -2 + 3X$ shows the „true" relationship between variables X and Y (we do not control (do not measure) all the other variables potentially influencing Y, therefore we treat them as a noise or disturbance or model's error ε). We would like to estimate (i.e., approximately restore from our data) the coefficients $\beta_0 (= -2)$ and $\beta_1 (= 3)$ and then draw a respective estimate of the regression line $Y = \hat{\beta}_0 + \hat{\beta}_1 X$.

A few words about the terminology. The estimation of β_1 (denoted by $\hat{\beta}_1$) is obtained with the help of certain formulas that use the sample values of Y and X; $\hat{\beta}_1$ is called an *estimate* of β_1 (it is a number). On the other hand, every time you repeat the above procedure you will get another sample and another estimate, thus $\hat{\beta}_1$ now can be treated as a random variable and in such a status it is called an *estimator* of β_1. To denote the estimator, we use the same symbol $\hat{\beta}_1$ but now we can speak (among other things) about its mean $E\hat{\beta}_1$ (it is desirable to have an estimation procedure such that $E\hat{\beta}_1 = \beta_1$; such an estimator is called *unbiased*) or about its variance $\operatorname{var} \hat{\beta}_1$ (it would be nice to have an estimator with the minimum possible variance; such an estimator is called *effective*).

How can we restore (or estimate) β_0 and β_1? We shall discuss a few procedures.

1. One can draw the line by eye and then deduce from the plot $\hat{\beta}_0$ and $\hat{\beta}_1$. However, this procedure is hardly reproducible. On the other hand, it is not quite clear how to explain the procedure of drawing to a computer. Even more so, sometimes one can quite succesfully draw a

straight line, but if we need a more general model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, to plot a plane in 3-dimensional space is quite a challenge and altogether impossible in more than four dimensional case.

2. To draw a straight line we need two points in plane which can be selected, for example, as follows: select 1/3 of the smallest X's and calculate the average value of respective X's and Y's; repeat the same with 1/3 of greatest X's; draw a „regression" line through these two points; restore the intercept and the slope of the line.

Quite possible that sometimes you will get a line close to the „true" one. However, in a more general model with, for example, two explanatory variables it is not clear how to split the (X_1, X_2) plane in three parts (we need three points to draw a plane) and draw a regression plane etc

3. In the next section we shall present the most popular and most promissing method of least squares.

3.2. The Method of Least Squares

Here we shall present a general method of (*ordinary*) *least squares* (OLS) which produces "best" estimators provided some conditions are satisfied. Assume that

U1. The DGP is described by the linear (in respect of β_0 and β_1) model $Y = \beta_0 + \beta_1 X + \varepsilon$ - thus the output random variable (r.v.) Y depends on two r.v.'s, X and ε (note that, for example, $Y = \beta_0 + \beta_1 \sqrt{X} + \varepsilon$ is also a linear model).

The further conditions U2-U4 describe the properties of unobservable r.v. ε and its relationship with observable r.v. X. Let $\vec{X} = (X_1, ..., X_N)$, $\vec{Y} = (Y_1, ..., Y_N)$, and $\vec{\varepsilon} = (\varepsilon_1, ..., \varepsilon_N)$ be our samples of, respectively, explanatory variable, response variable, and errors. In linear model, ε and Y are always dependent (because $\operatorname{cov}(Y, \varepsilon) \neq 0$), but ε may depend on X and may not. The U2 condition below demands them to be close to independent.

U2. $E(\varepsilon_i \mid X_j) \equiv 0,\ i, j = 1, ..., N$ – this means that whatever are the observations X_j, the errors on average equal 0; this claim also implies that $E(\varepsilon_i \mid X_i) \equiv 0$ and $E\varepsilon_i \equiv E\big(E(\varepsilon_i \mid X_i)\big) = 0$. For example, U2 is satisfied if for any $i = 1, ..., N$ and $j = 1, ..., N$, r.v.'s X_i and ε_j are independent[5] and $E\varepsilon = 0$ – then $E(\varepsilon_i \mid \vec{X}) = E\varepsilon_i = 0$. However, if ε_i and X_i are correlated, i.e., $\operatorname{cov}(\varepsilon_i, X_i) \neq 0$, then U2 will not hold. Indeed, we have $\operatorname{cov}(X_i, \varepsilon_i) = E(X_i - EX_i)\varepsilon_i = E\big((X_i - EX_i)E(\varepsilon_i \mid \vec{X})\big) \neq 0$ which contradicts U2. Note that U2 implies $E(Y_i \mid X_i) = \beta_0 + \beta_1 X_i$.

[5] Thus, say, ε_5 is influenced by neither X_5 nor X_{15}.

U3. $\mathrm{var}(\vec{\varepsilon} \mid \vec{X}) = \begin{pmatrix} \mathrm{var}(\varepsilon_1) & \mathrm{cov}(\varepsilon_1,\varepsilon_2) & \mathrm{cov}(\varepsilon_1,\varepsilon_3)\ldots. \\ \mathrm{cov}(\varepsilon_2,\varepsilon_1) & \mathrm{var}(\varepsilon_2) & \mathrm{cov}(\varepsilon_2,\varepsilon_3)\ldots. \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{pmatrix} = \sigma_\varepsilon^2 I$, i.e., the conditional variance-covariance matrix[6] of $\vec{\varepsilon}$ is a unit matrix times a constant σ_ε^2; to put it simpler, $\mathrm{var}(\varepsilon_i \mid \vec{X}) \equiv \sigma_\varepsilon^2$, i.e., (conditional) variance of ε_i is the same for all i's[7] and does not depend on the values of $\vec{X}$ and, also, $\mathrm{cov}(\varepsilon_i,\varepsilon_j) \mid \vec{X} = E(\varepsilon_i\varepsilon_j \mid \vec{X}) = 0$[8] for all $i \neq j$, i.e., disturbances do not interact (as a counterexample, if we describe the $(weight, height)$ DGP as $weight = \beta_0 + \beta_1 height + \varepsilon$, then probably $\mathrm{var}\,\varepsilon$ increases together with $height$ (see Fig. 3.2)[9], thus, this data does not satisfy U3).

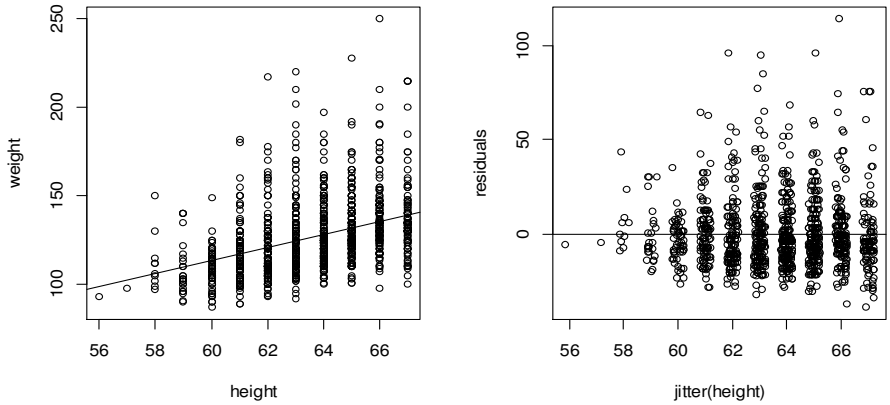


Figure 3.2. The spread of the residuals increases together with `height`

Usually we consider two cases where U3 does not hold:

U3h: if $\mathrm{var}(\varepsilon_i \mid (X_1,...,X_N))$ depends on X_i, the errors are called heteroskedastic;
U3a: if $\mathrm{cov}(\varepsilon_i,\varepsilon_j) \mid (X_1,...,X_N) \neq 0$, $i \neq j$, the errors are called autocorrelated.

These two cases deserve special treatments and will be considered later.

U4. Sometimes a requirement of normality is added: $\vec{\varepsilon} \mid \vec{X} \sim N(\vec{0}, \sigma_\varepsilon^2 I)$ (conditional density of $\vec{\varepsilon}$ (and, therefore, of any ε_i) is normal or, at least, close to symmetric) - this requirement simplifies some statistical properties of estimators (for example, $\hat{\beta}_i$ will be normal). Transforming the outcome is often successful for reducing the skewness of residuals. The rationale is that the more extreme values of the outcome are usually the ones with large residuals (defined as $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$); if we can "pull in" the outcome values in the tail of the distribution toward

[6] Attention: if $\vec{Z}$ is a random vector, $\mathrm{var}\,\vec{Z}$ is a matrix; here, inside the matrix, we have skipped conditioning.
[7] Such ε's are termed homoskedastic.
[8] If $\mathrm{cov}(\varepsilon_i,\varepsilon_j) \mid \vec{X} \neq 0$, such errors are termed serially correlated or autocorrelated.
[9] `weight` and `height` are taken from the (edited) dataset `babies` in R's package UsingR.

the center, then the corresponding residuals are likely to be smaller too (one such transformation is to replace the outcome Y with $\log Y$). Power transformations are a flexible alternative to the log transformation. In this case, Y is replaced by Y^{λ}, $\lambda > 0$ (smaller values of λ "pull in" the right tail more strongly). As an example, square ($\lambda = 1/2$) and cube ($\lambda = 1/3$) root transformations were commonly used in analyzing variables long tailed on the right. However, some outcome variables cannot be satisfactorily normalized by transformation, or there may be compelling reasons to analyze them on the original scale. In such a case, it is useful to construct bootstrap confidence intervals for β_i. ◀◀

In short – we say that the

Data Generating Process $Y = \beta_0 + \beta_1 X + \varepsilon$ satisfies U2-U3 if (conditionally on all X 's) $E\varepsilon_i \equiv 0$, $\operatorname{var}\varepsilon_i \equiv \sigma_\varepsilon^2$, and ε_i does interact neither with X_i nor with other ε_j

The „line" $Y = \beta_0 + \beta_1 X$ is called the *regression line*[10] (the words „regression line" is a shortcut for what is probably a *regression curve*, for example, $\log V = \beta_0 + \beta_1 \log U$ [11] or a similar curve). Under U2 we can also give another interpretation of regression line: since $Y = \beta_0 + \beta_1 X + \varepsilon$, $E(Y_i \mid X_1, ..., X_N) = \beta_0 + \beta_1 X_i + E(\varepsilon_i \mid X_1, ..., X_N) = \beta_0 + \beta_1 X_i$ thus the expected value of the response variable at X_i coincides with the regression line.

We do not know the coefficients β_0 and β_1 and in order to estimate these coefficients from our sample (X_i, Y_i), $i = 1, ..., N$, we would like to use the data in the best possible way to obtain the *estimate of a regression* line $Y = \hat{\beta}_0 + \hat{\beta}_1 X$. The „best" way is described by the *Gauss-Markov theorem*:

Under assumption that the conditions U1-U3 hold true, the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are BLUE&C

Here BLUE means Best Linear Unbiased Estimator and C means Consistent – in other words, OLS estimator is indeed very good[12]. However, in order to prove this we have first to define what is meant by the words *OLS estimator*.

[10] β_0 is called the intercept and β_1 the slope of the regression line.

[11] It is a curve $V = \exp(\beta_0) U^{\beta_1}$ in (U, V) coordinates, but a straight line in $(\log U, \log V)$ coordinates.

[12] Later we shall discuss these terms in more details.

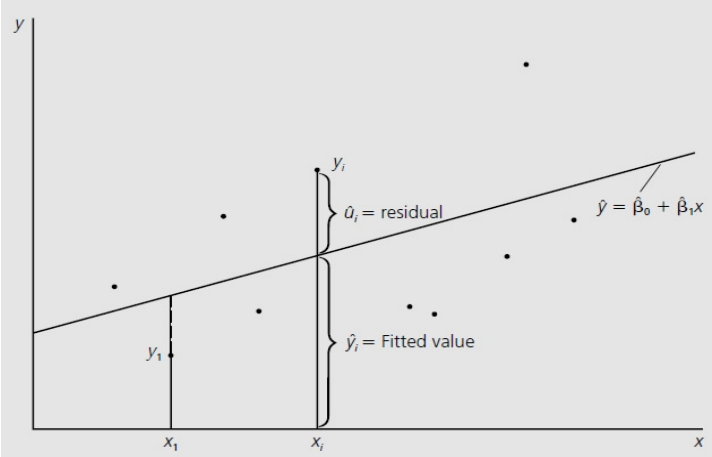


Figure 3.3. (X, Y) scatter diagram and the estimated regression line

To explain the notion, have a look at Fig. 3.3 where one can see the scatter diagram of our sample data. We want to draw a straight line through the „middle of the cloud“, i.e., choose b_0 and b_1 such that the residual sum of squares $RSS = RSS(b_0, b_1) = \sum_{i=1}^{N} (Y_i - (b_0 + b_1 X_i))^2$ were as small as possible. Below, we shall present a few techniques to achieve this goal.

Let X be a r.v. with distribution F_μ where μ is unknown mean. If the sample $(X_1, ..., X_N)$ is „good“ (for example, our observations X_i are independent identically distributed r.v.'s or iidrv's), then the OLS estimate of μ, $\hat{\mu} = \overline{X}$, is BLUE&C.

Let $Y = \beta_0 + \beta_1 X + \varepsilon$ where β_0, β_1 and σ_ε^2 are unknown. If the sample $\begin{pmatrix} Y_1 & \dots & Y_N \\ X_1 & \dots & X_N \end{pmatrix}'$ is "good" (for example, satisfies U2-U4), then the OLS estimators of unknown parameters will be BLUE&C.

- **The system of partial derivatives**

Recall that in order to minimize RSS, we have to differentiate it with respect to b_0 and b_1, equate the derivatives to zero and solve the following system of two equations:

$$\begin{cases} \dfrac{\partial RSS}{\partial b_0} = \dots = 0 \\ \dfrac{\partial RSS}{\partial b_1} = \dots = 0 \end{cases}. \qquad (3.1)$$

$$\begin{cases} \sum (Y_i - (b_0 + b_1 X_i)) = 0 \\ \sum X_i (Y_i - (b_0 + b_1 X_i)) = 0 \end{cases}. \qquad (3.2)$$

After some calculation, we obtain that the solutions $\hat{\beta}_0$ and $\hat{\beta}_1$ of the the system are equal to

$$\begin{cases} \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \\ \hat{\beta}_1 = \dfrac{\sum (Y_i - \bar{Y})(X_i - \bar{X})}{\sum (X_i - \bar{X})^2} = \dfrac{\sum (Y_i - \bar{Y}) X_i}{\sum (X_i - \bar{X})^2} = \dfrac{\sum (X_i - \bar{X}) Y_i}{\sum (X_i - \bar{X})^2} = \dfrac{\widehat{\mathrm{cov}}(X,Y)}{\widehat{\mathrm{var}} X} \end{cases} \qquad (3.3)$$

(first estimate $\hat{\beta}_1$ and then $\hat{\beta}_0$). Note that the estimate of the regression line goes through the point $(\bar{X}, \bar{Y})$. (Can you estimate β_0 in the simplest regression $Y = \beta_0 + \varepsilon$ and tell what is the meaning of $\hat{\beta}_0$?)

$\hat{\beta}_0$ *and* $\hat{\beta}_1$ *are called the OLS estimates* of the linear regression parameters (the procedure of calculating these estimates in R is performed with `lm(Y~X)` and in GRETL with `ols Y c X`). The straight line $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ is called an *estimated regression line* (or estimated Y or predicted Y) and $e_i = \hat{\varepsilon}_i = Y_i - \hat{Y}_i$ *residuals* of the model (they are hopefully close to the errors ε_i). In fact, there is one more parameter in the model we have to estimate, the variance of the error:

$$\hat{\sigma}_\varepsilon^2 = s_\varepsilon^2 = RSS(\hat{\beta}_0, \hat{\beta}_1)/(N-2) = \sum (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2 / (N-2) = \sum e_i^2 / (N-2) \qquad (3.4)$$

The second formula in (3.3) can still be simplified by introducing $x_i = X_i - \bar{X}$ and $y_i = Y_i - \bar{Y}$:

$$\hat{\beta}_1 = \begin{cases} \sum \dfrac{x_i}{\sum x_i^2} Y_i = \sum w_i Y_i \\ \dfrac{\sum x_i (\beta_0 + \beta_1 X_i + \varepsilon_i)}{\sum x_i^2} = \beta_1 + \sum \dfrac{x_i}{\sum x_i^2} \varepsilon_i = \beta_1 + \sum w_i \varepsilon_i \end{cases} \qquad (3.5)$$

This formula implies that the estimators $\hat{\beta}_1$ (and $\hat{\beta}_0$) are r.v.'s depending on two random vectors $\vec{X}$ and $\vec{\varepsilon}$.

3.3. Properties of the OLS Estimator

Now we are able to prove that $\hat{\beta}_1$ is BLUE&C[13] (the proof for $\hat{\beta}_0$ is similar).

L The linearity of the estimator with respect to Y_i follows from the first line in (3.5).

[13] The BLUE properties hold for any size samples whereas C (=consistency) is a large sample or asymptotic property.

U To prove unbiasedness, that is $E(\hat{\beta}_1 \mid \vec{X}) = \beta_1$, we use the second line in (3.5) and U2: $E(\hat{\beta}_1 \mid \vec{X}) = \beta_1 + \sum w_i E(\varepsilon_i \mid \vec{X}) = \beta_1$.

B Best means that, in the class of linear unbiased estimators, $\operatorname{var} \sum z_i Y_i$ attains its minimum when $z_i = w_i$ (in other words, $\hat{\beta}_1$ is an *efficient estimator* of β_1). The proof of this fact (assuming U3) is beyond the scope of the course but to calculate the $\operatorname{var}(\hat{\beta}_1 \mid \vec{X})$ itself under U3 is easy (in what follows we shall usually skip the conditioning):

$$\operatorname{var}(\beta_1 + \sum w_i \varepsilon_i) = \operatorname{var} \sum w_i \varepsilon_i = \sum_{i=1}^{N} \sum_{j=1}^{N} w_i w_j \operatorname{cov}(\varepsilon_i, \varepsilon_j) = \sigma_\varepsilon^2 \sum w_i^2 = \frac{\sigma_\varepsilon^2}{\sum x_i^2}. \qquad (3.6)$$

This formula implies that (according to the Law of Large Numbers), $\sum x_i^2 = N \cdot \left(\sum x_i^2 / N \right) \approx N \cdot \operatorname{var} X$, thus $\operatorname{var} \hat{\beta}_1 \to 0$ when $N \to \infty$, i.e., $\hat{\beta}_1$ becomes closer and closer to the true value of β_1.

Note that we do not know σ_ε^2 therefore we can use the *analogue principle* to estimate $\operatorname{var} \hat{\beta}_1$:

$$\widehat{\operatorname{var}} \hat{\beta}_1 = s_\varepsilon^2 / \sum x_i^2 \qquad (3.7)$$

(notice the difference between var (the true variance) and $\widehat{\operatorname{var}}$ (estimated variance)!).

Definition. The square root of the estimated variance of the estimator $\hat{\beta}_i$, that is, $\sqrt{\widehat{\operatorname{var}} \hat{\beta}_i}$ is called the standard error of respective estimator $\hat{\beta}_i$. ◀
The standard error describes the accuracy of the estimator (the smaller the better).

C We have just seen that $\hat{\beta}_1$ converges, in some sense[14], to the true value β_1. However, here we were dealing with the simplest Gauss-Markov model; in more general cases, it is not so easy to calculate $\operatorname{var} \hat{\beta}_1$. A similar but more convenient concept is convergence in probability: we say that a sequence of r.v.'s W_n *converges in probability* to a nonrandom constant w if $P(|W_n - w| > \lambda) \to 0$ (or, what is the same, $P(|W_n - w| < \lambda) \to 1$) as $n \to \infty$ for any positive number λ[15] (this is usually denoted by $W_n \xrightarrow{P} w$). The following properties are important:

- If $W_n \xrightarrow{P} w$ and $Z_n \xrightarrow{P} z$, then $W_n + Z_n \xrightarrow{P} w + z$ and $W_n Z_n \xrightarrow{P} wz$
- If $W_n \xrightarrow{P} w$, $Z_n \xrightarrow{P} z$, and $z \neq 0$, then $W_n / Z_n \xrightarrow{P} w / z$

[14] In variance sense.

[15] Even if λ is small (say, 0.0001), the inequality $|W_n - w| > 0.0001$ finally becomes improbable.

One more general definition: the statistics[16] Θ_n is a *consistent estimator* of unknown population's parameter θ if $\Theta_n \xrightarrow{P} \theta$ (hence Θ_n eventually collapses on θ).

To prove that (under U2) $\hat{\beta}_1$ is a consistent estimator of β_1 is easy:

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{N}\sum x_i \varepsilon_i}{\frac{1}{N}\sum x_i^2} \xrightarrow{P} \beta_1 + \frac{\operatorname{cov}(X,\varepsilon)}{\operatorname{var} X} = \beta_1 + \frac{0}{\operatorname{var} X} = \beta_1$$

*****

We have proved that the OLS estimator $\hat{\beta}_1$ has all the properties (under U1-U3) a good estimator must have. Note that

- To prove unbiasedness, it suffices to assume U2
- To prove efficiency, it suffices to assume U3 (homoskedasticity and uncorrelatedness of errors)

To illustrate the BLUE&C properties, we return to our DGP $Y = -2 + 3X + \varepsilon$. Clearly, the estimate $\hat{\beta}_1$ depends on sample. In Fig. 3.4 one can see four (out of 5000 generated) different samples and four different estimates of regression line (we use R; the code is in CL, p. 2-18).

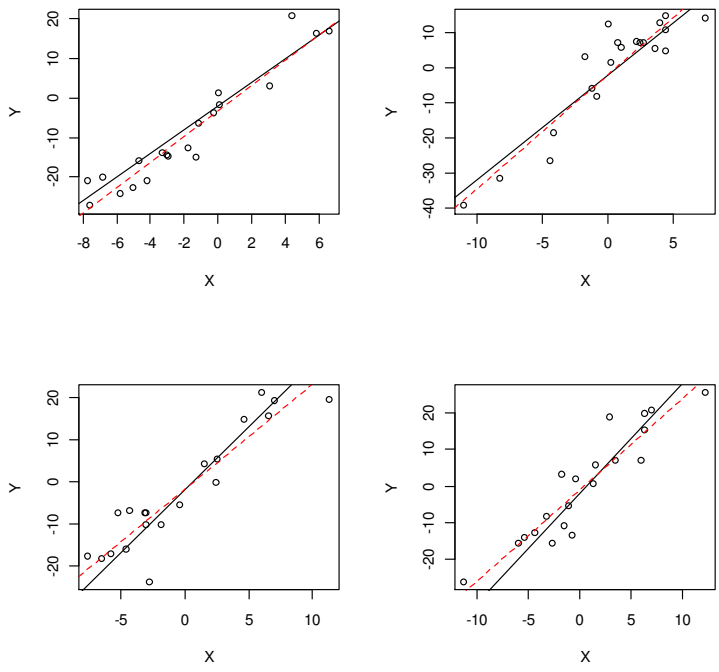


Figure 3.4. $N = 20$; four different samples and four different estimates (red) of the true regression line (black)

[16] Statistics is any function of a sample.

In order to demonstrate the unbiasedness of the OLS estimator, we must prove that $E\hat{\beta}_1=3$. We have already proved it "theoretically", but now we shall check the claim "empirically". To do this, we approximate $E\hat{\beta}_1$ by its sample mean – we generate many (5000) samples and calculate a sample mean `mean(beta1)=2.993`, i.e., $E\hat{\beta}_1$ ($\approx \widehat{E\hat{\beta}_1}$) is indeed very close to ($\beta_1 =$) 3. To illustrate that variance of $\hat{\beta}_1$ diminishes when the sample size N increases, we estimate $\text{var}(\hat{\beta}_1)$ by its sample variance – we use 5000 samples for $N = 10, 100$ and 1000:

$N = 10$ ($\widehat{\text{var}}(\hat{\beta}_1) =$) `var(beta1) = 0.210`
$N = 100$ ($\widehat{\text{var}}(\hat{\beta}_1) =$) `var(beta1) = 0.015`
$N = 1000$ ($\widehat{\text{var}}(\hat{\beta}_1) =$) `var(beta1) = 0.001`

A similar fact (consistency of $\hat{\beta}_1$) is explained by Fig. 3.5 – if the size of the sample increases, the deviations of $\hat{\beta}_1$ from ($\beta_1 =$) 3 become smaller and smaller (however, the rate of convergenge is rather slow).

Note that unbiasedness is true for any N while consistency is an asymptotic property (it holds when $N \to \infty$).

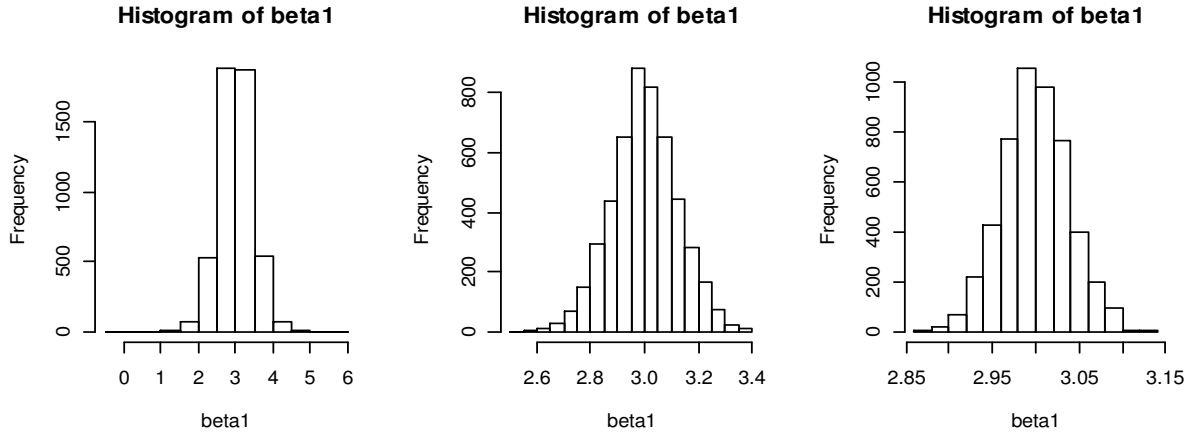


Figure 3.5. Three histograms of the estimates $\hat{\beta}_1$ (from left to right: $N = 10, 100, 1000$); consistency means that, when $N \to \infty$, the estimator $\hat{\beta}_1$ will collapse on β_1; also note that the histograms are bell-shaped

Figure 3.5 shows that the estimator $\hat{\beta}_1$ has a bell-shaped distribution. Under U4, we can prove even a more precise statement – since 1) $\hat{\beta}_1 = \beta_1 + \sum w_i \varepsilon_i$ and 2) uncorrelated ε_i have a normal distribution $N(0, \sigma_\varepsilon^2)$, the linear combination of ε's will be also normal: $\hat{\beta}_1 \sim N(\beta_1, \sigma_\varepsilon^2 / \sum x_i^2)$. A similar claim holds for $\hat{\beta}_0$.

3.4. Other Methods to Derive the OLS Estimates

The above formulas for $\hat{\beta}_1$ and $\hat{\beta}_0$ were obtained by directly solving two linear equations. This method is not convenient in multivariate case, therefore we shall present three more methods to derive respective formulas.

- **The method of moments (MM)**

If the population (or DGP) is described by two parameters (in our case, β_0 and β_1), in order to find them we have to create a system of two equations. To do this, the MM equates theoretical moments to sample moments. More specifically, we shall write the system

$$\begin{cases} \overline{\hat{\varepsilon}} = E\varepsilon & (=0) \\ \widehat{\mathrm{cov}}(X, \hat{\varepsilon}) = \mathrm{cov}(X, \varepsilon) & (=0) \end{cases}$$

which is equivallent (under U2 and some simplification of U3) to

$$\begin{cases} \sum (Y_i - (b_0 + b_1 X_i)) = 0 \\ \sum X_i (Y_i - (b_0 + b_1 X_i)) = 0 \end{cases} . \qquad (3.8)$$

Notice that this system coincides with (3.2), therefore we get the same solution. Next, (3.8) may be written as

$$\begin{cases} \sum \hat{\varepsilon}_i^{OLS} / N = 0 \\ \sum (X_i - \overline{X})(\hat{\varepsilon}_i^{OLS} - \overline{\hat{\varepsilon}}^{OLS}) / N = 0 \end{cases}$$

thus, by construction, i) the sample mean of the OLS residuals is always 0 and ii) the residuals do not correlate with *X*.

- **The matrix method**

It is convenient to use matrices when solving systems. Let $\vec{Y}$ be the column matrix $(Y_1, ...Y_N)'$, $\vec{\beta}$ the column matrix $(\beta_0, \beta_1)'$, $N \times 2$ *design matrix* $\mathbf{X} = \begin{pmatrix} 1 & X_1 \\ ..... & \\ 1 & X_N \end{pmatrix}$, and error column matrix $\vec{\varepsilon} = (\varepsilon_1, ..., \varepsilon_N)'$. The regression system now may be presented as $\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\varepsilon}$. Our goal is to find vector $\hat{\vec{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)$ such that $RSS(\vec{b}) = (\vec{Y} - \mathbf{X}\vec{b})'(\vec{Y} - \mathbf{X}\vec{b})$ attains its minimum at $\vec{b} = \hat{\vec{\beta}}$ ($\hat{\vec{\beta}}$ is the estimator of $\vec{\beta}$ we are looking for). The method to find $\hat{\vec{\beta}}$ is the same as earlier (partial derivatives) but the calculus is a bit tricky: as the derivative $\frac{\partial RSS}{\partial \vec{b}}$ is equal to $-2\mathbf{X}'\vec{Y} + 2\mathbf{X}'\mathbf{X}\vec{b}$, upon equating it to $\vec{0}$ we get $\mathbf{X}'\mathbf{X}\vec{b} = \mathbf{X}'\vec{Y}$ or

$$\hat{\vec{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\vec{Y}. \qquad (3.9)$$

Namely this formula is used by R and GRETL to calculate $\hat{\beta}_0$ and $\hat{\beta}_1$.

Note that in univariate regression

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ X_1 & X_2 & \dots & X_N \end{pmatrix}\begin{pmatrix} 1\,X_1 \\ 1\,X_2 \\ \dots\dots \\ 1\,X_N \end{pmatrix} = \begin{pmatrix} N & \sum X_i \\ \sum X_i & \sum X_i^2 \end{pmatrix},$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{N\sum X_i^2 - \left(\sum X_i\right)^2}\begin{pmatrix} \sum X_i^2 & -\sum X_i \\ -\sum X_i & N \end{pmatrix}$$

$$\mathbf{X}'\vec{Y} = \begin{pmatrix} \sum Y_i \\ \sum X_i Y_i \end{pmatrix}, \dots\dots\dots\dots\dots, \quad \hat{\vec{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} (\overline{X^2}\,\overline{Y} - \overline{X}\,\overline{XY}) / \widehat{\mathrm{var}}X \\ \widehat{\mathrm{cov}}(X,Y) / \widehat{\mathrm{var}}X \end{pmatrix},$$

that is, $\hat{\beta}_1 = \widehat{\mathrm{cov}}(X,Y) / \widehat{\mathrm{var}}X$ which is the same as in (3.3).

- **The „almost correct" method**

This method helps you to memorize (3.9). To estimate $\vec{\beta}$ in $\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\varepsilon}$, equate $\vec{\varepsilon}$ to $\vec{0}$ (anyway, it is on average zero): $\vec{Y} = \mathbf{X}\vec{\beta}$; to find $\hat{\vec{\beta}}$, multiply the equality from left by the inverse matrix $\mathbf{X}^{-1}$: $\mathbf{X}^{-1}\vec{Y} = \mathbf{X}^{-1}\mathbf{X}\vec{\beta} = \mathbf{I}\vec{\beta} = \hat{\vec{\beta}}$. However, our argument is faulty because the inverse $\mathbf{X}^{-1}$ is defined only for a square matrix, therefore we shall proceed as follows: $\mathbf{X}'\vec{Y} = \mathbf{X}'\mathbf{X}\vec{\beta}$ (now $\mathbf{X}'\mathbf{X}$ is a square matrix), $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\vec{Y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\vec{\beta} = \hat{\vec{\beta}}$ – this is exactly (3.9).

One final remark. The estimator $\hat{\vec{\beta}}$ is a (two-dimensional) r.v.[17], hence we can speak about its mean (we have already proven that it equals (β_0, β_1)) and its variance. With little matrix algebra it can be shown that the estimator of the variance-covariance matrix $\mathrm{var}\,\hat{\vec{\beta}} = \begin{pmatrix} \mathrm{var}\,\hat{\beta}_0 & \mathrm{cov}(\hat{\beta}_0, \hat{\beta}_1) \\ \mathrm{cov}(\hat{\beta}_1, \hat{\beta}_0) & \mathrm{var}\,\hat{\beta}_1 \end{pmatrix}$ is $\widehat{\mathrm{var}}\hat{\vec{\beta}} = s_\varepsilon^2\left(\mathbf{X}'\mathbf{X}\right)^{-1} = s_\varepsilon^2\left(\sum \vec{X}_i\vec{X}_i'\right)^{-1}$, where $\vec{X}_i = \begin{pmatrix} 1 \\ X_i \end{pmatrix}$. Note that $\widehat{\mathrm{var}}\hat{\beta}_1$ in this expression is the same as in (3.7), that is, $\widehat{\mathrm{var}}\hat{\beta}_1 = s_\varepsilon^2 / \sum x_i^2$.

[17] Because the formula contains random Y and $\vec{X}$.

3.5. Regression Model

In order to develop ideas of regression models, we are now going to use a simple but important economic example. Suppose that we are interested in studying the relationship between Y = `exp` = weekly food expenditure per person and X = `inc` = weekly household income. Economic theory tells us that expenditure on economic goods depends on income. Consequently we call Y the dependent or *response* variable and X independent or *explanatory* variable. An econometric analysis of expenditure relationship can provide answers to important questions, such as: if the income goes up by $100, how much will average weekly food expenditures rise? How much would we predict the weekly per person expenditure on food to be for a household with an income of $2000 per week? Such information is valuable for assesing existing market conditions, product distribution patterns, consumer buying habits, and consumer living conditions.

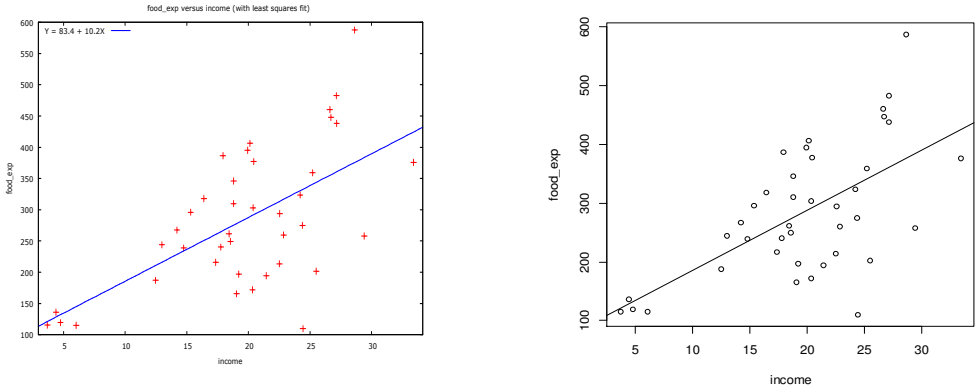


Figure 3.6. `inc-exp` scatter diagrams in GRETL (left) and R (right)

We begin with a data set food.gdt in GRETL\POE. It contains two variables, `exp` and `inc`, its XY scatterplot in GRETL is accompanied by an estimate of a regression line (blue). Respective regression model is given by the following output table:

```
Model 1: OLS, using observations 1-40
Dependent variable: exp

             coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const        83.4160      43.4102      1.922    0.0622   *
  inc          10.2096       2.09326     4.877    1.95e-05 ***

Mean dependent var   283.5735   S.D. dependent var   112.6752
Sum squared resid    304505.2   S.E. of regression   89.51700
R-squared            0.385002   Adjusted R-squared   0.368818
F(1, 38)             23.78884   P-value(F)           0.000019
Log-likelihood      -235.5088   Akaike criterion     475.0176
Schwarz criterion    478.3954   Hannan-Quinn         476.2389
```

Similar graph (see Fig. 3.6, right) and the output table in R look very much the same:

```
Call:
lm(formula = exp ~ inc)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   83.416     43.410   1.922   0.0622 .
inc           10.210      2.093   4.877 1.95e-05 ***

Residual standard error: 89.52 on 38 degrees of freedom
Multiple R-squared: 0.385,      Adjusted R-squared: 0.3688
F-statistic: 23.79 on 1 and 38 DF,  p-value: 1.946e-05
```

Is the equation $\widehat{exp} = 83.416 + 10.210\, inc$ a good model of our (unknown) DGP? In what follows, we shall try to answer this question.

3.6. Four important distributions

- Random variable $(X =)\ N(\mu, \sigma^2)$ is called a *normal* or *Gaussian* if its bell-shaped density function equals

$$f(x) = \varphi_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma} \right)^2 \right), -\infty < x < \infty .$$

Here μ and σ^2 are, respectively, the expectation $EX = \int_R x\varphi_{\mu,\sigma^2}(x)dx$ and variance $\operatorname{var} X = E(X - EX)^2 = \int_R (x - EX)^2 \varphi_{\mu,\sigma^2}(x)dx$ of $N(\mu, \sigma^2)$. The r.v. $(N(0,1) =)\ N$ is called standard normal. Note that not every bell-shaped density is normal (see, e.g., Student's density below).

- The random variable χ_n^2 is called a *chi squared* r.v. with n degrees of freedom (df) if its density function is the same as that of $\sum_{i=1}^{n} N_i^2$ where N_i are standard independent normal r.v.'s. Its expectation is n and variance $2n$, therefore according to CLT for big values of n the r.v. $\left(\chi_n^2 - n\right)/\sqrt{2n}$ is close to N. Note that $\chi_5^2 = (N_1^2 + N_2^2 + N_3^2) + (N_4^2 + N_5^2) = \chi_3^2 + \chi_2^2$ or, in general, $\chi_{r+s}^2 = \chi_r^2 + \chi_s^2$.

- Let $N, N_1, ..., N_n$ be independent standard normal r.v.'s. Random variable T_n is called the *Student* r.v. with n df if its density function is the same as that of $T_n = \frac{N}{\sqrt{\left(N_1^2 + ... + N_n^2\right)/n}} = \frac{N}{\sqrt{\chi_n^2 / n}}$. According to the Law of Large Numbers (LLN) $\chi_n^2 / n \xrightarrow{P} EN^2 = 1$, therefore for big values of n the r.v. T_n is close to N. Note that if $X_1, ..., X_n$ is a sample from normal population with mean μ and variance σ^2, then the so-called $t-$ratio statistics (of $\bar{X}$) $\left(\frac{\bar{X} - \mu}{s.e.\ \bar{X}} = \right) \frac{\bar{X} - \mu}{S/\sqrt{n}}$ (here $S^2 = \hat{\sigma}_X^2 = \sum (X_i - \bar{X})^2 / (n-1)$ is the estimator of σ_X^2) has the T_{n-1} distribution.

- Let χ_n^2 and χ_m^2 be independent chi squared r.v.'s. The ratio $\frac{\chi_n^2 / n}{\chi_m^2 / m}$ is called the Fisher r.v. with (n,m) df and denoted by $F_{n,m}$. Both χ_n^2 and $F_{n,m}$ are positive r.v.'s (they take only positive values which means that their density functions, for example $f_{F_{n,m}}(x)$, equals zero for $x \le 0$). If m is large, according to the LLN, $\chi_m^2 / m \approx E N^2 = 1$, therefore $F_{n,m} \approx \chi_n^2 / n$ for large m.

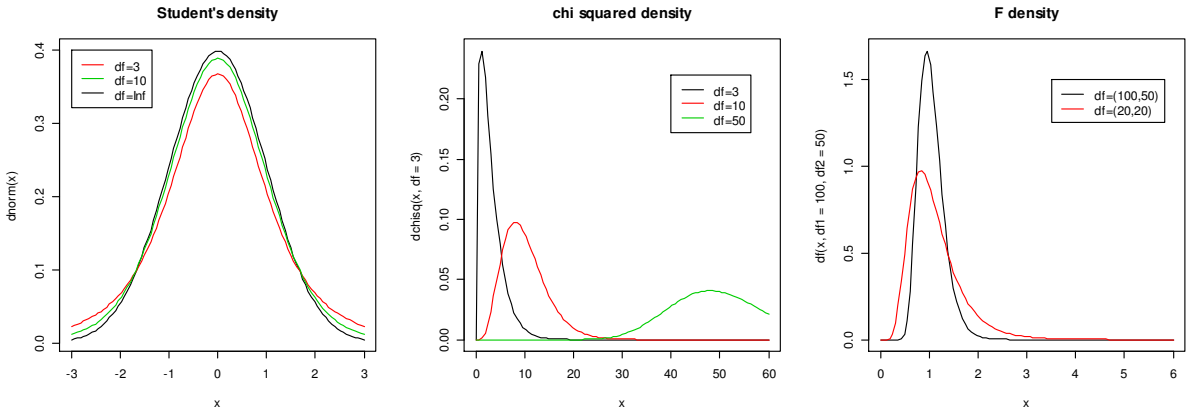


Figure 3.7. The densities of the Student T_n (they are close to normal and coincide with normal if df=inf), χ_n^2 (they equal 0 if $x < 0$ and have maximum at $n-2$) and $F_{n,m}$ distributions.

3.7. Hypothesis testing

Let us consider once again the DGP $Y = \beta_0 + \beta_1 X + \varepsilon$. We say that the variable X is *significant* in the model if $\beta_1 \neq 0$. Since we do not know the true value of β_1, we can only test the hypothesis $H_0 : \beta_1 = 0$ with different alternatives $H_1 : \beta_1 \neq 0$, $H_1 : \beta_1 > 0$ or $H_1 : \beta_1 < 0$. As always with hypotheses, we accept[18] the null if $\hat{\beta}_1$ is „close enough" to 0. It can be proved that, provided H_0 is true, the `t-ratio` statistics (of $\hat{\beta}_1$) $T = \frac{\hat{\beta}_1 - 0}{s.e.\,\hat{\beta}_1} = \hat{\beta}_1 / \sqrt{\widehat{\mathrm{var}}\hat{\beta}_1} =$ $= \hat{\beta}_1 / \sqrt{S_\varepsilon^2 / \sum x_i^2}$ does not depend on $\mathbf{X}$ and has the T_{N-2} distribution; thus, we reject H_0 and accept H_1 if $|t|$ exceeds[19] a respective Student's 5% critical value. The value depends on N, but the two-sided one is always close to 2:

```
> qt(.975, df = c(10,20,50,100,1000))
[1] 2.228139 2.085963 2.008559 1.983972 1.962339
```

[18] More correct is to say „we do not reject H_0".

[19] T is a r.v., its realization in a sample is a number denoted by *t*. In GRETL *t* is called `t-ratio`, in R `t value`.

Thus, the rule of thumb says – if the modulus of t exceeds 2, reject H_0, or, in other words, (with the 5% significance) conclude that the variable X is *significant* , i.e., the coefficient β_1 differs from 0.

In our `inc-exp` example `t-ratio` equals 4.877(=10.2096/2.09326) thus it definitely exceeds 2 and therefore `exp` depends on `inc`.

Equivalent approach is based on the concept of p - value. In testing a two-sided hypothesis, calculate $P(|T_{N-2}|>|t|)$ - if the probability is less than 0.05, reject H_0 and accept $H_1^{(1)}$, i.e., conclude that X is significant. In our example, the $p-$ value $P(|T_{40-2}|>4.877) =$ `2*pt(-4.877, df=38)` equals 1.948168e-05, exactly as given in the regression output table. Thus, there is no ground to remove `inc` from the model.

The third possibility to test $H_0: \beta_1 = 0$ with the alternative $H_1: \beta_1 \neq 0$ is to create the two-sided ~95% confidence interval $\left(\hat{\beta}_1 - 2\, s.e.\, \hat{\beta}_1,\ \hat{\beta}_1 + 2\, s.e.\, \hat{\beta}_1\right)$: if it covers 0, we do not reject H_0 at ~5% significance (in our `inc-exp` example, 0 does not belong to (10.21 - 2*2.09, 10.21 + 2*2.09), therefore we reject H_0).

A „good“ model should contain only significant variables

Sometimes it is important to test the hypothesis $H_0: \beta_1 = \beta_1^0$. The `t-ratio` statistics then equals $T = \left(\hat{\beta}_1 - \beta_1^0\right) / \sqrt{\widehat{\text{var}}\hat{\beta}_1}$ and one needs to take some extra steps to test H_0 (see Comp Labs).

3.8. Goodness of Fit (R^2)

Now we shall introduce two more parameters of the „goodness“ of a model.

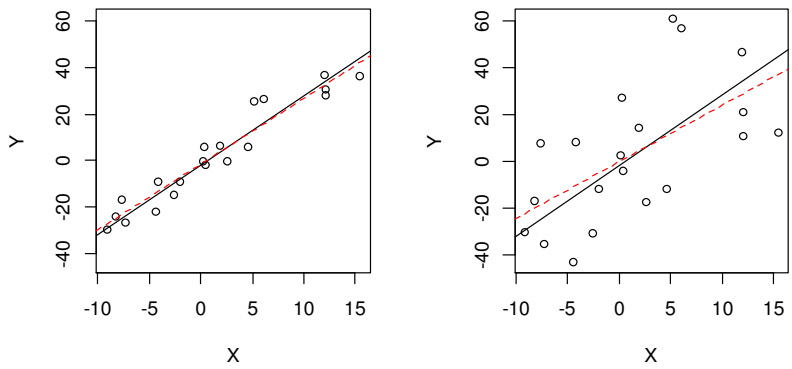


Figure 3.8. `sd=5` (left) and `sd=20` (right); black line is the true regression line $Y = -2 + 3X$ and the red one is the estimated regression line

We repeat the code in p. 3-4, first with sd=$\sigma_\varepsilon^2=5$ (Fig. 3.8, left) and then with sd=$\sigma_\varepsilon^2=20$ (right; note the greater (compared to the left graph) spread of the sample points around the regression line). The accuracy of the estimators $\hat{\beta}_i$ in the second variant is worse[20] (see the estimates of the coefficients of X and Std.errors in the output tables below).

```
   1) Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.5173     1.3691  -1.108    0.282
X             2.8500     0.1862  15.308 9.17e-12 ***

Residual standard error:6.018 on 18 degrees of freedom 6.018 is close to 5
Multiple R-squared: 0.9287,     Adjusted R-squared: 0.9247

   2) Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.06934    5.47652  -0.013  0.99004
X            2.39995    0.74470   3.223  0.00472 **  2.39 is further from 3
                                                     than 2.8500

Residual standard error:24.07 on 18 degrees of freedom 24.07 is close to 20
Multiple R-squared: 0.3659,     Adjusted R-squared: 0.3307
```

The *standard error of regression* or *residual standard error* is much bigger in the second case (24.07 compared to 6.018). To calculate it, we can use (3.4) and the following code:

```
YX.lm=lm(Y~X)
sqrt(sum(YX.lm$res^2)/YX.lm$df.res)
[1] 24.07187
```

Clearly, if the errors are big (i.e., their standard deviation σ_ε is big), the forecast of Y for a new value of X_0 (that is, forecasting with $\hat{Y}_0=\hat{\beta}_0+\hat{\beta}_1X_0$) will be less accurate. It can be shown that standard deviation of the error of the forecast ($se(f)$ in Fig. 3.9) is proportional to $\hat{\sigma}_\varepsilon$ and is given by the formula

$$\sqrt{\widehat{\mathrm{var}}(\hat{Y}_0-Y_0)}=\hat{\sigma}_\varepsilon\sqrt{1+\frac{1}{N}+\frac{(X_0-\bar{X})^2}{\sum(X_i-\bar{X})^2}},$$

thus 1) it depends on $|X_0-\bar{X}|$ and 2) the (approximate) 95% confidence *interval prediction* of Y_0 will be

$$\hat{\beta}_0+\hat{\beta}_1X_0\pm 2\hat{\sigma}_\varepsilon\sqrt{1+\frac{1}{N}+\frac{(X_0-\bar{X})^2}{\sum(X_i-\bar{X})^2}}$$

(the *point prediction* of Y_0 is $\hat{Y}_0=\hat{\beta}_0+\hat{\beta}_1X_0$, i.e., $(X_0,\hat{Y}_0)$ is a point on the estimated regression line).

[20] That is, the concrete estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ can be further from $(\beta_0=)-2$ and $(\beta_1=)3$.

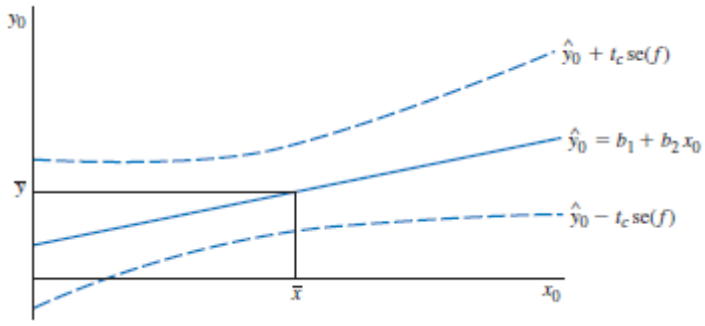


Figure 3.9. The point and interval prediction in regression model

The estimate $\hat{\sigma}_\varepsilon$ is a measure of absolute accuracy of the model, in fact we would like to compare it with the values of Y. One of the measures of overall accuracy of the model is the correlation between Y and $\hat{Y}$ ($\mathrm{cov}(U,V)$ is hardly a measure of dependence between U and V, but $\mathrm{cor}(U,V) = \mathrm{cov}(U,V)/\sqrt{\mathrm{var}\,U\,\mathrm{var}\,V}$ is). Still better measures are the modulus or square of $\widehat{cor}(Y,\hat{Y})$ (both will always be non-negative).

The number $\widehat{\mathrm{cor}}^2(Y,\hat{Y})$ (can you write it explicitely?) is called the *coefficient of determination* (of the model) and denoted by R^2

Assuming that U1-U3 holds and X increases from X_0 to $X_0 + h$, we shall estimate respective change in Y: $Y(X_0) - Y(X_0 + h) = \beta_1 h + \varepsilon_{X_0} - \varepsilon_{X_0+h}$ (the difference does not depend on X_0!). We do not know ε, thus we agree to estimate only the expected change in Y: $E(\beta_1 h + \varepsilon_{X_0} - \varepsilon_{X_0+h} \mid \vec{X}) = \beta_1 h$. The point estimate of this increment is $\hat{\beta}_1 h$, but we can also find the interval estimate: the $\alpha-$confidence interval is given by $((\hat{\beta}_1 - t_{N-1}(\alpha)\cdot s.e.\hat{\beta}_1)\cdot h, (\hat{\beta}_1 + t_{N-1}(\alpha)\cdot s.e.\hat{\beta}_1)\cdot h)$ where $t_{N-1}(0.95) \approx 2$ (the $t_{N-1}(0.99) = \texttt{qt}(0.995, N\text{-}1)$ depends on N and, for example, for $N = 20$ equals 2.8 and for $N = 100$ equals 2.6). ◀◀

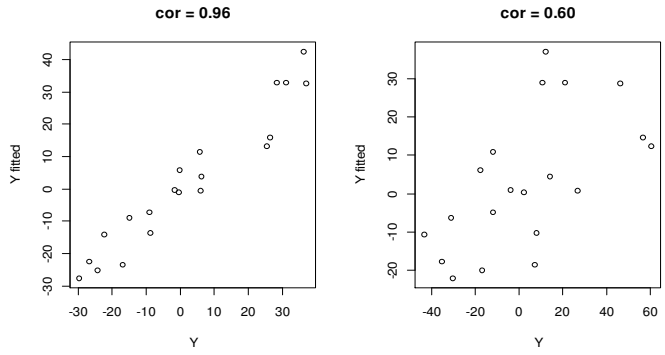


Figure 3.10. In the ideal case, $\hat{Y} = Y$, i.e., $cor(\hat{Y}, Y) = +1$; the correlation between Y and $\hat{Y}$ is stronger in `sd=5` case (left) than in `sd=20` case (right)

Usually R^2 is introduced in a different (but equivalent) way which allows more convenient interpretation of it. Denote by TSS the total sum of squares: $TSS = \sum (Y_i - \bar{Y})^2$. We have

$$\sum (Y_i - \bar{Y})^2 = \sum \left((Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}) \right)^2 = \sum (Y_i - \hat{Y}_i)^2 + 2\sum (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) + \sum (\hat{Y}_i - \bar{Y})^2.$$

It can be shown that the cross-product term $\sum (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y})$ is always 0, therefore[21]

$$TSS = RSS + ESS \text{ or } 1 = \frac{RSS}{TSS} + \frac{ESS}{TSS}.$$

It can be demonstrated that the ratio $\frac{ESS}{TSS}$ or $1 - \frac{RSS}{TSS}$ equals R^2 and it is again called the coefficient of determination

To get a feeling of what is the meaning of the coefficient of determination assume that

1. $R^2 = 0$. Then $\hat{Y}_i \equiv \bar{Y}$ or $\hat{Y}_i = \hat{\beta}_0 + 0 \cdot X_i$, or X has no explanatory power. In other words, all the fitted values, whatever is X, equal $\bar{Y}$.
2. Now assume that $R^2 = 1$. Then $Y_i = \hat{Y}_i$ which means that all the points of the scatter diagram are exactly on the regression line (prediction is ideal).

Thus $0 \le R^2 \le 1$ and the more is the better. If, for example, $R^2 = 0.17$, we shall say that X explains 17% of Y's variability. In our `inc-exp` example $R^2 = 0.385$ which means that 38.5% of the variation in `exp` is explained by `inc` (the rest is explained by unobservable ε). With cross-sectional data, R^2 values from 0.1 to 0.4 are very common.

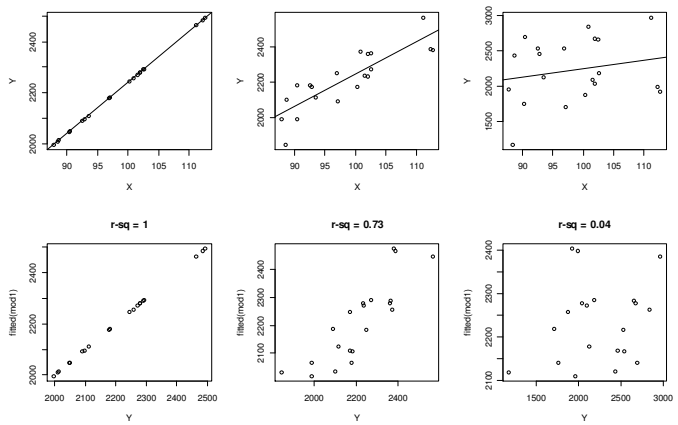


Figure 3.11. Explaining Y variation by X: all variation explained (left), most variation explained (centre), and little variation explained (right); in the upper row, the scatter diagrams are presented (they are of little value in the multivariate case) and in the lower $\hat{Y}$ vs Y.

[21] RSS=Residual Sum of Squares, ESS=Explained Sum of Squares

3.9. Choosing a Functional Form

In our `inc-exp` example we have assumed that the mean household food expenditure linearly depends on household income: $Y = \beta_0 + \beta_1 X$. However, in economic classes you have probably heard that, as income rises, we expect expenditures to increase at a decreasing rate. One candidate to such a behaviour is a linear-log function $Y = \beta_0 + \beta_1 \log X$.

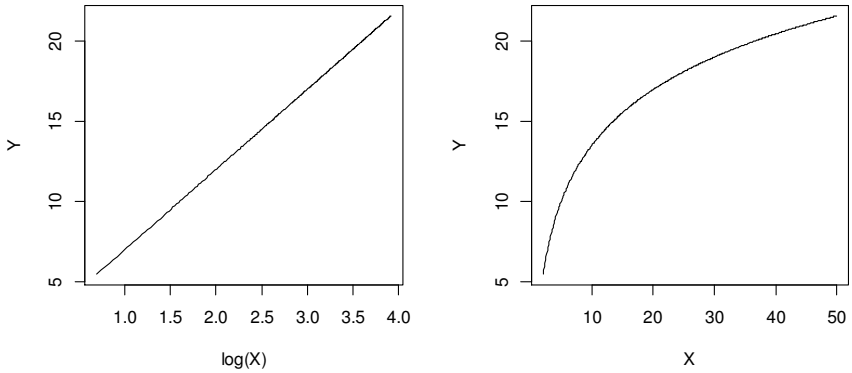


Figure 3.12. The graph of the function Y=2+5*log(X) in (log(X),Y) scale (left) and in (X,Y) scale (right)

For any curve $Y = f(X)$, its behaviour at x can be characterized by its derivative $Df(x) = \lim_{\Delta x \to 0} \frac{\Delta f(x)}{\Delta x} = \frac{df}{dX}(x)$ (which means the rate of increase at point x or the *marginal effect* of a change in the explanatory variable at x or the slope of the tangent to the curve at the point x) – in linear case it equals the constant β_1 and in linear-log case β_1 / X (thus the marginal effect diminishes with increasing X). In economics, even more popular characteristic of the rate of change is *elasticity* defined as $Ef(x) = \lim_{\Delta x \to 0} \frac{\Delta f(x) / f(x)}{\Delta x / x} = \frac{df}{dX}(x) \cdot \frac{x}{f(x)}$ which should be read as the *percentage change* in Y corresponding to 1% increase of X. Thus, in linear case elasticity equals $\beta_1 \cdot \frac{X}{\beta_0 + \beta_1 X}$ and in linear-log case $\frac{\beta_1}{X} \cdot \frac{X}{\beta_0 + \beta_1 X} = \frac{\beta_1}{\beta_0 + \beta_1 X}$.

Two more functions useful in regression are log-linear function $\log(Y) = \beta_0 + \beta_1 X$ or $Y = B_0 \exp(\beta_1 X)$ and log-log function $\log(Y) = \beta_0 + \beta_1 \log(X)$ or $Y = B_0 X^{\beta_1}$. The latter function is remarkable in the sense that its elasticity is constant: $Ef(X) = B_0 \beta_1 X^{\beta_1 - 1} \cdot (X / B_0 X^{\beta_1}) = \beta_1$. Note the meaning of the coefficient β_1 in all these four cases:

1. linear – a 1 unit change in X leads to a β_1 unit change in Y

2. linear-log – a 1% change in X leads to an (approximately) $\beta_1/100$ unit change in Y
3. log-linear – a 1 unit change in X leads to an (approximately) $100\,\beta_1$ % change in Y
4. log-log – a 1% change in X leads to an (approximately) β_1 % change in Y

As an example, we shall prove the 3*rd* proposition:

a) Let Z be any quantity; then the expression $(Z_{new} - Z_{old} =)\, Z_n - Z_o$ is called the *difference* or *change*.
b) The expression $(Z_n - Z_o)/Z_o$ is called a *relative change* and $(Z_n - Z_o)/Z_o * 100\%$ a *percentage change*.
c) If a relative change is „small", it can also be given as $\log Z_n - \log Z_o$. Indeed, $\log(Z_n/Z_o) = \log(1 + (Z_n - Z_o)/Z_o) \approx (Z_n - Z_o)/Z_o$ by the Taylor formula.
d) Let $X_o = X$ increases to $X_n = X + 1$; then $\log Y_n - \log Y_o = \beta_1$. ◀◀

Let us return to our four models: which of the four models describing the `inc-exp` relationship is the best? As the left-hand-side variable in cases 1 and 2 is the same, namely Y, we can use R^2 to choose between these two. However,

Do not evaluate the quality of the model based only on R^2. To evaluate the model it is important to consider factors such as its adequacy to economic theory, the correct signs and magnitudes of the estimated coefficients, their statistical and economic significance, the precision of their estimation, and the ability of the fitted model to predict values of the dependent variable that were not in the estimation sample.

In any case, we can start with R^2. Respective R code informs that the R^2 of the linear-log model 2 is only slightly less than that of linear model 1, but model 2 better reflects economic theory, therefore we choose it (model 2 is $\widehat{exp} = -97.19 + 132.17 \log(inc)$ and says that 1% increase in income will increase food expenditure by approximately \$1.32 per week or that 10% increase in income will increase food expenditure by approximately \$13.22 per week).

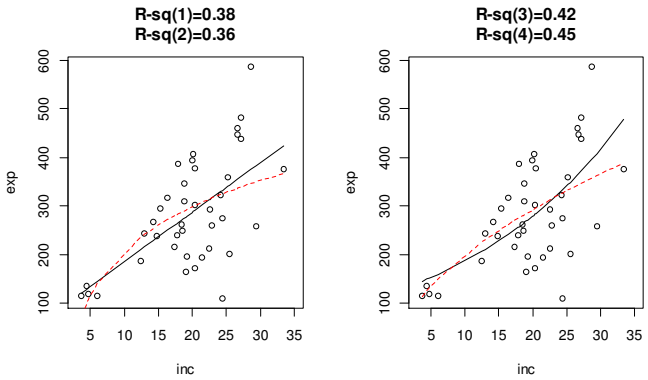


Figure 3.13. `inc-exp` models 1 and 2 (left; which one is model 2?) and models 3 and 4 (red) (right)

As we have already said, a model should not be chosen only on the basis of model fit with R^2 as the criterion. Here we append the above list of recommendations:

1. Choose a shape that is consistent with what economic theory tells us about the relationship.
2. Choose a shape that is sufficiently flexible to „fit" the data.
3. Choose a shape so that assumptions U1-U4 are satisfied (that is, test residuals for normality and homoskedasticity).

We shall return to issue #3 a bit later, but now let us compare models 3 and 4 – since their left hand sides are the same, namely $\log(Y)$, we can use R^2 and conclude that model 4 is more accurate (it also better reflects theory). However, these two groups (models1,2 and models 3,4) cannot be compared by R^2 because their lhs's are different (and thus TSS are different). To compare the best models 2 and 4, we shall rescale model 4 to the Y, instead of $\log(Y)$, axis. The log-log model $\log(Y)=\beta_0+\beta_1\log(X)+\varepsilon$ can be rewritten as $Y=\exp(\beta_0+\beta_1\log(X))\cdot\exp(\varepsilon)$, therefore we predict Y as $\hat{Y}_{(4)}=\exp(\hat{\beta}_0+\hat{\beta}_1X)\cdot E\exp(N(0,\hat{\sigma}_\varepsilon^2))=\exp(\widehat{\log Y})\cdot\exp(\hat{\sigma}_\varepsilon^2/2)$. Now $R^2_{(4)}=\widehat{cor}^{\,2}(\mathtt{exp},\widehat{\mathtt{exp}_{(4)}})=0.38$, thus model 4 is marginally better than model 2[22].

3.10. Does our model satisfy U3 and U4?

Recall that OLS estimators are BLUE&C if U1-U4 are satisfied. What if some of them, specifically, U3 and U4, are not satisfied?

U3. Let us start with U3: $\operatorname{var}\varepsilon_i\equiv\sigma_\varepsilon^2$, i.e., the variance of errors does not depend on i or, maybe, on any other Z_i (the model satisfying this condition is called *homoskedastic*, otherwise it is *heteroskedastic*). If U3 is violated, the OLS estimators of the coefficients are still unbiased, consistent and the interpretation of R^2 is unaffected. However, the OLS estimator $\hat{\beta}_1$ has no longer minimum variance and also the OLS estimator of the variance, $\widehat{\operatorname{var}}\hat{\beta}_1^{OLS}$, is biased and thus the t- ratios and p- values are unreliable even in large samples.

We propose two solutions to heteroskedasticity problem.

1. Recall that the OLS estimator of $\operatorname{var}\hat{\beta}_1$ (under assumption $\operatorname{var}\varepsilon_i\equiv\sigma_\varepsilon^2$) is (see (3.6)) $\widehat{\operatorname{var}}\hat{\beta}_1=\left(1/\sum x_i^2\right)^2\sum x_i^2\,\widehat{\operatorname{var}}\varepsilon_i=\left(1/\sum x_i^2\right)^2\sum x_i^2\cdot\left(\sum e_i^2/(N-2)\right)$. If the variances of errors are not equal, use another formula: $\widehat{\operatorname{var}}\hat{\beta}_1=\dfrac{N}{N-2}\dfrac{\sum x_i^2e_i^2}{\left(\sum x_i^2\right)^2}$. It is called White's heteroskedasticity-consistent standard errors or heteroskedasticity robust standard errors or simply robust

[22] Another opportunity to compare models 2 and 4 is to use the Box-Cox transformation described in [AST, p. 166].

standard errors and it gives a consistent variance estimator. Note that we change neither $\hat{\beta}_1$ nor residuals, we only correctly the estimate $\widehat{\text{var}}\hat{\beta}_1$.

To perform the procedure in GRETL, one has just to check respective box. Import the `food` data set from POE, add logarithms of the two variables and go to Model→Ordinary Least squares, fill in variable windows and uncheck „Robust standard errors“:

```
Dependent variable: l_food_exp

            coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       3.96357      0.294373    13.46     4.84e-016 ***
  l_income    0.555881     0.100660     5.522    2.57e-06  ***
```

It seems that the errors of our model are heteroskedastic (see Fig. 3.14). Therefore we shall repeat the same procedure with „Robust standard errors“ checked:

```
Dependent variable: l_food_exp
Heteroskedasticity-robust standard errors, variant HC1

            coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       3.96357      0.163363    24.26     9.87e-025 ***
  l_income    0.555881     0.0666362    8.342    4.08e-010 ***
```

The differences in estimations are colored in yellow (in both cases p - values are much less than 0.05, therefore our conclusion concerning the significance of `l_income` is the same).

2. The OLS β_1 estimate in **1**. is not efficient. To get an efficient estimator, we have to use another, *weighted least squares* or WLS procedure. Assume that in $Y = \beta_0 + \beta_1 X + \varepsilon$, the standard deviation $\sigma_i = \sqrt{\text{var}\,\varepsilon_i}$ is not a constant but the function of a certain variable, for example $\sigma_i = f(X_i)$; most often $\sigma_i = cX_i$ or $\sigma_i = c\sqrt{X_i}$. Assume that $\sigma_i = cX_i$ and divide the ith equation by X_i (the numbers $1/\sigma_i^2 = (1/f(X_i))^2$ $(= 1/X_i^2)$ are called *weights*):

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$
$$Y_i / X_i = \beta_0 / X_i + \beta_1 + \varepsilon_i / X_i$$
$$Y_i^* = \beta_1 + \beta_0 X_i^* + \varepsilon_i^*$$

Since $\text{var}\,\varepsilon_i^*$ now is constant (thus, this new model is homoskedastic), we can safely apply OLS to estimate all the parameters and then return to the original equation. Note that in WLS we minimize $RSS(b_0, b_1) = \sum \left(1/f(X_i)\right)^2 (Y_i - (b_0 + b_1 X_i))^2$.

The usual problem is to find right weights (we shall use graphs and relevant tests to this end).

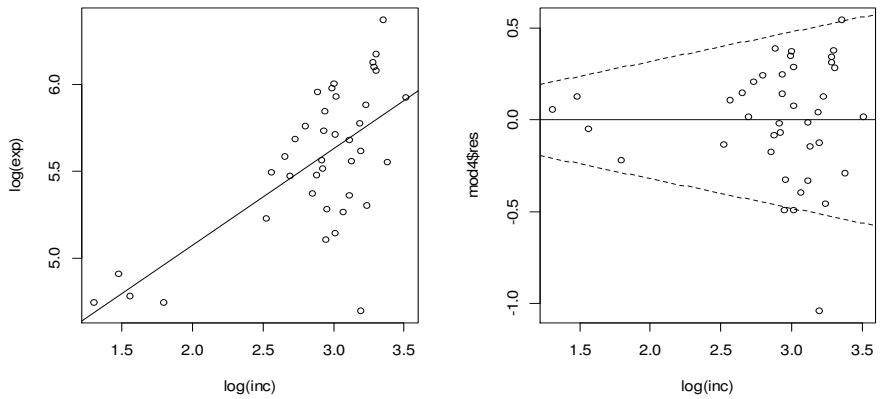


Figure 3.14. `log(inc)-log(exp)` scatter diagram together with a regression line (left) and residuals of the log-log model (right)

In Fig. 3.14 (right) one can see that the spread of the residuals around their mean ($\equiv 0$) is increasing with `log(inc)`, thus, probably, σ_ε is proportional to it. To test the null homoskedasticity hypothesis $H_0 : \sigma_i^2 \equiv c_0$ with alternative $H_1 : \sigma_i^2 = c_1 \log^2(inc)$, we shall use the *Breusch-Pagan test*:

1. Estimate the model $\log(\text{exp}) = \beta_0 + \beta_1 \log(inc) + \varepsilon$ with OLS.
2. Run the regression $\hat{\varepsilon}_i^2 = \gamma_0 + \gamma_1 \log^2(inc_i) + u_i$ and test the hypothesis $H_0 : \gamma_1 = 0$ (if we reject it, errors are heteroskedastic and we have to use WLS with the $1/\log^2(inc)$ weight).

Note that instead of looking for the $p-$value of γ_1 in the printout of the second regression, we have to go, in the first model window, to Tests| Heteroskedasticity| Breusch-Pagan – the answer there is

```
Test statistic: LM = 3.761727,
with p-value = P(Chi-square(1) > 3.761727) = 0.052438
```

which means that we should not care much about heteroskedasticity. Nevertheless, in order to demonstrate the use of WLS, go to Model| Other linear models| Weighted Least Squares...| (create in advance the weight variable with `series ww = 1/log(income)^2` and) fill in the Weight variable window with `ww` etc:

```
Dependent variable: l_food_exp
Variable used as weight: ww

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       3.99897      0.159356     25.09     2.95e-025 ***
  l_income    0.543296     0.0604561     8.987    6.10e-011 ***

Statistics based on the weighted data:

Sum squared resid    0.419384   S.E. of regression    0.105054
R-squared            0.680025   Adjusted R-squared    0.671605
```

```
Schwarz criterion   -61.42103   Hannan-Quinn        -63.57750

Statistics based on the original data:

Mean dependent var   5.565019   S.D. dependent var   0.424068
Sum squared resid    3.892513   S.E. of regression   0.320054
```

Note a slightly smaller `std.error` (more efficient estimator) now.

In general, it is recommended to take care of heteroskedasticity only in severe cases. We shall also extent the analysis of this problem in the next chapter. [redacted]

U3. The errors are called uncorrelated if $\text{cov}(\varepsilon_i, \varepsilon_j) = E\varepsilon_i\varepsilon_j \equiv 0$ for all $i \neq j$. This condition is often violated in time series context. Specifically, assume that

$$\begin{cases} Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t & (3.10a) \\ \varepsilon_t = \rho\varepsilon_{t-1} + u_t, \ |\rho| < 1 & (3.10b) \end{cases}$$

(we say that the model has *autocorrelated* or *serially correlated* errors). Note that now yesterday's error ε_{t-1} contains some information about today's error ε_t: $\text{cov}(\varepsilon_t, \varepsilon_{t-1}) = \text{cov}(\rho\varepsilon_{t-1} + u_t, \varepsilon_{t-1}) = \rho \,\text{var}\,\varepsilon_{t-1} \neq 0$, therefore this information can be used in regression. Indeed, subtracting $\rho Y_{t-1} = \rho\beta_0 + \rho\beta_1 X_{t-1} + \rho\varepsilon_{t-1}$ from the first equation we get a multivariate model

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \alpha_2 X_t + \alpha_3 X_{t-1} + u_t$$

with Y's and X's lags in rhs and uncorrelated errors u_t. We shall analyze the model in the next chapter, here we mention only that estimating (3.10a) with OLS and ignoring violation of U3, we get unbiased and consistent estimators $\hat{\beta}_i^{OLS}$ (and forecasts based on them). However, the estimator is inefficient and the variances of the regression coefficients will be biased.

The (pseudo) autocorrelation in errors is often the result of the misspecification of the model. For example, in GRETL go to File| Open data| Sample file...| Ramanathan| data6-6 (the data set contains the percent of the total US population that lives on farms).

The model $farmpop = \beta_0 + \beta_1 year + \varepsilon$ is not very succesful (compare the actual and fitted graphs of `farmpop`, Fig. 3.15, left; see also the graph of the residuals (Fig. 3.15, right - inertia or persistency of the residual curve is the first sign of autocorrelation; as the first formal autocorrelation test we shall use the DW or Durbin-Watson statistics – it is always between 0 and 4; if it is not close to 2, errors are autocorrelated). Note that here the `std.error` was estimated with HAC = Heteroskedasticity and Autocorrelation Consistent procedure.

```
Dependent variable: farmpop
HAC standard errors, bandwidth 2 (Bartlett kernel)

            coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const      646.257       64.0368      10.09     8.53e-013 ***
  year        -0.324848     0.0324780   -10.00     1.11e-012 ***
  rho          0.944462      Durbin-Watson          0.055649
```

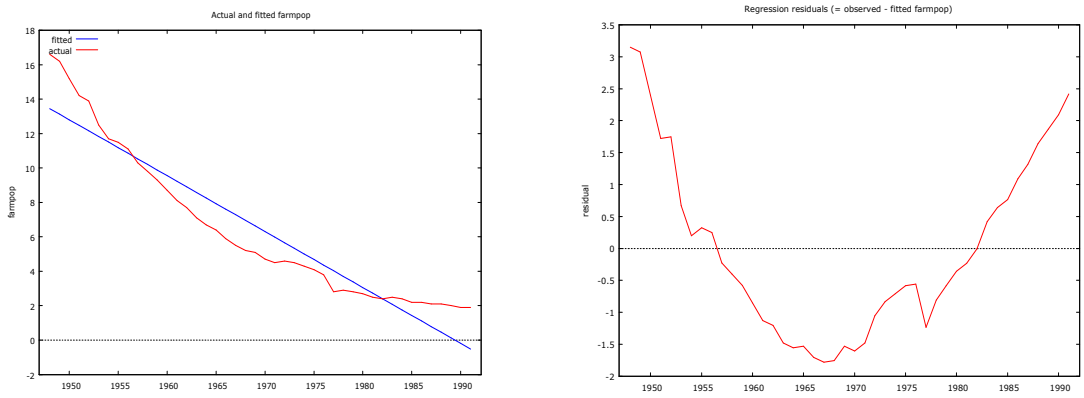


Figure 3.15. Graphs of `farmpop` and inadequate regression line (left); persistent residuals of the linear model (right)

Most probable explanation of the observed „autocorrelation“ is the wrong specification of the model. Let us try a quadratic model $farmpop = \beta_0 + \beta_1 year + \beta_2 year^2 + \varepsilon$ (Durbin-Watson statistics is still far from 2, the graph of the residuals (Fig. 3.16, right) still demonstrates persistency).

```
Dependent variable: farmpop
HAC standard errors, bandwidth 2 (Bartlett kernel)

             coefficient       std. error     t-ratio     p-value
  ------------------------------------------------------------------
  const      37165.9           2125.09          17.49     1.28e-020 ***
  year         -37.4115           2.15670       -17.35     1.72e-020 ***
  sq_year        0.00941526       0.000547170    17.21     2.31e-020 ***
  rho            0.695629         Durbin-Watson             0.601455
```

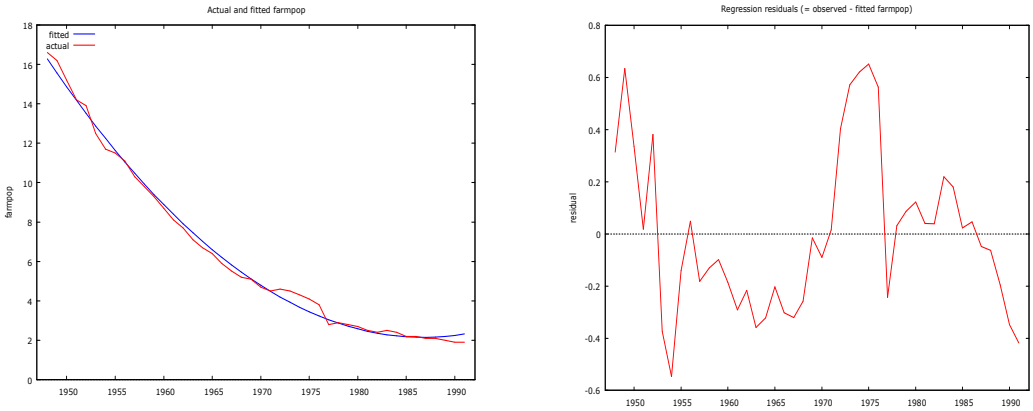


Figure 3.16. The model with a quadratic trend (note the „wrong“ behavior of the blue parabola around 1990, left) and its persistent residuals (right); the left model suggests to explore the exponential trend $farmpop = \beta_0 + \beta_1 \exp(\beta_2 year) + \varepsilon$, but this is an example of a rather complicated nonlinear regression model.

We shall respecify the model once again: $ld\_farmpop_t (= \log(farmpop_t) - \log(farmpop_{t-1}))$ $= \beta_0 + \beta_1 year_t + \varepsilon_t$ (the meaning of the lhs is the percentage growth of `farmpop` in year t).

```
Model 3: OLS, using observations 1949-1991 (T = 43)
Dependent variable: ld_farmpop

             coefficient    std. error    t-ratio   p-value
  ---------------------------------------------------------
  const      -1.19588       1.24153      -0.9632    0.3411
  year        0.000581457   0.000630206   0.9226    0.3616

rho                -0.148626   Durbin-Watson        2.266147
```

It seems that autocorrelation has gone but now the model has insignificant variable `year`, therefore the model can still be improved (remove `year` (then the percentage growth will be the same for all the years, it is `Model 4`); to forecast $\log(\text{farmpop}_t)$, use the formula $\widehat{\log(\text{farmpop}_t)} = \log(\text{farmpop}_{t-1}) + forecast\_of(\text{ld\_farmpop}_t)$).

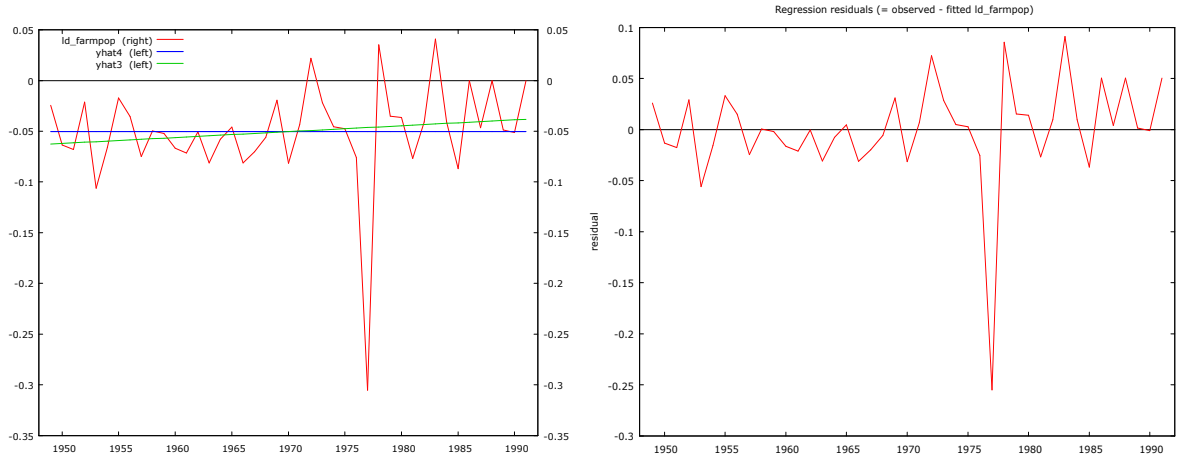
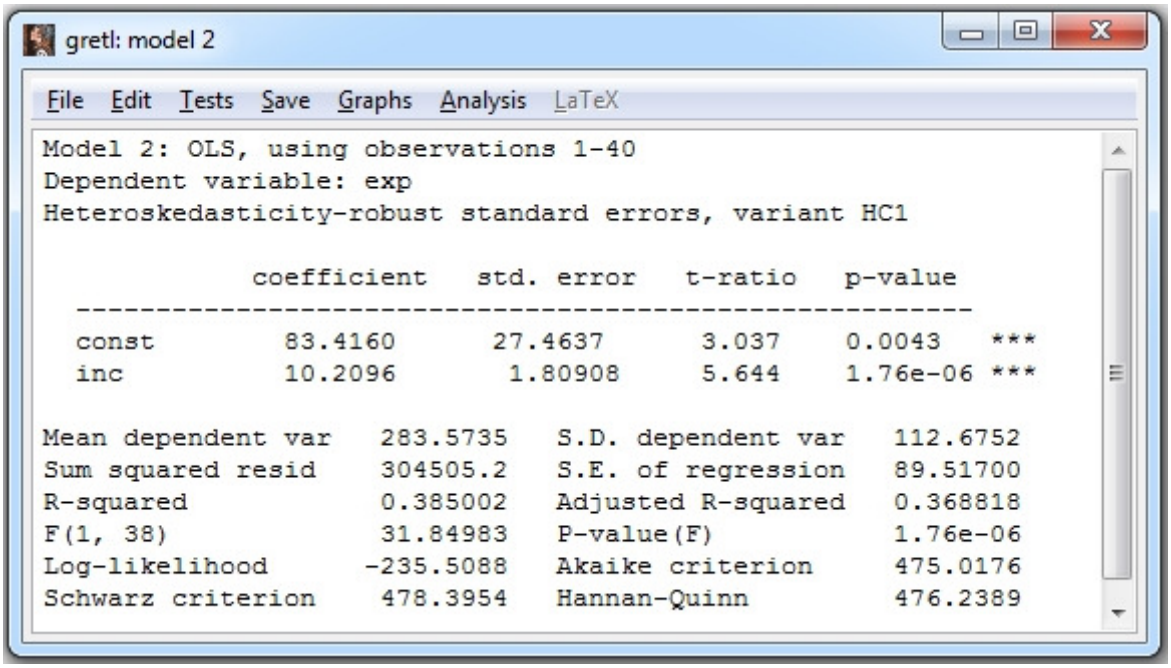


Figure 3.17. The graph of `ld_farmpop` and `yhat`'s of two growth models (left); uncorrelated residuals of `Model 4` (right)

U4. The requirement of normality, $\varepsilon \sim N(0, \sigma_\varepsilon^2)$, is necessary if one wants to obtain correct p - values when testing $H_0 : \beta_i = \beta_i^{(0)}$. We shall apply the χ^2 goodness-of-fit test to validate normality (we do not know ε, therefore we use their estimates $\hat{\varepsilon}$ which are hopefully close to ε). In GRETL, upon creating the model



go to Tests| Normality of residual – in the Fig. 3.16 one can see the bell-shaped histogram together with the p - value of the H_0 : *errors have normal distribution* ; it equals 0.7068, therefore we have no ground to reject H_0. Note that if the errors do not depart much from nor-

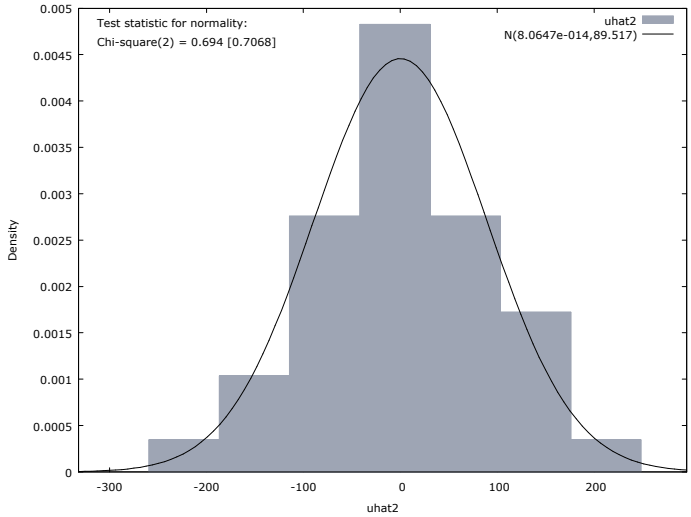


Figure 3.18. Histogram of residuals and the p - value of normality test

mality, in large samples, the p - values presented in the 4*th* column will be close to the correct ones.

3.11. Nonlinear regression

The word *nonlinear* is a bit confusing – here we mean a regression nonlinear in parameters, e.g., exponential $Y=\beta_0+\beta_1\exp(\beta_2 X)+\varepsilon$ or power $Y=\beta_0+\beta_1 X^{\beta_2}+\varepsilon$ (it is close to $Y=B_0X^{\beta_1}e^{\varepsilon}$ which is equivallent to the linear model $\log(Y)=\beta_0+\beta_1\log(X)+\varepsilon$) or the logistic regression $Y_t=\beta_0/\left(1+\exp(-\beta_2(t-\beta_3))\right)+\varepsilon_t$. The estimation procedure is similar to that of OLS, but now we may have problems with minimizing RSS. In the first example of exponential regression we get a system of nonlinear equations

$$\begin{cases}\dfrac{\partial RSS}{\partial b_0}(b_0,b_1,b_2)=\sum(-2)\left(Y_i-(b_0+b_1\exp(b_2X_i))\right)=0\\ \dfrac{\partial RSS}{\partial b_1}(b_0,b_1,b_2)=\sum(-2)\left(Y_i-(b_0+b_1\exp(b_2X_i))\right)\cdot\exp(b_2X_i)=0\\ \dfrac{\partial RSS}{\partial b_2}(b_0,b_1,b_2)=\sum(-2)\left(Y_i-(b_0+b_1\exp(b_2X_i))\right)\cdot b_1\exp(b_2X_i)\cdot X_i=0\end{cases}$$

which can be solved only by the use of iterative numeric methods (the common problem is to choose a good zero*th* iteration).

We shall present a short explanation on how the iterative method works. Assume for simplicity that we have only two parameters: $Y_i=f(X_i,b_0,b_1)+\varepsilon_i$, $b_0^{(0)}$ and $b_1^{(0)}$ (the zero*th* iteration) are numbers close to the minimizing ones and

$$RSS(b_0,b_1)\left(=\sum(Y_i-f(X_i,b_0,b_1))^2\right)\approx RSS(b_0^{(0)},b_1^{(0)})+$$
$$(b_0-b_0^{(0)})\frac{\partial RSS}{\partial b_0}(b_0^{(0)},b_1^{(0)})+(b_1-b_1^{(0)})\frac{\partial RSS}{\partial b_1}(b_0^{(0)},b_1^{(0)})$$
$$\frac{(b_0-b_0^{(0)})^2}{2!}\frac{\partial^2 RSS}{\partial b_0^2}(b_0^{(0)},b_1^{(0)})+\frac{(b_0-b_0^{(0)})(b_1-b_1^{(0)})}{1!1!}\frac{\partial^2 RSS}{\partial b_0\partial b_1}(b_0^{(0)},b_1^{(0)})+\frac{(b_1-b_1^{(0)})^2}{2!}\frac{\partial^2 RSS}{\partial b_1^2}(b_0^{(0)},b_1^{(0)})$$

(we approximate the surface $z=RSS(b_0,b_1)$ by a quadratic one given by its Taylor expansion). To minimize RSS, calculate partial derivatives of this quadratic expression and equate them to zero. It is convenient to rewrite the system in matrix notation:

$$\begin{pmatrix}\dfrac{\partial^2 RSS}{\partial b_0^2} & \dfrac{\partial^2 RSS}{\partial b_0\partial b_1}\\ \dfrac{\partial^2 RSS}{\partial b_1\partial b_0} & \dfrac{\partial^2 RSS}{\partial b_1^2}\end{pmatrix}\begin{pmatrix}b_0-b_0^{(0)}\\ b_1-b_1^{(0)}\end{pmatrix}=-\begin{pmatrix}\dfrac{\partial RSS}{\partial b_0}\\ \dfrac{\partial RSS}{\partial b_1}\end{pmatrix}$$

(all the derivatives are estimated at the point $\left(b_0^{(0)}, b_1^{(0)}\right)$). To simplify the above expression, rewrite it as $M_2 S = -M_1$ or $S = -M_2^{-1} M_1$, i.e., $\begin{pmatrix} b_0^{(1)} \\ b_1^{(1)} \end{pmatrix} = \begin{pmatrix} b_0^{(0)} \\ b_1^{(0)} \end{pmatrix} - M_2^{-1} M_1$. We repeat the procedure at point $\begin{pmatrix} b_0^{(1)} \\ b_1^{(1)} \end{pmatrix}$ etc until $\begin{pmatrix} b_0^{(n)} \\ b_1^{(n)} \end{pmatrix}$ (almost) stops to change – this will be a solution to our RSS minimizing problem. More subtle variants of this gradient descent procedure are implemented in the `nls` function present in both GRETL and R.

As an example, we shall consider the data set USPop from the car package in R. The set contains two variables

`year`	census year (once in 10 years, 1790 through 2000)
`population`	US population in million

We shall create two models, exponential and logistic, based on censuses from 1790 till 1930 and then predict `population` till 2000. The code is in Practicals, both models are not very succesful in the long run.

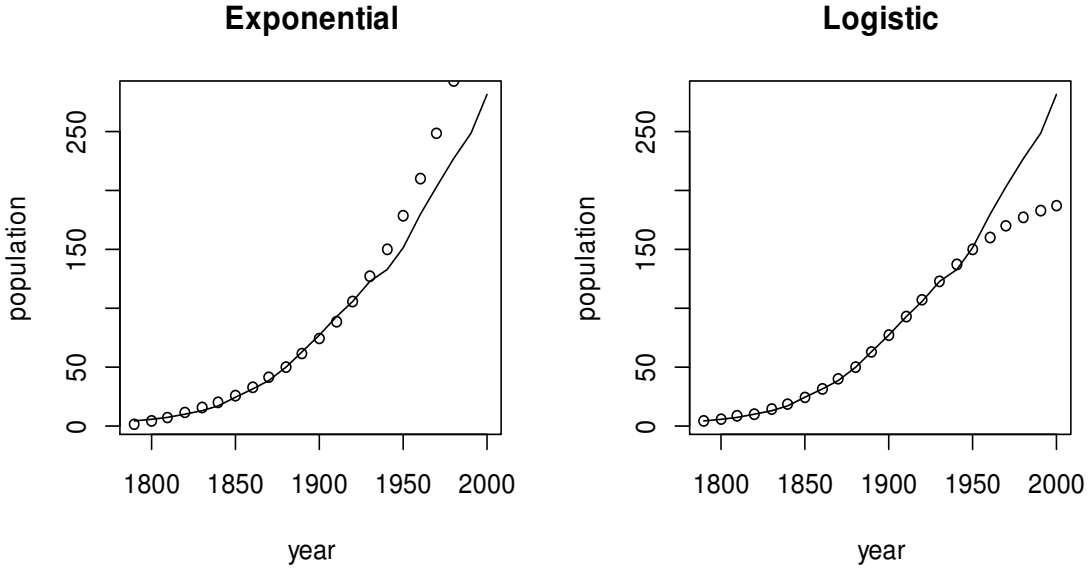


Figure 3.19. The US population graph and two trends based on the forecasts of the nonlinear models for 1790 through 1930

Exponential: $population = -15.292 + 13.897 * 10\,\hat{}\,(-12) * \exp(1.552/100 * year)$

Logistic: $population = 202.1/\left(1 + \exp(-0.031 * (year - 1916))\right)$

Note that in the time series setting, if the predictive variable is time, the regression curve is called a *trend*.

4. Multivariate Regression

It is rarely the case that economic relationships involve just two variables. Rather, a dependent variable Y can depend on a whole series of explanatory variables or regressors. For example, the demand for a good does not just depend on its price but also on the prices of close substitutes or complements, the general level of prices and on the resouces of consumers. Thus in practice we are normally faced with relationships of the form

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \varepsilon \tag{4.1a}$$

where Xs stand for causes and Y for the effect. The sample produced by this DGP can be written as a system of equations

$$\begin{cases} Y_1 = \beta_0 + \beta_1 X_{11} + ... + \beta_k X_{k1} + \varepsilon_1 \\ ..... \\ Y_N = \beta_0 + \beta_1 X_{1N} + ... + \beta_k X_{kN} + \varepsilon_N \end{cases}$$

or

$$\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\varepsilon} \tag{4.1b}$$

where

$$\vec{Y} = \begin{pmatrix} Y_1 \\ ... \\ Y_N \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1\, X_{11} \,...\, X_{k1} \\ ..... \\ 1\, X_{1N} ... \, X_{kN} \end{pmatrix}, \vec{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ .... \\ \beta_k \end{pmatrix}, \vec{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ ... \\ \varepsilon_N \end{pmatrix}.$$

Our observations can be presented by the matrix $\begin{pmatrix} Y_1 \;\; X_{11} \,...\, X_{k1} \\ ..... \\ Y_N \; X_{1N} ... \, X_{kN} \end{pmatrix}$, $\vec{X}_i = (1, X_{1i}, ..., X_{ki})$ denotes the vector of the ith observation of explanatory variables, $\mathbf{X}$ is the *design* matrix, $\vec{\beta}$ the vector of coeffients, and ε the *error* of the DGP.

4.1. Ordinary Least Squares (OLS)

The formula $Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k$ describes (unknown) population regression equation[1], we want to use the sample data to estimate its unknown coefficients. Whatever estimation method we use, the „plane“ $Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + ... + \hat{\beta}_k X_k \; (= \hat{Y})$ is called a sample regression equation or a fit to the scatter diagram and $\hat{Y}$ is a *predicted value* of Y. The difference $Y - \hat{Y}$ is

[1] If $k = 1$, it is a line in (X_1, Y) plane; if $k = 2$ a plane in three-dimensional (X_1, X_2, Y) space, and if $k \geq 3$ a „hyperplane“ in R^{k+1} space.

referred to as a *residual* and denoted by $\hat{\varepsilon}$ or e. The slope coefficient β_1 equals $Y(X_1+1, X_2, ..., X_k) - Y(X_1, X_2, ..., X_k)$ (verify), thus, β_1 reflects the isolated (holding the other variables constant) influence of X_1 on Y (or, as it is often called, influence of X_1 on Y *ceteris paribus*, i.e., when all the rest of X 's do not change).

To estimate the coefficients $\beta_m, m = 0,1,...,k,$ we shall again use the OLS method, but first we formulate conditions under which the OLS estimators will be BLUE&C[2]. Note that in fact the conditions coincide with the univariate case, the only new requirement is M5.

M1. The DGP is described by the linear model $\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\varepsilon}$ where the output random variable (r.v.) Y depends on two r.v.'s, $\vec{X}$ and ε.

The further conditions M2-M4 describe the properties of unobservable ε and its relationship with observable $\vec{X}$.

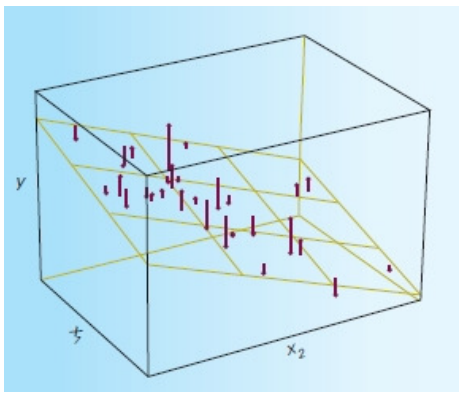
M2. $E(\vec{\varepsilon} \mid \mathbf{X}) = \vec{0}$ (strict exogeneity) – this means that whatever are the observations, the errors on average equal 0: $\int_{R^N} \vec{z}\, f_{\vec{\varepsilon}}(\vec{z} \mid \vec{X}_1, ..., \vec{X}_N) d\vec{z} = \vec{0}$. Note that M2 implies $E(\varepsilon_i \mid \vec{X}_i) \equiv 0$, $E\varepsilon_i \equiv 0$, $E\varepsilon_i \vec{X}_i = \vec{0}$ (ε_i and any $X_{m,i}$ do not correlate) and $E(Y_i \mid \mathbf{X}) = E(Y_i \mid \vec{X}_i) = \vec{X}_i\vec{\beta}$.

M3. $\mathrm{var}(\vec{\varepsilon} \mid \mathbf{X}) = \sigma_\varepsilon^2 \mathbf{I}$ – this means that the conditional distribution of the errors given the matrix of explanatory variables has constant variances and zero covariances.In particular, this means that each error has the same variance and that any two error terms are uncorrelated.

M4. Sometimes a requirement of normality is added: $\vec{\varepsilon} \mid \mathbf{X} \sim N(\vec{0}, \sigma_\varepsilon^2 \mathbf{I})$ which says that conditional density of $\vec{\varepsilon}$ (and, therefore, of any ε_i) is normal (this is equivalent to saying that $\vec{Y} \mid \mathbf{X} \sim N(\mathbf{X}\vec{\beta}, \sigma_\varepsilon^2 \mathbf{I})$ or $Y_i \mid \mathbf{X} \sim N(\vec{X}_i\vec{\beta}, \sigma_\varepsilon^2)$).

The following assumption has no analogue in univariate case:

M5. There exist no exact linear relationship between the sample values of any two or more of the explanatory variables[3] - we mean that it must not be the case that, for example, $X_{3i} = 4 + 2X_{1i}$ for all i (that is, it must not be the case that the fourth column in the design matrix $\mathbf{X}$ is a linear combination of the first and second columns); another way to express this condition is to say that $rank\, \mathbf{X} = k+1$ or $\det \mathbf{X}'\mathbf{X} \neq 0$.



The OLS estimator of the coefficients β_m, $m = 0,1,...,k,$ in multivariate case is defined in the same way as previously: find $b_0, b_1, ..., b_k$ such that the residual sum of

[2] Recall that the BLUE properties hold for any size samples whereas C (=consistency) is a large sample or asymptotic property.
[3] If the condition is not satisfied, the model is called *multicollinear*.

squares $RSS = RSS(b_0, b_1, ..., b_k) = \sum_{i=1}^{N} (Y_i - (b_0 + b_1 X_{1i} + ... + b_k X_{ki}))^2 = \sum \hat{\varepsilon}_i^2$ were as small as possible (the estimated regression plane must be as close to the sample points as possible).

In other words, we have to equate all the partial derivatives to zero and solve the system

$$\begin{cases} \sum (Y_i - (b_0 + b_1 X_{1i} + ... + b_k X_{ki})) = 0 \\ \sum X_{1i} (Y_i - (b_0 + b_1 X_{1i} + ... + b_k X_{ki})) = 0 \\ ..... \\ \sum X_{ki} (Y_i - (b_0 + b_1 X_{1i} + ... + b_k X_{ki})) = 0 \end{cases}$$

or just to use the formula (3.9)

$$\hat{\vec{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \vec{Y} . \qquad (4.2)$$

The expressions of the coefficients are rather complicated, for example, if $k = 2$,

$$\hat{\beta}_1 = \frac{\widehat{\mathrm{cov}}(X_1, Y)\, \widehat{\mathrm{var}} X_1 - \widehat{\mathrm{cov}}(X_2, Y)\, \widehat{\mathrm{cov}}(X_1, X_2)}{\widehat{\mathrm{var}} X_1\, \widehat{\mathrm{var}} X_2 - \left(\widehat{\mathrm{cov}}(X_1, X_2)\right)^2},$$

therefore all statistical programs prefer to use the matrix expression (4.2). Another variant of (4.2) which helps us to memorize the formula is to recall univariate regression and (3.3):

$$\begin{cases} \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \\ \hat{\beta}_1 = \dfrac{\sum (Y_i - \bar{Y})(X_i - \bar{X})}{\sum (X_i - \bar{X})^2} = \dfrac{\sum (Y_i - \bar{Y}) X_i}{\sum (X_i - \bar{X})^2} = \dfrac{\sum (X_i - \bar{X}) Y_i}{\sum (X_i - \bar{X})^2} = \dfrac{\widehat{\mathrm{cov}}(X, Y)}{\widehat{\mathrm{var}} X} \end{cases} .$$

In multivariate case, assuming $\vec{X}_i = (1, X_{1i}, ..., X_{ki})'$, the OLS estimator of $\vec{\beta}$ can be expressed similarly as $\hat{\vec{\beta}} = \left(\sum_{i=1}^{N} \vec{X}_i \vec{X}_i' \right)^{-1} \cdot \sum_{i=1}^{N} \vec{X}_i Y_i$ [4].

In multivariate case, the *Gauss-Markov theorem* is true again:

> Under assumption that the conditions M1-M3 and M5[5] hold true, the OLS estimators $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ are (both conditionally and unconditionally) BLUE&C

As earlier, if the null $H_0 : \beta_m = \beta_m^0$ is true, the t- ratio $(\hat{\beta}_m - \beta_m^0) / s_{\hat{\beta}_m}$ has the $T_{N-(k+1)}$ distribution[6] (note that the number of degrees of freedom is N-(k +1)!) where $s_{\hat{\beta}_m}^2 = s^2 X^{mm}$,

[4] Let $\vec{X}_i = (1, X_i)$. Prove that $\hat{\beta}_1$ is the same as above.

[5] Note that the assumption M5 is needed just to ensure that the inverse matrix $(\mathbf{X}'\mathbf{X})^{-1}$ in (4.2) exists.

$s^2(=\hat{\sigma}_\varepsilon^2) = \sum e_i^2 / (N-(k+1))$,and X^{ij} is the the element in the ith row and j th column of the inverse matrix $(\mathbf{X}'\mathbf{X})^{-1}$. In general, $\widehat{\text{var}}\hat{\vec{\beta}} = s_\varepsilon^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$, i.e., $\widehat{\text{cov}}(\hat{\beta}_m, \hat{\beta}_l) = s^2 X^{ml}$. To describe the closeness of fit in multiple regression, we again use the coefficient of determination $R^2 = 1 - \dfrac{RSS}{TSS}$ which, as in a univariate case, equals $\widehat{cor}^{\,2}(Y, \hat{Y})$.

4.2. An Example

What are the criteria of a good model? The first ones are 1) all the variables must be significant, 2) their signs can be explained by the economic theory, and 3) if we can choose among several models, choose the one with the highest R^2 or similar coefficient of goodness of fit (in fact, we should also take into account the number of variables on the rhs, see Sec. 4.4 as well as some other factors and this is what we shall do in the next chapters).

As the first example of multivariate regression, consider the household.txt file where we have a random sample of 25 four-person (two adults) households. It contains information in hundreds of dollars on their annual disposable income `inc` and their annual total expenditure `cons` on nondurable goods and services. The expenditure data in the file excludes spending on durable goods, and hence is a reasonable approximation to the economist's definition of consumption[7]. Our aim is to create a model of consumption.

`cons`	total expenditure
`inc`	disposable income
`las`	stocks of liquid assets (around the 1st of July)
`fexp`	expenditure on food

We import the data to GRETL and start with univariate regression.

```
Model 1: OLS, using observations 1-25
Dependent variable: cons

             coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const       30.7063      20.6438       1.487    0.1505
  inc          0.812402     0.113151     7.180    2.60e-07 ***

Mean dependent var   163.2936   S.D. dependent var   81.31436
Sum squared resid    48958.73   S.E. of regression   46.13719
R-squared            0.691479   Adjusted R-squared   0.678065
F(1, 23)             51.54928   P-value(F)           2.60e-07
Log-likelihood      -130.2217   Akaike criterion     264.4434
Schwarz criterion    266.8811   Hannan-Quinn         265.1195
```

[6] All the distribution results in our model hold conditional on $\mathbf{X}$, for example $\hat{\beta}_m \mid \mathbf{X} \sim N(\beta_m, \sigma_\varepsilon^2 X^{mm})$.

[7] Both the life-cycle hypothesis and the permanent income hypothesis of consumer behaviour suggest that consumption depends not so much on income as on some measure of total lifetime resources. At the very least, some measure of consumers wealth needs to be included in a consumption function (we use `las`).

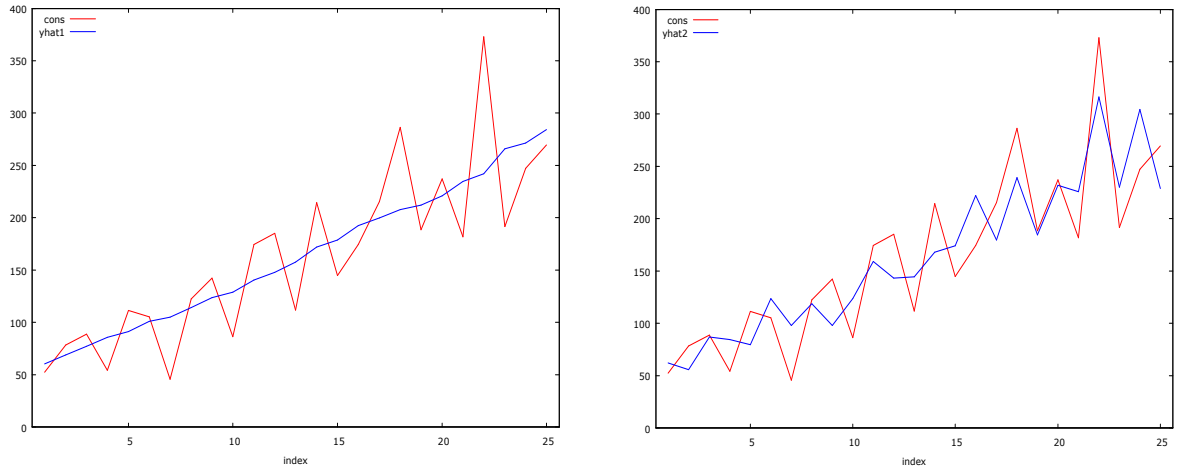


Figure 4.1. Actual (red) and fitted (blue) consumption: model 1 (left) and model 2 (right); model 2 is more accurate (blue line is closer to the actual `cons` line)

Formally speaking, this is an acceptable model – the coefficient $\hat{\beta}_1$ has the right sign, it is significant, R^2 is quite impressive. On the other hand, it contradicts to our economic theory because it is improbable that most of income goes to immediate consumption.

We extend our model by including `las` to the rhs:

```
Model 2: OLS, using observations 1-25
Dependent variable: cons

             coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const       36.7901      17.2945       2.127    0.0449   **
  inc          0.331830     0.172101     1.928    0.0668   *
  las          0.125786     0.0376880    3.338    0.0030   ***

Mean dependent var   163.2936   S.D. dependent var   81.31436
Sum squared resid    32501.96   S.E. of regression   38.43646
R-squared            0.795184   Adjusted R-squared   0.776564
F(2, 22)             42.70676   P-value(F)           2.66e-08
Log-likelihood      -125.1007   Akaike criterion     256.2014
Schwarz criterion    259.8580   Hannan-Quinn         257.2156
```

Recall the interpretation of the coefficients: if you take any two stratas of population where `inc` in a second one is 1 unit higher, `cons` there will be on average 0.3318 units greater, ceteris paribus, and if you take any two stratas of population where `las` in a second one is 1 unit higher, `cons` in the second strata will be on average 0.1258 units greater, ceteris paribus.

When presenting regression results, it is customary to abbreviate the above table and express it as

$$\widehat{cons} = 36.79 + 0.33\,inc + 0.13\,las$$
$$(1.82) \quad (2.59)$$

where, beneath estimated coefficients, the relevant $t-$ ratios are placed.

Is Model 2 better than Model 1? Let us begin with the coefficient at `inc` – it is much smaller and even insignificant at 5% level, but this is not surprising. In life-cycle and permanent income theories of consumption, it is some measure of overall lifetime resources (for example, `las`) rather than current measured income that is held to be the major influence on consumption. The $t-$ratios now are also smaller compared with 6.93 in model 1 which reflects the fact that `inc` and `las` are close to multicollinear (see Sec. 4.3) since $cor(inc, las) = 0.84$.

****

What will happen to the coefficient of, say, `inc` if we change the units of its measurement, e.g., instead of "hundreds of dollars" our records will be in "thousands of dollars" (this effectively means that all observations must be divided by 10, that is, `inc10=inc/10`)?

```
Dependent variable: cons
Heteroskedasticity-robust standard errors, variant HC1

             coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const       36.7901      14.4557       2.545    0.0185  **
  inc10        3.31830      1.81934      1.824    0.0818  *
  las          0.125786     0.0484841    2.594    0.0166  **

Mean dependent var   163.2936   S.D. dependent var   81.31436
Sum squared resid    32501.96   S.E. of regression   38.43646
R-squared            0.795184   Adjusted R-squared   0.776564
F(2, 22)             35.57516   P-value(F)           1.28e-07
Log-likelihood      -125.1007   Akaike criterion     256.2014
Schwarz criterion    259.8580   Hannan-Quinn         257.2156
```

Thus, only the coefficient β_1 and its `std.error` were modified, both they have increased 10 times. Note that if the model contains $\log(\texttt{inc})$ then, when passing to $\log(\texttt{inc}/10)$, the only change will be in the `const`, i.e., the intercept of the regression line will remain the same (why?).

4.3. Multicollinearity

Sometimes explanatory variables are tightly connected and it is impossible to disentangle the individual influences of explanatory variables. For example, consider the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ and assume that $X_{2i} = a_0 + a_1 X_{1i}$ for all $i = 1, ..., N$. Substituting this expression into the model, we get $Y = (\beta_0 + \beta_2 a_0) + (\beta_1 + \beta_2 a_1) X_1 + \varepsilon = \gamma_0 + \gamma_1 X_1 + \varepsilon$. Thus, however good are our <u>two</u> estimates of γ_0 and γ_1, we will never be able to obtain estimates of <u>three</u> original parameters β_0, β_1 and β_2. Fortunately, this situation virtually never arises in practice and can be disregarded. What frequently happens with real-world data is that an <u>approximate</u> linear relationship occurs among the sample values of explanatory variables. If one of the columns of the design matrix $\mathbf{X}$ is an approximate linear function of one or more the others, then the matrix $\mathbf{X}'\mathbf{X}$ will be close to singularity – that is, its determinant $\det \mathbf{X}'\mathbf{X}$ will be close to zero. Recall that the estimator of the variance-covariance matrix of $\hat{\hat{\beta}}$ is given by $\widehat{\text{var}}\hat{\hat{\beta}} = S_\varepsilon^2 (\mathbf{X}'\mathbf{X})^{-1}$, that is, if $\det \mathbf{X}'\mathbf{X}$ is „close to zero" (or the design matrix close to

singular or our data close to multicollinear), then $(\mathbf{X}'\mathbf{X})^{-1}$ will be „big" and every estimator $\hat{\beta}_m$ will be imprecise. Exactly collinear data are very rare, sometimes they are only close to collinear and, in econometrics, namely this case is termed multicollinearity. If this is the case, our model is „strange" - R^2 is big, but all (or some of) the variables are „insignificant" and/or their signs are wrong, and even small changes to the data set may have a marked effect on such estimates. The problem may be explaines by the fact that variables are tightly connected, it is impossible to disentangle the individual influences of explanatory variables.

> The sample is called multicollinear if one of the columns of the design matrix **X** is close to a linear function of one or more the others.

A popular measure of multicollinearity is the *variance inflation factor* (VIF) defined as $VIF(\hat{\beta}_m) = 1/(1-R_m^2)$ where R_m^2 is the coefficient of multiple determination when the variable X_m is regressed on all the other explanatory variables (one can show that $\operatorname{var}\hat{\beta}_m = VIF(\hat{\beta}_m)\,\sigma_\varepsilon^2 / \sum (X_{mi} - \bar{X}_m)^2$). If a certain $VIF(\hat{\beta}_m)$ exceeds 10 and respective p - value is >0.05, this may mean that $\hat{\beta}_m$ is only spuriously insignificant and, in fact, there is no need to remove X_m from the model.

What to do if we detect (approximate) multicollinearity? If our purpose is to predict Y and R^2 is high, there is no need to change anything; however, if our purpose is to estimate the individual influence of each variable, then such a model is bad. Are you sure that you need every variable in the model? Maybe you can drop some of them and eliminate multicollinearity? Another method is to combine all or some multicollinear variables into groups and to use the method of *principal components* (see an example in Computer Labs).

Note that except when it is perfect, multicollinearity does not imply any violation of the classical assumptions M1-M3 – the OLS estimators retain all the desired properties of unbiasedness, efficiency, consistency etc (they are still BLUE&C), we just need many more observations to get precise estimates of the coefficients.

Returning to our Model 2:

```
Variance Inflation Factors
Values > 10.0 may indicate a collinearity problem

            inc    3.333
            las    3.333
```

thus we should not care much about the multicollinearity problem

As a final note: multicollinearity inflates all the variances, thus deflates all the $t-$values, and makes variables „insignifican t". How can we distinguish true insignificance from multicollinearity? One of the solutions is to use the $F-$test (see Sec. 4.7): if the seemingly insignificant parameters are truly zero, then the $F-$test should not reject the joint hypothesis[8] involved. If it does, we have an indication that the low $t-$values are due to multicollinearity.

[8] The H_0 hypothesis is *all the relevant coefficients are jointly zeros*.

4.4. AIC, SIC and Similar Measures of Fit

Once again, is Model 2 better than Model 1? We could use `R - squared` to compare these models (the more the better, thus it seems that Model 2 with `R-squared=0.795184` is better than Model 1 with `R-squared=0.691`). However, R^2 always increases when we add more variables. Indeed, recall that $R^2 = 1 - RSS/TSS$ and

$$\min_{b_0,b_1,b_2} \sum (Y_i - (b_0 + b_1 X_{1i} + b_2 X_{2i}))^2 \le \min_{b_0,b_1} \sum (Y_i - (b_0 + b_1 X_{1i}))^2 .$$

This means that if we add one more variable to the rhs (in our case, we added `las`), even if the variable is in no way connected with Y, R^2 will increase. To test it further, add a variable `norm1` generated with `series norm1 = randgen(N,0,1)` to the rhs of the model:

```
Model 3
Dependent variable: cons

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       37.5095      16.7214       2.243     0.0358  **
  inc          0.330694     0.187518     1.764     0.0924  *
  las          0.124530     0.0480249    2.593     0.0170  **
  norm1       -1.00167     10.9164      -0.09176   0.9278

Mean dependent var   163.2936   S.D. dependent var   81.31436
Sum squared resid    32489.51   S.E. of regression   39.33344
R-squared            0.795263   Adjusted R-squared   0.766014
F(3, 21)             22.95634   P-value(F)           7.88e-07
Log-likelihood      -125.0959   Akaike criterion     258.1918
Schwarz criterion    263.0673   Hannan-Quinn         259.5441
```

Here, as one can see, despite the fact that `norm1` has no relation with `cons`, R^2 now is bigger that in Model 2. It means that we have to look for a "better" criterion of fit which would penalize for new variables and increase only if the new variable significantly lowers RSS = `Sum squared resid` (in our example RSS has only slightly decreased from 32501.96 to 32489.51).

Commonly reported criteria for model comparison are[9]:

Adjusted R^2 $\quad \bar{R}^2 = 1 - RSS(N-1)/TSS(N-k)$ (the more the better)

Akaike Information Criterion $\quad AIC = (RSS/N)\exp(2k/N)$ (the less the better)

Schwarz Bayesian Criterion $\quad SIC = (RSS/N)\,N^{k/N}$ (the less the better; it penalizes model's complexity more heavily than AIC)

Note that some statistical programs print out not AIC and SIC but, under the same name, their logarithms.

[9] Note that all they increase with the number of variables k; as a consequence, AIC and SIC will decrease only if the new RSS is considerable smaller.

Applying these criteria to our three models, we see that Model 2 is best:

```
Akaike criterion      264.4434   256.2014   258.1918
Schwarz criterion     266.8811   259.8580   263.0673
```

4.5. Categorical Variables on the Right Hand Side

Regression line $Y = \beta_0 + \beta_1 X$ describes the average value of Y for a given X. So far we have used a continuous or numeric variable, but what if X is a categorical or nominal variable such as gender or race? For example, the file hsb.txt (hsb=high school and beyond) contains 15 variables describing a nationally representative sample of 600 high school seniors:

```
          MALE       RACE       SES        SCTYP      HSP        LOCUS
          CONCPT     MOT        CAR        RDG        WRTG       MATH
          SCI        CIV
          0          1          1          1          3          0.29
          0.88       0.67       10         33.6       43.7       40.2
          39.0       40.6
          1          1          1          1          1          -0.42
          0.03       0.33       2          46.9       35.9       41.9
          36.3       45.6
          0          1          1          1          1          0.71
          0.03       0.67       9          41.6       59.3       41.9
          44.4       45.6
.....................................................................................
```

where

MALE[10]	1=male, 0=female
RACE	1=hispanic, 2=asian, 3=black, 4=white
SES	socio-economic status: 1=lower, 2=middle, 3=upper
SCTYP	school type (2 levels)
HSP	high school program (3 levels)
LOCUS	Locus of control
CONCPT	self concept
MOT	motivation
CAR	career choice (17 levels)
RDG	reading t-score
WRTG	writing t-score
MATH	math t-score
SCI	science t-score
CIV	civics t-score

We want to examine whether writing skills WRTG depend on sex, i.e., what are the average values of WRTG in these two groups and do these values differ significantly? We can create two subgroups of males and females, estimate sample means in both groups, and use $T-$test

[10] Even if one uses 1 and 0 to code male and female, these „numbers“ are labels rather than numbers.

to test the equality of these two means. Alternatively, we can use regression analysis to do this in one step. What is even more important, we can generalize it to the case of many levels (are the differences among RACEs significant?) where the $T-$test is not applicable and also include other variables into equation.

To tell GRETL that MALE is a label rather than number, select MALE and go to Variable| Edit attributes and check „Treat this variable as discrete" button. The standard procedure to deal with nominal variables is, instead of one discrete variable MALE to introduce (in our case, two) numerical *dummy (*or *indicator) variables* DMALE_1 and DMALE_2 for, respectively, females and males:

DMALE_1 = 1 if MALE=0 and =0 if MALE=1 (for females)
DMALE_2 = 1 if MALE=1 and =0 if MALE=0 (for males)

and to replace the original design matrix

$$\begin{pmatrix} 1\ WRTG_1\ \ MALE_1 \\ 1\ WRTG_2\ \ MALE_2 \\ \ldots\ldots\ldots\ldots\ldots\ldots \end{pmatrix} \text{ by } \mathbf{X}=\begin{pmatrix} 1\ WRTG_1\ \ DMALE\_1_1\ \ DMALE\_2_1 \\ 1\ WRTG_2\ DMALE\_1_2\ \ DMALE\_2_2 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{pmatrix} \text{ or, specifically, by } \begin{pmatrix} 1\ \ 43.7\ \ 0\ 1 \\ 1\ \ 35.9\ \ 1\ 0 \\ 1\ \ 59.3\ \ 0\ 1 \\ \ldots\ldots\ldots\ldots\ldots \end{pmatrix}.$$

Unfortunately, the regression equation $WRTG=\beta_0+\beta_1 DMALE\_1+\beta_2 DMALE\_2+\varepsilon$ is exactly multicollinear because in $\mathbf{X}$ the first column of 1's equals $DMALE\_1+DMALE\_2$ (this is called the *dummy variable trap*). Remember: always drop one dummy variable from the model, i.e., instead of $WRTG=\beta_0+\beta_1 DMALE\_1+\beta_2 DMALE\_2+\varepsilon$ estimate the equation $WRTG=\gamma_0+\gamma_1 DMALE\_2+\varepsilon$; the level (or group) where $DMALE\_2=0$ (women) is called the *base*, γ_1 shows the average change of $WRTG$ in the second group of males compared with the base group of women and the $p-$ value shows whether the change is significant.

In GRETL, to create the two dummy variables, select SEX (remember, it is now a discrete variable) and go to Add| Dummies for selected dummy variables. We get

```
Model 1: OLS, using observations 1-600
Dependent variable: WRTG

             coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const       54.5544       0.522008    104.5     0.0000   ***
  DMALE_2     -4.76835      0.773876     -6.162   1.32e-09 ***

Log-likelihood      -2197.306   Akaike criterion      4398.612
Schwarz criterion    4407.406   Hannan-Quinn          4402.036
```

which should be interpreted as follows: in the base group (where DMALE_2=0, i.e., in the females' group) the average value of WRTG is 54.554; in the group where DMALE_2=1 (i.e., in males' group) WRTG decreases on average by 4.768 and equals 49.786, and the increase significantly differs from 0 (because the $p-$ value 1.32e-09<0.05).

If we repeat similar procedure with RACE, we get[11]

```
Dependent variable: WRTG

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const        46.7197       1.09700     42.59     5.88e-183 ***
  DRACE_2       8.98028      1.92780      4.658    3.93e-06  ***
  DRACE_3      -0.419718     1.63602     -0.2565   0.7976
  DRACE_4       7.13520      1.18276      6.033    2.83e-09  ***

Log-likelihood      -2183.708   Akaike criterion      4375.417
Schwarz criterion    4393.005   Hannan-Quinn          4382.264
```

which reads as follows: in the base group of hispanic students the mean WRTG is 46.719, in the second group of asian students WRTG increases (with respect to hispanic group) by 8.980 (and the change is definitely not 0); in black group it is lower than in hispanic group but the difference (compared with the base group) is not significant, and in white group it is significantly higher.

If at least one dummy variable DRACE_i is significant, the model should contain the dummy variables of the RACE group. Another approach advices to remove insignificant terms, DRACE_3 in our case. This is equivallent to combining the base subgroup with the 3rd subgroup (the base now will consist of subgroups 1 and 3):

```
Dependent variable: WRTG

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const        46.5310       0.813207    57.22     1.63e-244 ***
  DRACE_2       9.16899      1.78056      5.150    3.55e-07  ***
  DRACE_4       7.32391      0.925484     7.914    1.22e-014 ***

Log-likelihood      -2183.742   Akaike criterion      4373.483
Schwarz criterion    4386.674   Hannan-Quinn          4378.618
```

Both Akaike and Schwarz criteria now have decreased, therefore we can take this model as the best one among those based on race only.

Thus, to describe (predict) WRTG we can use different models (which of the three is the best?):

$$WRTG = \beta_0 + \beta_1\, DMALE\_2 + \varepsilon$$
$$WRTG = \beta_0 + \beta_1\, DRACE\_2 + \beta_2 DRACE\_4 + \varepsilon$$
$$WRTG = \beta_0 + \beta_1 DMALE\_2 + \beta_2 DRACE\_2 + \beta_3 DRACE\_4 + \varepsilon$$

[11] Note, in order to avoid the dummy variable trap, we exclude the base variable DRACE_1 from the model.

Regression models also allow us to combine categorical and continuous variables. For example, two models which regress *WRTG* on *RDG* and also takes into account sex effects, could be written as follows:

$$WRTG = \beta_0 + \beta_1\, DMALE\_2 + \beta_3 RDG + \varepsilon = \begin{cases} \beta_0 + \beta_1 + \beta_3 RDG + \varepsilon & \text{if } DMALE\_2 = 1 \\ \beta_0 + \beta_3 RDG + \varepsilon & \text{if } DMALE\_2 = 0 \end{cases}$$

$$WRTG = \beta_0 + \beta_1\, DMALE\_2 + (\beta_3 + \beta_4 DMALE\_2) * RDG + \varepsilon =$$
$$\begin{cases} \beta_0 + \beta_1 + (\beta_3 + \beta_4) * RDG + \varepsilon & \text{if } DMALE\_2 = 1 \quad \text{(male)} \\ \beta_0 + \beta_3 * RDG + \varepsilon & \text{if } DMALE\_2 = 0 \quad \text{(female)} \end{cases}$$

In the first model, only the intercept is different for the two sex subgroups whereas in the second both the intercept and slope differs in the two subgroups (this is the *model with interaction* of DMALE_2 and RDG). We get

```
Dependent variable: WRTG

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       17.5275      1.56570      11.19     1.54e-026 ***
  DMALE_2      5.28965     0.582836      9.076    1.63e-018 ***
  RDG          0.616056    0.0287517    21.43     5.56e-076 ***

Log-likelihood      -2026.178   Akaike criterion     4058.355
Schwarz criterion    4071.546   Hannan-Quinn         4063.490
```

and (here DMR_2=DMALE_2*RDG)

```
Dependent variable: WRTG

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       14.3001      2.23308       6.404    3.08e-010 ***
  DMALE_2    -11.3273      3.04212      -3.723    0.0002    ***
  RDG          0.677691    0.0418525    16.19     3.85e-049 ***
  DSR_2       -0.116185    0.0574622    -2.022    0.0436    **

Log-likelihood      -2024.127   Akaike criterion     4056.254
Schwarz criterion    4073.842   Hannan-Quinn         4063.100
```

Thus, the best model so far (according to Akaike) is the last one. Note that the stricter Schwarz criterion chooses the above, more parsimoniuos, model.

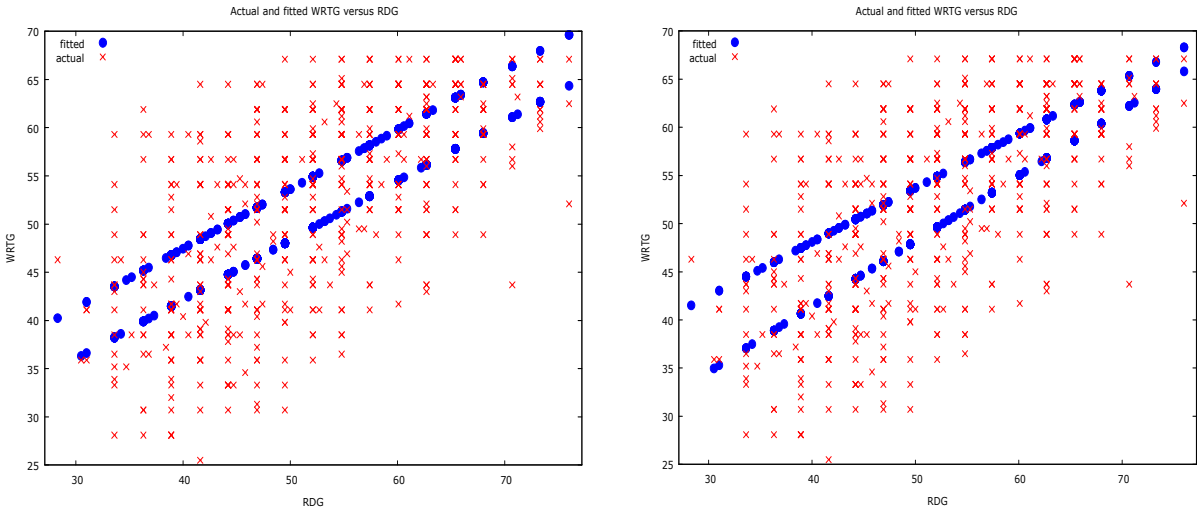


Figure 4.2. Model without interaction (only the intercepts for males and females differ, left) and with interaction (both intercepts and slopes differ, right)

In R we get exactly the same models, however the commands differ:

```
hsb=read.table(file.choose(),header=TRUE) # go to hsb.txt
head(hsb)
attach(hsb)
summary(lm(WRTG~factor(MALE)))       # factor recodes numbers to labels
summary(lm(WRTG~factor(MALE)*RDG))   # * stands for the interaction
```

4.6. Testing Hypotheses: One Linear Restriction

We have already mentioned that to test the hypothesis $H_0: \beta_m = \beta_m^0$ we use the the t- ratio statistics $(\hat{\beta}_m - \beta_m^0)/s_{\hat{\beta}_m}$ which has the $T_{N-(k+1)}$ distribution if disturbances are normal (the distribution is only approximately Student's if disturbances are not normal). However, in multivariate case we can test many more hypothesis, for example, $H_0: \beta_{k-(r-1)} = 0, ..., \beta_k = 0$ or $H_0: \beta_1 = \beta_2$ (which is equivalent to $H_0: 1\cdot\beta_1 + (-1)\cdot\beta_2 = 0$) or $H_0: a_1\beta_1 + a_3\beta_3 = 1$ etc. In the first case, we have r restrictions (we shall study this case in the next section) and in the second and third cases one (linear[12]) restriction. There are several methods to test such hypotheses.

Student or $T-$test. If $H_0: a_1\beta_1 + a_3\beta_3 = 1$ is true, the difference $a_1\hat{\beta}_1 + a_3\hat{\beta}_3 - 1$ should not differ much from zero, or, more precisely, the $t-$ratio

$$T = \frac{a_1\hat{\beta}_1 + a_3\hat{\beta}_3 - 1}{\sqrt{\widehat{\mathrm{var}}(a_1\hat{\beta}_1 + a_3\hat{\beta}_3)}} = \frac{a_1\hat{\beta}_1 + a_3\hat{\beta}_3 - 1}{S_\varepsilon\sqrt{a_1^2X^{11} + 2a_1a_3X^{13} + a_3^2X^{33}}}$$

[12] $H_0: \beta_1\beta_3 - 2\beta_2 = -1$ is an example of nonlinear restriction.

should not be „big". Random variable T has Student's distribution with $N-(k+1)$ d.f., therefore if the $p-$value $P(|T_{N-(k+1)}|\geq|t|)$ is less than 0.05 (here t is the realization of T in our sample), we reject H_0 with the 5% significance (for the definition of X^{11}, X^{13} etc, see Sec. 4.1; this is where we need the variance-covariance matrix $\widehat{\text{var}}\hat{\beta}$). Note that in practice this test is rarely performed.

Fisher or $F-$test. Let us consider two models – unrestricted

$$Y=\beta_0+\beta_1X_1+...+\beta_kX_k+\varepsilon \qquad \text{(UR)}$$

and restricted, assuming that $H_0: a_1\beta_1+a_3\beta_3=1$ is true:

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\left((1-a_1\beta_1)/a_3\right)X_3+\beta_4X_4+...+\beta_kX_k+\varepsilon \qquad \text{(R)}$$

If H_0 is true, both models should be of more or less of the same accuracy, i.e., $RSS_R-RSS_{UR}\approx 0$ or, more precisely,

$$F=\frac{\left(RSS_R-RSS_{UR}\right)/r}{RSS_{UR}/(N-(k+1))}$$

should not be big (here r is the number of restrictions, in our case it is 1). Both numerator and denominator are sums of squares of normal r.v.'s, therefore it does not come as a surprise that F has Fisher's distribution $F_{r,N-(k+1)}$. The p - value of this test is calculated by virtually all statistical programs and if $p-$value is less than 0.05, we reject H_0.

Wald (large sample) test

Recall that the denominator in

$$TS=\left(\frac{a_1\hat{\beta}_1+a_3\hat{\beta}_3-1}{\sqrt{\widehat{\text{var}}(a_1\hat{\beta}_1+a_3\hat{\beta}_3)}}\right)^2$$

(here TS stands for the Test Statistics) is a square root of the consistent estimator of $\text{var}(a_1\hat{\beta}_1+a_3\hat{\beta}_3)$, thus, when N is big, it is safe to treat $\widehat{\text{var}}(a_1\hat{\beta}_1+a_3\hat{\beta}_3)$ as a nonrandom true variance which implies that TS is a square of a standard normal r.v. In other words, TS has a χ_1^2 distribution; now, if the p - value $P(\chi_1^2>ts)$ of the test is less than 0.05, we reject H_0.

4.7. Testing Hypotheses: r Linear Restrictions

The hypotheses $H_0 : \beta_{k-(r-1)} = 0, ..., \beta_k = 0$ is frequently met in econometrics, it tests whether we have to keep the last r, $1 \le r \le k$, variables in our regression model (thus, if we accept it, at, say, 5% significance level, we remove all of them from the model). Note that to accept this hypothesis is not the same as to accept r separate hypothesis $H_0 : \beta_m = 0$, $m = k-(r-1), ..., k$, at the same significance level. Another example of several restrictions is the hypothesis $H_0 : a_1^{(1)}\beta_1 + ... + a_k^{(1)}\beta_k = c^{(1)}$, $a_1^{(2)}\beta_1 + ... + a_k^{(2)}\beta_k = c^{(2)}$ which contains two restrictions.

Student or $T-$test. It is not applicable in the case of more than one restriction.

Fisher or $F-$test. The only difference with the case of one restriction is in calculating RSS_R (now there will be r restrictions on the coefficients). Again, if H_0 is true, the F statistics should not be „big" or, more precisely, if $P(F_{r,N-(k+1)} > f) < 0.05$, where f is the realization of F in our sample, we reject H_0. Specifically, the $p-$value of the the null $H_0 : \beta_1 = ... = \beta_k = 0$ (i.e., Y is essentially a constant β_0 plus some unpredictable ε, or, in other words, knowing the values of $X_1, ..., X_k$ does not affect the expected value of Y), is presented in the regression output table in the line (see a few pages above)

```
F(3, 596) 210.8339      P-value(F) 3.44e-93
```

(thus we reject H_0). Note that the $F-$statistics there was

$$\frac{(RSS_R - RSS_{UR})/k}{RSS_{UR}/(N-k-1)} = \frac{(R^2_{UR} - R^2_R)/k}{(1-R^2_{UR})/(N-k-1)} = \frac{R^2/k}{(1-R^2)/(N-k-1)}$$

where R^2 is the usual $R-$squared from the regression of Y on $X_1, ..., X_k$.

4.1 example. (Chow test) Let us assume that we use use two samples and create two regression models. For instance, consider a system

$$\begin{cases} Y_i = \beta_0' + \beta_1' X_{1i} + ... + \beta_k' X_{ki} + \varepsilon_i', \ i = 1, ..., n \\ Y_i = \beta_0'' + \beta_1'' X_{1i} + ... + \beta_k'' X_{ki} + \varepsilon_i'', \ i = n+1, ..., n+m \end{cases}$$

where, for example, Y is wage and X's are age, education, experience etc. Now say that the first sample is that of males and the second of females and our question is whether these two models, i.e., their coefficients, coincide? Calculate $RSS_{UR} = RSS_1 + RSS_2$ first; next, to obtain RSS_R treat two samples as one and estimate RSS_R of this model – the $F-$statistics equals

$$F = \frac{(RSS_R - RSS_{UR})/k}{RSS_{UR}/(n+m-2k)}.$$

The Chow test can be performed in an equivalent manner as follows – write the above system as

$$Y_i = (\beta_0 + \alpha_0 M_i) + (\beta_1 + \alpha_1 M_i) X_{1i} + ... + (\beta_k + \alpha_k) X_{ki} + \varepsilon_i, i = 1,..., m+n,$$

where M is a male's dummy variable and test $H_0 : \alpha_0 = ... = \alpha_k = 0$ (what are unrestricted and restricted models?). To perform the Chow test in R, use $\texttt{anova}(\texttt{mod}_{\texttt{UR}}, \texttt{mod}_{\texttt{R}})$.

Wald (large sample) test

To test the hypothesis $H_0 : a_1^{(1)}\beta_1 + ... + a_k^{(1)}\beta_k = c^{(1)},\ a_1^{(2)}\beta_1 + ... + a_k^{(2)}\beta_k = c^{(2)}$ (can you write a simple variant of H_0?) with two restrictions, we use the fact that in large samples the test statistics

$$TS = \left(\frac{a_1^{(1)}\hat{\beta}_1 + ... + a_k^{(1)}\hat{\beta}_k - c^{(1)}}{\sqrt{\widehat{\text{var}}\left(a_1^{(1)}\hat{\beta}_1 + ... + a_k^{(1)}\hat{\beta}_k\right)}} \right)^2 + \left(\frac{a_1^{(2)}\hat{\beta}_1 + ... + a_k^{(2)}\hat{\beta}_k - c^{(2)}}{\sqrt{\widehat{\text{var}}\left(a_1^{(2)}\hat{\beta}_1 + ... + a_k^{(2)}\hat{\beta}_k\right)}} \right)^2,$$

provided H_0 holds true, has χ_2^2 distribution (if we have r restrictions, similar statistics will have χ_r^2 distribution). Thus, if $P(\chi_2^2 > ts) < 0.05$, we reject H_0.

Lagrange multiplier (LM) (large sample) test

This is a modification of the $F-$test in large samples to test the hypothesis $H_0 : \beta_{k-(r-1)} = 0,..., \beta_k = 0$. The restricted model now is

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_{k-(r+1)} X_{k-(r+1)} + \varepsilon^R \qquad \text{(R)}$$

and unrestricted

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_{k-(r+1)} X_{k-(r+1)} + ... + \beta_k X_k + \varepsilon^{UR}. \qquad \text{(UR)}$$

(can you write down the F statistics?).

The LM test is performed as follows:

Step 1 estimate the restricted model and save its residuals e_i^R, $i = 1,..., N$;

Step 2 regress e_i^R against <u>all</u>[13] $X_1,..., X_k$ plus a constant (the model is termed *auxiliary regression*) and calculate NR_{aux}^2; it has been shown that this random variable,

[13] If the omitted variables $X_{k-r},..., X_k$ truly have zero coefficients in DGP, then the residuals e^R, at least approximately, should be uncorrelated with each of these variables in the sample. This suggests running a regres-

provided H_0 is true, has the χ_r^2 distribution, therefore, if the $p-$value $P(\chi_r^2 > NR_{aux}^2)$ is less than 0.05, we reject H_0.

To reason the $LM-$test, note that if the DGP is described by the unrestricted model, the effects of omitted variables $X_{k-(r+1)},...,X_k$ should be captured by e_i^R, thus regressing them on all the X's should lead to a good fit and nonzero R_{aux}^2; thus big values of NR_{aux}^2 mean rejection of H_0.

****

As a general note, in order to simplify the model, choose one, the least significant, variable and, if its $p-$value is greater than 0.05, remove it from the original model; then repeat the procedure with a new model. Another possibility is the bulk removal of several variables at a time (use the F or LM tests to this end; simple rule of thumb says - add a variable to this list if its $p-$value is more than 0.5 (not 0.05!)).

There exist an automated procedure to simplify the model by removing insignificant variables (it is called *stepwise regression*). It is a risky procedure, especially in multicollinearity case, because sometimes just adding a few new observations could lead to a quite different model. To perform this regression, in GRETL, in model window, go to Tests| Omit variable, check „sequential...“ box; in R it is implemented by `stepAIC` function in MASS package and is performed with a step-by-step simplification of the model based on minimizing AIC.

****

Now we shall apply some of the above discussed tests to two examples.

4.2 example. The popular Cobb-Douglas production function may be expressed as a nonlinear multiplicative regression of the form $X_t = aK_t^{\beta_1} L_t^{\beta_2} e^{\varepsilon_t}, t = 1,...,T$. It can be transformed to a simpler linear log-log model $\log X_t = \beta_0 + \beta_1 \log K_t + \beta_2 \log L_t + \varepsilon_t$ where, for example, β_1 is the elasticity of output X with respect to capital K. Clearly, if the quantities of capital and labor inputs are doubled, then the output becomes $2^{\beta_1+\beta_2} X$ and we say that economy has *constant returns to scale* if $\beta_1 + \beta_2 = 1$. We shall test the assumption for the data in klein_maddala.txt where we have two inputs for both capital and labor. The model in GRETL is

```
Dependent variable: l_X

             coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
```

sion of these residuals on those independent variables excluded under H_0, which is almost what the *LM* test does. However, it turs out that, to get a usable test statistic, we must include all of the independent variables in the regression (we must include all regressors because, in general, the omitted regressors in the restricted model are correlated with the regressors that appear in the restricted model).

```
  const       -0.711093     1.31102     -0.5424   0.5911
  l_L1         1.40931      0.392238     3.593    0.0010  ***
  l_L2        -0.451694     0.408992    -1.104    0.2772
  l_K1         0.911616     0.228243     3.994    0.0003  ***
  l_K2        -0.546145     0.206974    -2.639    0.0125  **
```

In the model window, go to Tests| Linear restrictions, type in `b[2]+b[3]+b[4]+ b[5]=1` and get

```
Restriction:
 b[l_L1] + b[l_L2] + b[l_K1] + b[l_K2] = 1

Test statistic: Robust F(1, 34) = 2.7806, with p-value = 0.104599

Restricted estimates:

            coefficient   std. Error   t-ratio    p-value
  ---------------------------------------------------------
  const        1.37524     0.610646      2.252   0.0307    **
  l_L1         1.31848     0.345917      3.812   0.0005    ***
  l_L2        -0.693092    0.303169     -2.286   0.0284    **
  l_K1         1.25294     0.111731      11.21   3.89e-013 ***
  l_K2        -0.878326    0.0811165    -10.83   1.02e-012 ***
```

The constant returns to scale one-restriction hypothesis $H_0 : \beta_1 + \beta_2 + \beta_3 + \beta_4 = 1$ is tested here with the $F-$ test; the $p-$ value $P(F_{1,34} > 2.78) = 0.10 > 0.05$, therefore we have no ground to reject H_0 (in the table, the restricted model is presented). Similar answer is given by the $T-$ test: go to Tests| Sum of coefficients and get

```
Variables: l_L1 l_L2 l_K1 l_K2
   Sum of coefficients = 1.32309
   Standard error = 0.193755
```

Since the interval $1.32 \pm 2 * 0.19$ covers 1, we do not reject H_0 with 5% significance.

4.3 example. Open in GRETL hsb.txt (see Section 4.5). Our aim is to create a model for `WRTG`. As a first step, we shall, in two steps, dummify all <u>nominal</u> variables: in GRETL's main window select `SEX` and go to Variable| Edit attributes and check „Treat this variable as discrete" (repeat the same with `RACE` etc); then select all nominal variables and go to Add| Dummies for selected discrete variables and check „Skip the lowest value". Since we have many variables, probably we shall have to remove some insignificant variables from the model. Go to Model| Ordinary Least Squares| choose `WRTG` as dependent and all the rest as independent variables.

```
Dependent variable: WRTG
Heteroskedasticity-robust standard errors, variant HC1

            coefficient   std. error   t-ratio     p-value
  ---------------------------------------------------------
  const      10.6918       2.32765       4.593     5.38e-06  ***
  LOCUS       0.333209     0.434551      0.7668    0.4435
  CONCPT     -0.131674     0.421186     -0.3126    0.7547
```

```
  MOT          1.92976      0.817126      2.362     0.0185     **
  RDG          0.187808     0.0394520     4.760     2.46e-06   ***
  MATH         0.252978     0.0415044     6.095     2.02e-09   ***
  SCI          0.153911     0.0431880     3.564     0.0004     ***
  CIV          0.163680     0.0354927     4.612     4.94e-06   ***
  DSEX_2       4.22260      0.619358      6.818     2.37e-011  ***
  DRACE_2      1.33683      1.27671       1.047     0.2955
  DRACE_3     -1.71909      1.15637      -1.487     0.1377
  DRACE_4      0.753261     0.814667      0.9246    0.3556
  DSES_2      -0.0912213    0.708517     -0.1287    0.8976
  DSES_3      -0.178942     0.788942     -0.2268    0.8207
  DSCTYP_2     1.27310      0.709316      1.795     0.0732     *
  DHSP_2       0.409338     0.727809      0.5624    0.5740
  DHSP_3      -0.357951     0.784231     -0.4564    0.6483
  DCAR_2      -4.45071      1.42617      -3.121     0.0019     ***
  ..................................................
  DCAR_17     -3.19713      2.10858      -1.516     0.1300

Log-likelihood      -1915.428   Akaike criterion     3896.855
Schwarz criterion    4041.954   Hannan-Quinn         3953.339
```

We begin improving the model by the bulk exclusion of the „most insignificant“ variables (those with $p-$value > 0.5 (not 0.05!) they are marked in yellow color) in one step: in the model window, go to Tests→Omit variables and select „yellow“ variables -

```
  Test on Model 1:

  Null hypothesis: the regression parameters are zero for the variables
  CONCPT, DSES_2, DSES_3, DHSP_2, DHSP_3
  Test statistic: Robust F(5, 567) = 0.231201, p-value 0.948854
  Omitting variables improved 3 of 3 model selection statistics.

Model 2: OLS, using observations 1-600
Dependent variable: WRTG
Heteroskedasticity-robust standard errors, variant HC1

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const        9.94476     2.13700       4.654    4.06e-06  ***
  LOCUS        0.370853    0.418428      0.8863   0.3758
  MOT          1.89264     0.769885      2.458    0.0143    **
  RDG          0.190454    0.0390748     4.874    1.42e-06  ***
  MATH         0.260353    0.0409793     6.353    4.30e-010 ***
  SCI          0.151334    0.0429677     3.522    0.0005    ***
  CIV          0.168753    0.0351944     4.795    2.08e-06  ***
  DSEX_2       4.27932     0.609212      7.024    6.13e-012 ***
  DRACE_2      1.32978     1.27842       1.040    0.2987
  DRACE_3     -1.67773     1.15199      -1.456    0.1458
  DRACE_4      0.741039    0.813787      0.9106   0.3629
  DSCTYP_2     1.39801     0.669593      2.088    0.0373    **
  DCAR_2      -4.42289     1.40775      -3.142    0.0018    ***
  ..................................................
  DCAR_17     -3.23395     2.12697      -1.520    0.1290

Log-likelihood      -1916.061   Akaike criterion     3888.122
Schwarz criterion    4011.236   Hannan-Quinn         3936.048
```

The $p-$ value of the $F-$ test is greater than 0.05, therefore we accept the hypothesis that all five yellow variables together are insignificant. Now, we repeat the procedure with blue variables (in the previous case, the answer was predictable but now we have to test respective hypothesis (which hypothesis?)):

```
  Null hypothesis: the regression parameters are zero for the variables
  LOCUS, DRACE_2, DRACE_3, DRACE_4
  Test statistic: Robust F(4, 572) = 1.63755, p-value 0.163262
  Omitting variables improved 2 of 3 model selection statistics.

Model 3: OLS, using observations 1-600
Dependent variable: WRTG
Heteroskedasticity-robust standard errors, variant HC1

             coefficient   std. error   t-ratio     p-value
  ---------------------------------------------------------
  const         8.22222     1.98614       4.140     4.00e-05  ***
  MOT           1.78128     0.756928      2.353     0.0189    **
  RDG           0.188642    0.0394039     4.787     2.15e-06  ***
  MATH          0.277458    0.0400766     6.923     1.18e-011 ***
  SCI           0.179913    0.0412264     4.364     1.51e-05  ***
  CIV           0.166843    0.0352533     4.733     2.79e-06  ***
  DSEX_2        4.41259     0.602006      7.330     7.84e-013 ***
  DSCTYP_2      1.53279     0.671763      2.282     0.0229    **
  DCAR_2       -4.25045     1.40136      -3.033     0.0025    ***
  ...........................................................
  DCAR_17      -3.26318     2.17952      -1.497     0.1349

R-squared             0.626687   Adjusted R-squared    0.611781
Log-likelihood       -1920.171   Akaike criterion      3888.341
Schwarz criterion     3993.868   Hannan-Quinn          3929.421
```

Again $p-$ value of the $F-$ test is greater than 0.05, therefore we accept the hypothesis that all four blue variables are together insignificant. The final model contains only significant variables, its Schwarz criterion is minimum, R-squared is quite large. We can stop here or to continue with removing step by step all the insignificant DCAR_· In this case the final model is

```
Dependent variable: WRTG
Heteroskedasticity-robust standard errors, variant HC1

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const         6.44443     1.72918       3.727    0.0002    ***
  MOT           2.02391     0.741207      2.731    0.0065    ***
  RDG           0.190966    0.0398450     4.793    2.09e-06  ***
  MATH          0.282662    0.0398036     7.101    3.58e-012 ***
  SCI           0.179969    0.0406589     4.426    1.14e-05  ***
  CIV           0.173323    0.0348552     4.973    8.68e-07  ***
  DSEX_2        4.55269     0.549151      8.290    7.72e-016 ***
  DCAR_2       -3.02935     1.07918      -2.807    0.0052    ***
  DCAR_4       -2.31168     1.16972      -1.976    0.0486    **
  DCAR_5       -4.47529     2.04220      -2.191    0.0288    **
  DCAR_7       -3.65748     1.59903      -2.287    0.0225    **
  DCAR_12      -3.11489     1.86644      -1.669    0.0957    *
  DCAR_16      -3.43138     1.04889      -3.271    0.0011    ***
```

```
Log-likelihood      -1925.887   Akaike criterion     3877.775
Schwarz criterion    3934.935   Hannan-Quinn         3900.026
```

This final model does not suffer from multicollinearity (all VIFs are less than 3), its Akaike is the smallest one in our list of models. We spent a lot of time when creating the model but we can automate the selection procedure if in the very first model window go to Tests| Omit variables| and check the „Sequential elimination of variables using two-sided p-value:“ box - the final model has even smaller AIC=3870.79. Thus the procedure based on removing several variables at once can not always be the best. On the other hand, the criteria based on „leave only significant terms in the model“ and „look for a model with minimum AIC“ not always lead to the same result.

4.8. Violation of M1

Until now we have assumed that the multiple regression equation we are estimating includes all the relevant explanatory variables from the DGP. In practice, it is rarely the case. Sometimes some relevant variables are not included due to oversight or our ignorance, or lack of observations. On the other hand, sometimes some irrelevant variables are included.

4.8.1. Omission of Relevant Variables

Suppose that our DGP is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon, \qquad (4.3.1)$$

but you run

$$Y = \gamma_0 + \gamma_1 X_1 + \tilde{\varepsilon} \qquad (4.3.2)$$

(thus, an important X_2 is omitted; this happens quite often, especially when the rhs contains many variables). To partially fill this gap, run a fictitious regression $X_2 = \alpha_0 + \alpha_1 X_1 + u$ (recall that $\alpha_1 = \mathrm{cov}(X_2, X_1) / \mathrm{var}(X_1)$, cf. (3.3)); then

$$\begin{gathered} Y = \beta_0 + \beta_1 X_1 + \beta_2(\alpha_0 + \alpha_1 X_1 + u) + \varepsilon = \\ (\beta_0 + \beta_2\alpha_0) + (\beta_1 + \beta_2\alpha_1) X_1 + (\beta_2 u + \varepsilon) = \\ \gamma_0 + \gamma_1 X_1 + \tilde{\varepsilon}. \end{gathered}$$

The coefficient γ_1, estimated in (4.3.2), equals β_1 only in the case where $\alpha_1 = 0$, that is, when X_2 is not correlated with X_1. In any other case, $E(\hat{\gamma}_1 \mid \mathbf{X}) \neq \beta_1$ thus $\hat{\gamma}_1^{OLS}$ is generally biased[14] and inconsistent. The result is expected, because if X_1 and X_2 correlates, then X_1

[14] This is called the *omitted variable bias*. This bias is important: if our DGP has, for example, 10 variables, but we know only 7 – our decision to remove some variables will be biased and our final model may be wrong.

and the model's (4.3.2) error $\tilde{\varepsilon} = \beta_2 X_2 + \varepsilon$ correlates which implies bias[15] of $\hat{\gamma}_1$. The bottom line is: when searching for the „right" model, do not start with too simple model.

To better understand the claim, imagine that you generate N random numbers $X_{11},...,X_{1N}$, then[16] $X_{21},...,X_{2N}$, and finally $\varepsilon_1,...,\varepsilon_N$; calculate respective Y_i according to the, say, formula $Y_i = -0.5 + 3X_{1i} - 7.2X_{2i} + \varepsilon_i$. However, instead of using the matrix $\begin{pmatrix} Y_1 & X_{11} & X_{21} \\ ..... & & \\ Y_N & X_{1N} & X_{2N} \end{pmatrix}$ as your data set to estimate β_0, β_1, and β_2 from (4.3.1), you use (because of lack of X_2 data or your ingorance of the DGP) the matrix $\begin{pmatrix} Y_1 & X_{11} \\ ..... & \\ Y_N & X_{1N} \end{pmatrix}$ and estimate $\hat{\gamma}_0^{(1)}$ and $\hat{\gamma}_1^{(1)}$ of (4.3.2). Repeat the procedure n times where n is a „big" number - you will get that $\left(\hat{\beta}_1^{(1)} + ... + \hat{\beta}_1^{(n)}\right)/n$ is „not close" to $\beta_1\ (=3)$ from (4.3.1), i.e., the estimator $\hat{\gamma}_1$ is a biased estimator of β_1 from (4.3.1) as we have proved it earlier "theoretically".

4.8.2. Inclusion of Irrelevant Variables

Assume that our (X,Y) data are generated by $Y = \beta_0 + \beta_1 X_1 + \varepsilon$, but we possess more data and begin with the model $Y = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + u$ instead. Is it true that $\hat{\gamma}_1$ and $\hat{\gamma}_2$ are unbiased estimators of β_1 and $\beta_2\ (=0)$? The answer is „yes": recall that (see p. 4-3)

$$\hat{\gamma}_1 = \frac{\widehat{\mathrm{cov}}(X_1,Y)\,\widehat{\mathrm{var}}X_1 - \widehat{\mathrm{cov}}(X_2,Y)\,\widehat{\mathrm{cov}}(X_1,X_2)}{\widehat{\mathrm{var}}X_1\,\widehat{\mathrm{var}}X_2 - \left(\widehat{\mathrm{cov}}(X_1,X_2)\right)^2};$$

now, since $Y = \beta_0 + \beta_1 X_1 + \varepsilon$, after some simplification we get $E\hat{\gamma}_1 = \beta_1$ and, similarly, $E\hat{\gamma}_2 = 0$.

This result, coupled with the result from the previous section, might lead us to believe that it is better to include variables (when in doubt) rather to exclude them. However, this is not exactly so, because though the inclusion of irrelevant variables has no effect on the bias of the estimators, it does affect the variances.

The variance of $\hat{\beta}_1$ is given by $\mathrm{var}\,\hat{\beta}_1 = \sigma^2 / \sum x_{1i}^2$ and $\mathrm{var}\,\hat{\gamma}_1 = \sigma^2 / (1 - r_{12}^2) \sum x_{1i}^2$ where r_{12} is the correlation between X_1 and X_2. Thus $\mathrm{var}\,\hat{\gamma}_1 > \mathrm{var}\,\hat{\beta}_1$ unless $r_{12} = 0$. Hence we will be

[15] Recall that $\hat{\gamma}_1$ is BLUE&C only if U1-U3 holds (see p. 3-6). In particular, $\hat{\gamma}_1$ is unbiased if X_1 does not correlate with the error.

[16] To make X_2 correlate with X_1, use, for example, the formula $X_2 = 1 + 2X_1 + v$.

getting unbiased but inefficient estimators by including irrelevant variables. It can be also shown that the estimator for the residual variance we use is an unbiased estimator of σ^2.

To get a better understanding of the claim, imagine that you generate N random numbers $X_{11},...,X_{1N}$, then $\varepsilon_1,...,\varepsilon_N$, and, finally, calculate respective Y_i according to the, say, formula $Y_i = -0.5 + 3X_{1i} + \varepsilon_i$. However, instead of using the design matrix $\begin{pmatrix} 1\ X_{11} \\ ..... \\ 1\ X_{1N} \end{pmatrix}$ to estimate β_0 and β_1 from (4.3.2), you use (because you do not know the true DGP) the formula $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ and the design matrix $\begin{pmatrix} 1\ X_{11}\ X_{21} \\ ..... \\ 1\ X_{1N}\ X_{2N} \end{pmatrix}$ (here $X_{21},...,X_{2N}$ are generated by some arbitrary rule) to calculate $\hat{\beta}_0^{(1)}$, $\hat{\beta}_1^{(1)}$ and $\hat{\beta}_2^{(1)}$ of (4.3.1). Repeat the procedure n times where n is a „big" number: you will find that $\left(\hat{\beta}_1^{(1)} + ... + \hat{\beta}_1^{(n)}\right)/n$ is „close" to $\beta_1\ (=3)$ from (4.3.1) and $\left(\hat{\beta}_2^{(1)} + ... + \hat{\beta}_2^{(n)}\right)/n$ is „close" to $\beta_2\ (=0)$, i.e., the estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ are unbiased estimator of β_1 from (4.3.2) as we have proved earlier.

If a relevant variable is omitted, estimators of the coefficients are biased and inconsistent (however, if the excluded variable is not correlated with the included variable(s), then there will be no omitted variable bias).
If irrelevant variable is included, estimators of the coefficients are unbiased, but less precise.

4.9. Generalized Least Squares (GLS)

Let

$$\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\varepsilon}\,. \qquad (4.4.1)$$

So far we have analyzed the case where the condition M3, namely, $\text{var}(\vec{\varepsilon} \mid \mathbf{X}) = \sigma_\varepsilon^2 I_N$ holds (in words – errors $\varepsilon_i, i = 1,...,N,$ are uncorrelated and all have the same variance σ_ε^2). In many instances, the behavior of errors is close to the above described, thus the OLS estimator $\hat{\vec{\beta}}^{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\vec{Y}$ is a „good" one (what does it mean „good"?). However, sometimes[17] $\text{var}\,\vec{\varepsilon} = \mathbf{V} \neq \sigma_\varepsilon^2 I_N$ and then $\hat{\vec{\beta}}^{OLS}$ is no longer BLUE (in fact, it is unbiased and consistent but no longer efficient, and also the OLS estimate $\widehat{\text{var}}\hat{\vec{\beta}}^{OLS}$ is biased). If $\mathbf{V}$ is known (it is rarely so), define the matrix $\mathbf{P}$ as a solution to the equation[18] $\mathbf{P}'\mathbf{P} = \mathbf{V}^{-1}$, multiply from the left the

[17] We omit conditioning.
[18] What are the dimensions of the matrix **V**? And P?

both sides of (4.4.1) by $\mathbf{P}$, and define $\vec{Y}^* = \mathbf{P}\vec{Y}$, $\mathbf{X}^* = \mathbf{PX}$, $\vec{\varepsilon}^* = \mathbf{P}\vec{\varepsilon}$. Then (4.4.1) transforms to

$$\vec{Y}^* = \mathbf{X}^*\vec{\beta} + \vec{\varepsilon}^* \qquad (4.4.2)$$

where now $\text{var}\,\vec{\varepsilon}^* = I_N$. Applying the usual OLS to (4.4.2), we obtain the formula of what is called a generalized least squares estimator $\hat{\vec{\beta}}^{GLS} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\vec{Y}$. It can be shown that $\hat{\vec{\beta}}^{GLS}$ is BLUE in the class of linear unbiased estimators (and, of course, $\hat{\vec{\beta}}^{GLS} = \hat{\vec{\beta}}^{OLS}$ if $\mathbf{V} = I_N$).

To make the matters more transparent, let us assume that $\mathbf{V} = diag(\sigma_1^2,...,\sigma_N^2)$; then $\mathbf{V}^{-1} = diag(1/\sigma_1^2,...,1/\sigma_N^2)$, $\mathbf{P} = diag(1/\sigma_1,...,1/\sigma_N)$, and multiplication by $\mathbf{P}$ is equivalent to dividing the ith equation by (known!) σ_i which results in $\text{var}\,\varepsilon_i^* \equiv 1$. We get

$$Y_i/\sigma_i = \beta_0 \cdot 1/\sigma_i + \beta_1 \cdot X_{1,i}/\sigma_i + ... + \beta_k \cdot X_{k,i} + \varepsilon_i/\sigma_i,\ i = 1,...,N,$$

or

$$Y_i^* = \beta_0 \cdot 1/\sigma_i + \beta_1 \cdot X_{1,i}^* + ... + \beta_k \cdot X_{k,i}^* + \varepsilon_i^*,\ i = 1,...,N,$$

where to obtain the estimators of $\vec{\beta}$ the usual OLS can be applied). Note that in univariate case we have already used this procedure (called WLS, see p.3-25).

What to do if the $\mathbf{V}$ is not known? Even in the simplest case $\mathbf{V} = diag(\sigma_1^2,...,\sigma_N^2)$, we have to estimate $N + (k+1)$ parameters what is impossible since we have only N observations[19] (the rule of thumb says that we usually need at least 5 to 10 observations per one parameter to get a satisfactory estimator of $\vec{\beta}$). Therefore, we have to assume some simplifying conditions concerning $\mathbf{V}$, for example, if $\mathbf{V}$ is a diagonal matrix, suppose that $\sigma_i^2 = \alpha_0 + \alpha_1 X_{m,i}$ for some m, $m = 1,...,k$, (now, we have to estimate only two extra parameters). The case where we use estimated $\hat{\mathbf{V}}$ is called the *feasible* or *estimable generalized least* squares[20], the estimators have good properties in large sample. In what follows, we shall consider two most popular cases of GLS.

4.9.1. Heteroskedastic Errors

The case where $\text{var}(\vec{\varepsilon}\,|\,\mathbf{X})$ is diagonal, but not equal to σ_ε^2 times the identity matrix I_N, is referred to as *heteroskedasticity*. It means that the error terms are mutually uncorrelated, while

[19] Thus, the procedure of estimating $\vec{\beta}$ is unfeasible.

[20] Denoted, respectively, FGLS and EGLS.

the conditional variance of ε_i may vary over the observations[21]. Heteroskedasticity does not cause bias or inconsistency in the $\hat{\beta}_m^{OLS}$, whereas something like omitting an important variable would have this effect. However, without the homoskedasticity assumption $\hat{\beta}_m^{OLS}$ are no longer best, the OLS estimators of the $\text{var}\,\hat{\beta}_m$ are biased, and the usual OLS $t-$statistics do not have Student's distribution even in large samples. Similarly, F statistic no longer has Fisher distribution, and the *LM* statistic no longer has an asymptotic χ^2 distribution.

Our strategy will be as follows:

1. Assume that $Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \varepsilon$ is the true model and test the hypothesis H_0: *the model is homoskedastic* or, in other words, $H_0 : E(\varepsilon_i^2 \mid \mathbf{X}) \equiv \sigma^2, i = 1,..., N.$

2. If we accept the hypothesis[22], do nothing (i.e., use the usual OLS estimators).

3. If we reject the hypothesis, there are two variants:

   3a) stick to the OLS estimators $\hat{\beta}_m$, but correct the estimators of $\text{var}\,\hat{\beta}_m$ (White correction);

   3b) instead of OLS, use the weighted least squares (WLS) and get another model with better[23] estimators $\hat{\beta}_m^{WLS}$ and $\widehat{\text{var}}\hat{\beta}_m^{WLS}$.

1. Heteroskedasticity tests.

The most popular are the Breusch-Pagan (BP) and White tests. Generally speaking, if H_0 is false, the variance $\sigma_i^2 = \text{var}(\varepsilon_i \mid \mathbf{X})$ can be virtually any function of the X_m's. The BP test assumes a <u>linear</u> function, i.e., $\text{var}(\varepsilon_i \mid \mathbf{X}) = \pi_0 + \pi_1 X_{1,i} + ... + \pi_k X_{k,i}$; then the null hypothesis of <u>homo</u>skedasticity is formulated as $H_0 : \pi_1 = ... = \pi_k = 0$. To test it, we use the approximate equality $\sigma_i^2 \approx \varepsilon_i^2 \approx \hat{\varepsilon}_i^2$ and run a regression

$$\hat{\varepsilon}_i^2 = \pi_0 + \pi_1 X_{1,i} + ... + \pi_k X_{k,i} + u_i, i = 1,..., N; \qquad (4.5)$$

both the F or LM statistics for the overall significance of explanatory variables in explaining $\hat{\varepsilon}^2$ can be used to test H_0. The *LM* version of this test is typically called the *Breusch-Pagan test for heteroskedasticity*, its test statistics is which, under the null hypothesis, is distributed asymptotically as χ_k^2 (thus, if $P(\chi_k^2 > LM) < 0.05$, we reject H_0). Koenker sugges-

[21] If $\text{var}(\varepsilon_i \mid \mathbf{X}) \equiv \sigma_\varepsilon^2$, the errors are (or the model is) called *homoskedastic*.

[22] Remember that here (and always) the correct wording ought to be „if we fail to reject H_0".

[23] „Better" means with smaller variance compared to the estimators obtained by the OLS formulas.

ted another form of the *LM* statistic which is generally preferred (it is less dependent on the deviation of the errors ε from normality).

The alternative hypothesis in the *White test for heteroskedasticity* assumes that $\text{var}(\varepsilon_i \mid \mathbf{X})$ is described not by the linear function like in (4.5), but a more complicated one which can be approximated by its Taylor expansion up to the second power; the testing procedure begins with the regression

$$\hat{\varepsilon}^2 = \pi_0 + \pi_1^{(1)} X_1 + ... + \pi_k^{(1)} X_k + \pi_1^{(2)} X_1^2 + ... + \pi_{12}^{(2)} X_1 X_2 + ... + u\,. \qquad (4.6)$$

The number of variables in (4.6) increases very quickly (if $k = 3$, it has 9 terms; if $k = 6$ the White regression would generally involve 27 regressors, unless some are redundant). This abundance of regressors is a weakness in the pure form of the White test, therefore there exists a simplified form of the test which involves only square terms (and no cross products). In any case, we use the *LM* statistics to test the hypothesis that all π 's are zeros.

4.4 example. We use the data in hprice.txt to test for heteroskedasticity and create a relevant model in a housing price equation.

`price`	house price, $1000s
`assess`	assessed value, $1000s
`bdrms`	number of bedrooms
`lotsize`	size of lot in square feet
`sqrft`	size of house in square feet
`colonial`	=1 if home is colonial style
`lprice`	log(price)
`lassess`	log(assess)
`llotsize`	log(lotsize)
`lsqrft`	log(sqrft)

We start with the OLS Model 1, using levels and excluding `assess`:

$$price = \beta_0 + \beta_1 bdrms + \beta_2 lotsize + \beta_3 sqrft + \beta_4 colonial + \varepsilon\,.$$

Some explanatory variables are insignificant, therefore, in Model 1 window, go to Tests| Omit variables and check the "Sequential elimination ..." box. The final Model 2 is

```
Model 2: OLS, using observations 1-88
Dependent variable: price

             coefficient    std. error    t-ratio    p-value
  ---------------------------------------------------------
  const      5.93241       23.5124        0.2523   0.8014
  lotsize    0.00211349     0.000646560   3.269    0.0016    ***
  sqrft      0.133362       0.0113969    11.70     2.11e-019 ***

R-squared            0.663143   Adjusted R-squared   0.655217
Log-likelihood      -484.0985   Akaike criterion     974.1970
Schwarz criterion    981.6290   Hannan-Quinn         977.1912
```

To test H_0 : *the (original) model is homoskedastic*, we use the modified BP (or Koenker) test (in Model 1 window, go to Tests| Heteroskedasticity| Koenker):

```
Breusch-Pagan test for heteroskedasticity
Dependent variable: scaled uhat^2 (Koenker robust variant)

             coefficient     std. error    t-ratio   p-value
  ------------------------------------------------------------
  const      -8376.57       3398.84        -2.465    0.0158  **
  bdrms       1637.59       1092.47         1.499    0.1377
  lotsize        0.214738      0.0737843    2.910    0.0046  ***
  sqrft          1.27670       1.53140      0.8337   0.4069
  colonial   -2848.36       1680.54        -1.695    0.0938  *

Test statistic: LM = 16.150571,
with p-value = P(Chi-square(4) > 16.150571) = 0.002824
```

The small $p-$value evidences against the null (thus the variance depends on lotsize). This means that the usual standard errors reported in Model 1 table (and further sequential procedure) may be not reliable.

As a side note, heteroskedasticity is often reduced when passing to logarithms:

```
Model 3: OLS, using observations 1-88
Dependent variable: lprice

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const      -1.34959      0.651041    -2.073     0.0413    **
  colonial    0.0537962    0.0447732    1.202     0.2330
  llotsize    0.167819     0.0381806    4.395     3.25e-05  ***
  lsqrft      0.707193     0.0928020    7.620     3.69e-011 ***
  bdrms       0.0268304    0.0287236    0.9341    0.3530
```

Now, the respective Koenker test rejects heteroskedasticity:

```
Breusch-Pagan test for heteroskedasticity
Dependent variable: scaled uhat^2 (Koenker robust variant)
Test statistic: LM = 5.913380,
with p-value = P(Chi-square(4) > 5.913380) = 0.205711
```

and sequential elimination using two-sided alpha = 0.10 ends in

```
Model 4: OLS, using observations 1-88
Dependent variable: lprice

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       -1.64007     0.601880    -2.725     0.0078    ***
  llotsize     0.168457    0.0384596    4.380     3.37e-05  ***
  lsqrft       0.762369    0.0808862    9.425     7.44e-015 ***
```

We cannot compare directly Model 2 and Model 4 (left sides of the models differ) but we can go back to `price` in Model 4 as follows:

```
ols lprice 0 llotsize lsqrft # the log-log model
yhatt=exp($yhat+$sigma^2/2)  # return to the fitted price
genr cor2_log = corr(price,yhatt)^2 # Replaced scalar cor2_log = 0.72
```

Thus, upon comparing Model 2 (its `R-squared=0.663143`) with Model 4, we choose, at the moment, Model 4 (we can compare Model 2 and Model 4 by their `R-squared` because the number of variables on their rhs is the same). ◀◀

2. If we do not reject the homoskedasticity hypothesis, use the usual OLS estimators.

3a. If we reject the heteroskedasticity hypothesis, the first variant of our behavior is to stick to the OLS estimators $\hat{\beta}_m$, but correct the estimators of $\operatorname{var}\hat{\beta}_m$. This *White correction* will not change $\hat{\beta}_m$'s (remember, they are unbiased, but ineffective), it will only correct $\widehat{\operatorname{var}}\hat{\beta}_m$.

To motivate this correction, recall that in univariate case

$$\hat{\beta}_1 = \beta_1 + \sum \frac{x_i}{\sum x_i^2}\varepsilon_i = \beta_1 + \sum w_i\varepsilon_i$$

and $\operatorname{var}\hat{\beta}_1 = \sum w_i^2 \operatorname{var}\varepsilon_i$. If $\operatorname{var}\varepsilon_i \equiv \sigma^2$, we replace all $\operatorname{var}\varepsilon_i$ by $s^2 = \sum e_i^2/(N-2)$ and get $\widehat{\operatorname{var}}\hat{\beta}_1 = \sum w_i^2 s^2$. If $\operatorname{var}\varepsilon_i \not\equiv \sigma^2$, we estimate $\operatorname{var}\varepsilon_i$ by e_i^2 and get the White-corrected formula:

$$\widehat{\operatorname{var}}\hat{\beta}_1 = \frac{N}{N-2}\sum w_i^2 e_i^2$$

(the *White heteroskedasticity-consistent standard error* or *heteroskedasticity robust standard error* for $\hat{\beta}_1$ is given by the square root of this quantity). Note that this $\widehat{\operatorname{var}}\hat{\beta}_1$ is a consistent estimator of $\operatorname{var}\hat{\beta}_1$, that is, in large samples it is a good approximation to $\operatorname{var}\hat{\beta}_1$. However, do not use robust standard errors in small samples if there is no heteroskedasticity; for example, the corrected t - statistics will not necessarily have Student's distribution.

In multiple regression models, the formulas are more complex, but the correction principle is the same.

In GRETL, the correction is done by checking the „Robust standard errors" box in Model| Ordinary Least Squares... window. In R, use the `hccm` function from the car package or `vcovHC` function from the sandwich package.

As an example, we correct Model 1 which suffers from heteroskedasticity and then apply sequential procedure:

```
Model 5: OLS, using observations 1-88
Dependent variable: price
Heteroskedasticity-robust standard errors, variant HC1

             coefficient   std. error    t-ratio    p-value
  ---------------------------------------------------------
  const       5.93241       33.6711       0.1762    0.8606
  lotsize     0.00211349     0.00120361   1.756     0.0827    *
  sqrft       0.133362       0.0168342    7.922     8.12e-012 ***

Log-likelihood      -484.0985   Akaike criterion      974.1970
Schwarz criterion    981.6290   Hannan-Quinn          977.1912

Breusch-Pagan test for heteroskedasticity (robust variant) -
  Null hypothesis: heteroskedasticity not present
  Test statistic: LM = 13.8412
  with p-value = P(Chi-square(2) > 13.8412) = 0.000987224
```

We got the same model, i.e., the same coefficients (but different standard errors and, consequently, $p-$values). It means that `price`'s reaction to, say, `sqrft` is the same but now we are more confident that about the model.

3b. If we reject the homoskedasticity hypothesis, the second variant of our behavior is to use, instead of OLS, the *weighted least squares* (WLS) procedure which allows us to get another model with better estimators $\hat{\beta}_m^{WLS}$ and $\widehat{\text{var}}\hat{\beta}_m^{WLS}$.

Assume that $\text{var}(\varepsilon \mid \vec{X}) = \sigma^2 h(\vec{X})$ where $h(\vec{X})$ is some function of $\vec{X} = (X_1, ..., X_k)$. Provided we know $h(\vec{X})$, replace the regression equations

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \varepsilon_i \qquad (4.7)$$

with

$$Y_i / \sqrt{h_i} = \beta_0 / \sqrt{h_i} + \beta_1 X_{1i} / \sqrt{h_i} + ... + \beta_k X_{ki} / \sqrt{h_i} + \varepsilon_i / \sqrt{h_i} \text{ (here } h_i = h(\vec{X}_i)\text{)},$$

or

$$Y_i^* = \beta_0 X_{0i}^* + ... + \beta_k X_{ki}^* + \varepsilon_i^* . \qquad (4.8)$$

Now $\text{var}\,\varepsilon_i^* \equiv \sigma^2$, therefore the last equation does not suffer from heteroskedasticity and its OLS estimators $\hat{\beta}_m$ are BLUE&C. Note that, mathematically, the OLS estimation of (4.8) is equivalent to minimizing

$$RSS^{WLS} = \sum (Y_i - b_0 - b_1 X_{1i} - ... - b_k X_{ki})^2 / h_i \qquad (4.9)$$

with respect to b_m in original variables. The procedure is called the Weighted Least Squares (WLS), $1/h_i$ the weights, and respective solutions $\hat{\beta}_m^{WLS}$ the WLS estimators of β_m. Clearly, <u>if</u> we know $h(\vec{X})$, the estimators $\hat{\beta}_m^{WLS}$ are BLUE&C.

The basic problem with WLS is that usually we do not know $h(\vec{X})$. However, if we have established (for example, with the BP test) that $h(\vec{X})$ is a certain linear function of $\vec{X}$, we can replace h_i in (4.9) by $\hat{h}(\vec{X}_i)$. This procedure of correcting for heteroskedasticity is called the *Feasible Generalized Least Squares* (FGSL) and can be described as follows:

1. Run the OLS regression of Y on $X_1,...,X_k$ and obtain the residuals $\hat{\varepsilon}$.
2. Create $\log(\hat{\varepsilon}^2)$.
3. Run the regression of [24] $\log(\hat{\varepsilon}^2)$ on $X_1,...,X_k$ and obtain the fitted values, $\hat{g}$.
4. Exponentiate the fitted values: $\hat{h} = \exp(\hat{g})$
5. Estimate the equation $Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \varepsilon_i$ by WLS, using weights $1/\sqrt{\hat{h}}$.

The steps 3 and 4 are necessary to guarantee that the weights $\hat{h}$ would be positive. Now, when we use $\hat{h}$ instead of h, the FGLS estimators are biased but consistent and asymptotically (i.e., in large samples) more efficient than OLS. If you have some doubt about the variance specified in steps 3 and 4, use heteroskedasticity-robust standart errors of 3a.

The White correction does not change the coefficients of the original model, it only correctly estimates standard errors (and, thus, $p-$values) of the coefficients.
The weighted regression changes the coefficients and makes them more accurate.
Whatever technique we use, it does not remove heteroskedasticity from the data.

In GRETL, the FGLS procedure is done automatically through Model| Other linear models| Weighted Least Squares... or with the help of the following script:

```
ols price 0 bdrms lotsize sqrft colonial
series logres = log($uhat^2)
ols logres 0 bdrms lotsize sqrft colonial
series hhat = exp($yhat)
series ww = 1/sqrt(hhat)
wls ww price 0 bdrms lotsize sqrft colonial --robust
omit --auto=0.10
series yhat6 = $yhat
```

The output is presented below.

```
Model 6: WLS, using observations 1-88
Dependent variable: price
Heteroskedasticity-robust standard errors, variant HC1
Variable used as weight: ww
```

[24] This procedure (called the Harvey-Godfrey test) is similar to the BP test and can be used interchangeably with it.

```
            coefficient   std. error    t-ratio    p-value
  ------------------------------------------------------------
  const       40.5412       26.3001       1.541    0.1269
  lotsize      0.00320714    0.00161870   1.981    0.0508    *
  sqrft        0.110514      0.0125189    8.828    1.21e-013 ***

Statistics based on the weighted data:

R-squared             0.557315   Adjusted R-squared   0.546899
Log-likelihood       -328.9427   Akaike criterion     663.8854
Schwarz criterion     671.3174   Hannan-Quinn         666.8796
```

The standard errors in Model 6 are comparable with the those from Model 2, but now, in Model 6, we have better (that is, more effective or more precise) estimates of β's. Thus, we can use Model 6 as our final model of `price` (note that `yhat6` is very close to `yhatt` thus both models can be used to fit `price`).

```
Model 2: OLS, using observations 1-88
Dependent variable: price

            coefficient    std. error    t-ratio    p-value
  ------------------------------------------------------------
  const       5.93241       23.5124       0.2523   0.8014
  lotsize     0.00211349     0.000646560  3.269    0.0016    ***
  sqrft       0.133362       0.0113969    11.70    2.11e-019 ***

R-squared             0.663143   Adjusted R-squared   0.655217
Log-likelihood       -484.0985   Akaike criterion     974.1970
Schwarz criterion     981.6290   Hannan-Quinn         977.1912
```

****

The WLS estimators require an assumption about the form of heteroskedasticity. If that assumption is correct, the generalized least squares estimator is minimum variance. If that assumption is wrong (what is quite probable in multivariate case), then, like the OLS estimator, the WLS estimator will not be minimum variance, and its standard errors will be incorrect. This problem can be avoided by using OLS with White standard errors where an assumption about the form of heteroskedasticity is not needed, but then the potential reduction in variance of $\hat{\beta}_m$'s from generalized least squares will not be realized. Thus, which variant to choose?

After correcting for heteroskedasticity via WLS, one can test the residuals from the transformed model to see if any evidence of heteroskedasticity remains. If there is no evidence of remaining heteroskedasticity, then we can expect that generalized least squares has improved the precision of estimation, and that the chance of obtaining incorrect standard errors has been reduced. However, if we wish to err on the side of caution, or if further modeling fails to eliminate heteroskedasticity, we can use robust standard errors in conjunction with the generalized least squares estimator. Robust standard errors can be used not only to guard against the possible presence of heteroskedasticity when using least squares, they can be used to guard against the possible misspecification of a variance function when using generalized least squares.

****

As a final practical note – take care of heteroskedasticity only in severe cases, otherwise the usual OLS procedure gives quite satisfactory and reliable results.

4.9.2. Autoregressive Errors

We will now look at another case where the assumption $\text{var}(\vec{\varepsilon} \mid \mathbf{X}) = \sigma^2 I_N$ is violated, namely: what happens if the condition M3: $\text{cov}(\varepsilon_i, \varepsilon_j) = E\varepsilon_i\varepsilon_j = 0$ for all $i \neq j$ fails (if this assumption does not hold, the errors are called autocorrelated). Autocorrelation is held to occur most frequently when estimating equations using time series data (and to underline the case, we will use the index t instead of i). With such a data, there may be a tendency for random errors or shocks, or disturbancies to „spill over" from one time period to the next. For example, if inflation Y_t in one quarter is rather high then it is quite probable that the inflation will also be high next quarter, i.e., the correlation $cor(Y_t, Y_{t+1})$ will not be zero.

The consequences of autocorrelation are much the same as in homoskedastic case: the OLS estimators $\hat{\beta}_m^{OLS}$ remain unbiased and consistent, but are no longer best or asymptotically efficient. More seriously, the usual OLS formulas for estimating the variances of the estimators become biased, thus invalidating the customary OLS inferential procedures.

Perhaps the most popular way of modelling autocorrelated (or serially correlated) disturbances has been to replace the classical assumptions concerning the disturbances ε_t by the model

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t, |\rho| < 1, \tag{4.10}$$

where $u_t, t = 1,...,T,$ is a sequence of uncorrelated r.v.'s with zero mean and constant variance σ_u^2. The process $\{u_t\}$ is called the *white noise* (WN) process and $\{\varepsilon_t\}$ in (4.10) the *first-order autoregressive process* AR(1). The coefficient ρ indicates the strenght of relationship between ε_{t-1} and ε_t (in fact, it equals $cor(\varepsilon_t, \varepsilon_{t-1})$): if it is close to 0, $\{\varepsilon_t\}$ will be close to WN (thus the „no correlation" condition will be „almost true") and if ρ is close to +1, then the trajectories of $\{\varepsilon_t\}$ will have the property of persistency or inertia (this is the first sign of „nonwhiteness"). The autoregressive disturbances satisfy $E\varepsilon_t \equiv 0$ and $\text{var}\,\varepsilon_t \equiv \sigma_u^2/(1-\rho^2)$, i.e., the zero-mean and homoskedasticity conditions, but nevertheless do not satisfy M3 in full. The parameters ρ and σ_u^2 are typically unknown, and, along with β_m, we may wish to estimate them.

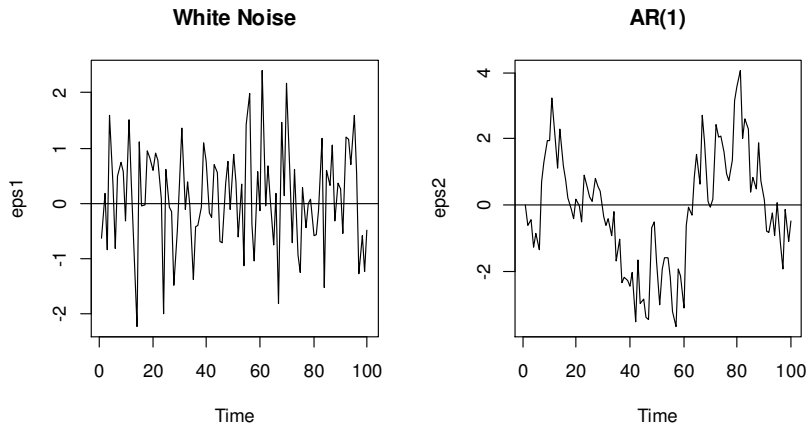


Figure 4.3. One trajectory of WN ($\rho = 0$, left) and AR(1) process ($\rho = 0.8$, right)

Note that our case of autoregressive errors is a particular case of GLS with

$$\mathbf{V} = \frac{\sigma_u^2}{1-\rho^2}\begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{N-1} \\ \rho & 1 & \rho & \dots & \rho^{N-2} \\ \dots & \dots & \dots & \dots & \dots \\ \rho^{N-1} & \rho^{N-2} & \rho^{N-3} & \dots & 1 \end{pmatrix}.$$

Our strategy in dealing with autoregressive models will be as follows:

1. Assume that

$$\begin{cases} Y_t = \beta_0 + \beta_1 X_{1t} + \dots + \beta_k X_{kt} + \varepsilon_t \\ \varepsilon_t = \rho\varepsilon_{t-1} + u_t \end{cases}$$

is the true model and test the hypothesis H_0 : *the errors constitute WN* or, in other words, $H_0 : \rho = 0$.

2. If we fail to reject the hypothesis, do nothing (i.e., the usual OLS procedure applies).

3. If we reject the hypothesis, there are two variants:

3a) stick to the OLS estimators $\hat{\beta}_m$, but correct the estimators of $\operatorname{var}\hat{\beta}_m$ (respective standard errors are known as HAC (heteroskedasticity and autocorrelation consistent) errors);

3b) instead of OLS, use *quasi-differenced* variables and get another model with better[25] estimators $\hat{\beta}_m^{GLS}$ and $\widehat{\operatorname{var}}\hat{\beta}_m^{GLS}$.

[25] „Better“ means with smaller variance compared to the estimators obtained by the OLS formulas.

1. Testing for autocorrelation.

The most popular are the Durbin-Watson (DW) and Breusch-Godfrey (BG) tests.

- The ***Durbin-Watson statistics*** is defined as

$$DW = \sum_{t=2}^{T}(e_t - e_{t-1})^2 / \sum_{t=1}^{T} e_t^2 . \qquad (4.11)$$

The numerator in (4.11) can be written as $\sum e_t^2 - 2\sum e_t e_{t-1} + \sum e_{t-1}^2 \approx 2\left(\sum e_t^2 - \sum e_t e_{t-1}\right)$, therefore

$$DW \approx 2(1-\hat{\rho}) = \begin{cases} 2, & \text{if } \hat{\rho} = 0 \\ 0, & \text{if } \hat{\rho} = 1 \end{cases}.$$

The DW statistics (which is present in GRETL regression table) can serve as a rough guide to test the hypothesis $H_0 : \rho = 0$ – if DW is „close“ to 2, the disturbances are, most probably, close to WN. The problem with DW is that, given $H_0 : \rho = 0$ is true, the critical values of its sampling distribution depends on the sample size T, the number of explanatory variables k, and on the values taken by those explanatory variables[26]. Even more so, the DW test is a test for first-order autocorrelation only, therefore it does not test the null $H_0 : \rho_1 = 0, \rho_2 = 0$ for a second-order process AR(2) such as $\varepsilon_t = \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + u_t$ and also for higher order processes such as $\varepsilon_t = \rho \varepsilon_{t-4} + u_t$ (such disturbances are quite common if your time series is quarterly). Another serious disadvantage of the DW statistics is that it is biased towards 2 when a lagged response variable Y_{t-1} is included among the regressors of an equation. For example, if $Y_t = \beta_0 + \beta_1 X_t + \beta_2 Y_{t-1} + \varepsilon_t$ and ε_t follows a first-order autoregressive process, then it is very likely that the DW statistic would fail to detect the autocorrelation. Therefore another test is far more applicable than the DW test.

- Suppose that we wish to test $H_0 : \rho_1 = \rho_2 = 0$ in the model

$$\begin{cases} Y_t = \beta_0 + \beta_1 X_t + \beta_2 Y_{t-1} + \varepsilon_t \\ \varepsilon_t = \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + u \end{cases}$$

1) Run the OLS regression of Y_t on X_t and Y_{t-1} and obtain the OLS residuals $\hat{\varepsilon}_t = e_t$, $t = 1,...,T$.

2) Run the auxiliary regression of e_t against X_t, Y_{t-1}, e_{t-1}, and e_{t-2}, $t = 3,...,T$, to obtain the ***F* test** for joint significance of e_{t-1} and e_{t-2}. If these two lags are jointly significant at a small enough, say, 5% level then we reject H_0 and conclude that the errors are serially correlated

[26] R can estimate the p – value with the `dwtest` function from the lmtest package.

(to find the respective $p-$value in GRETL, in the OLS model window go to Tests| Autocorrelation| Lag order for test:2→OK; there you will find the LMF statistics and respective p-value).

3) An alternative to computing the F test is to use the Lagrange multiplier form of the statistic. The LM statistic for testing H_0 is simply $LM = (T-2)R^2_{aux}$ where R^2_{aux} is just the usual $R-$squared from the auxiliary regression (under the null hypothesis, $LM \overset{asymp}{\sim} \chi^2_2$). This is usually called the **Breusch-Godfrey** test for AR(2) serial correlation (in GRETL, to apply the LM test, go to the same window and look for T·R^2; it is followed by respective $p-$value).◀

In many cases, the presence of autocorrelation is not an indication that the model has autocorrelated errors but rather that it is misspecified, for example, suffering from omitted variables or lagged terms, or just because of wrong functional form (e.g., X was used instead of $\log X$). We shall discuss the issue later.

4.5 example. The file icecream.dat contains five time series (30 four-weekly observations):

`cons`	consumption of ice cream per head (in pints)
`income`	average family income per week (in US Dollars)
`price`	price of ice cream (per pint)
`temp`	average temperature (in Fahrenheit)
`time`	index from 1 to 30

The model used to explain consumption of ice cream is a linear regression model with `income`, `price`, and `temp` as explanatory variables.

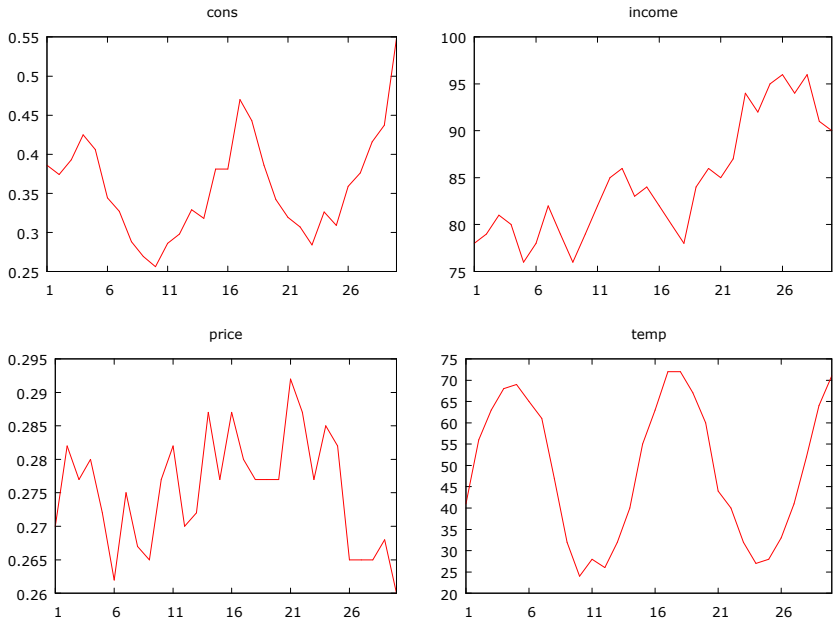


Figure 4.4. It seems that `cons` most closely follows `temp`

```
Model 1: OLS, using observations 1-30
Dependent variable: cons

             coefficient   std. error    t-ratio   p-value
  ------------------------------------------------------------
  const       0.197315     0.270216      0.7302   0.4718
  income      0.00330776   0.00117142    2.824    0.0090   ***
  price      -1.04441      0.834357     -1.252    0.2218
  temp        0.00345843   0.000445547   7.762    3.10e-08 ***

rho                 0.400633   Durbin-Watson        1.021170
```

While the coefficient estimates have the expected signs, the DW statistics is computed as 1.02 which is quite far from 2; thus, most probably, the null hypothesis $H_0 : \rho = 0$ should be rejected against the alternative of positive autocorrelation. Note that $\hat{\rho} = 0.401$ which can also be obtained by running the regression $\texttt{uhat}_t = \rho\,\texttt{uhat}_{t-1} + v_t$:

```
Dependent variable: uhat1

             coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  uhat1_1     0.400633      0.177417    2.258    0.0319   **
```

The $p-$value is <0.05, thus once again we reject $H_0 : \rho = 0$. To double check, in the Model 1 window go to Tests→Autocorrelation:

```
Breusch-Godfrey test for first-order autocorrelation
OLS, using observations 1-30
Dependent variable: uhat

             coefficient    std. error    t-ratio   p-value
  ------------------------------------------------------------
  const       0.0615530     0.257165      0.2394   0.8128
  income     -0.000115792   0.00110852   -0.1045   0.9176
  price      -0.147641      0.791862     -0.1864   0.8536
  temp       -0.000203334   0.000432839  -0.4698   0.6426
  uhat_1      0.428282      0.211215      2.028    0.0534   *

  Unadjusted R-squared = 0.141235

Test statistic: LMF = 4.111588,
with p-value = P(F(1,25) > 4.11159) = 0.0534

Alternative statistic: TR^2 = 4.237064,
with p-value = P(Chi-square(1) > 4.23706) = 0.0396

Ljung-Box Q' = 3.6,
with p-value = P(Chi-square(1) > 3.6) = 0.0578
```

All the $p-$values are less than or just marginally greater than 0.05, therefore we stick to the assumption that the errors are AR(1).

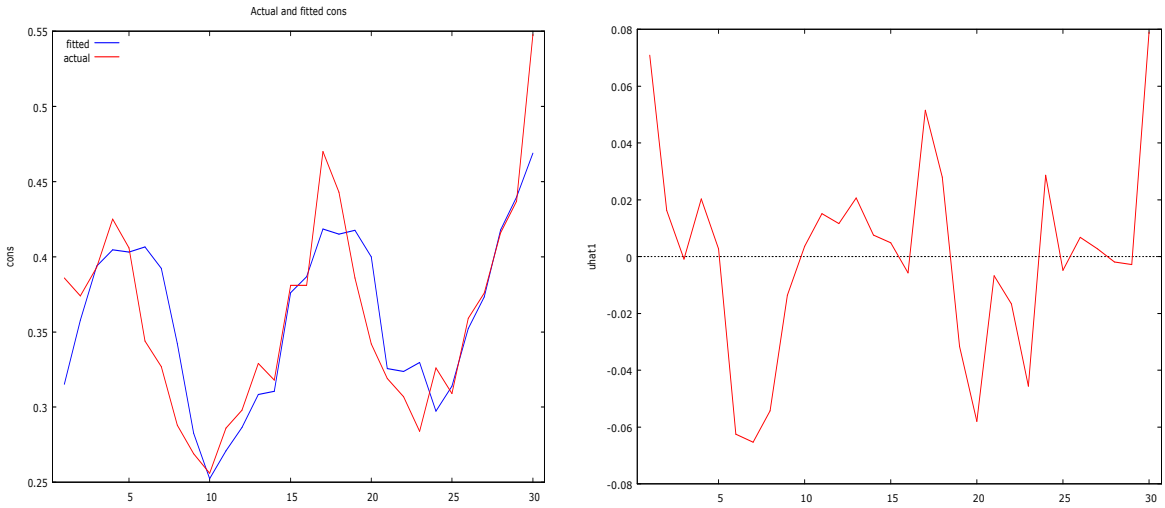


Figure 4.5. Both graphs indicate that residuals have the persistency property (for a long time they have the same sign), therefore they are, probably, AR(1)

2. If we do not reject the hypothesis that errors make WN, use the usual OLS estimators.

3a. If we reject the WN hypothesis, the first variant of our behavior is to stick to the OLS estimators $\hat{\beta}_m$, but correct the estimators of $\operatorname{var}\hat{\beta}_m$ (these standard errors are known as the Newey-West consistent estimator or HAC or *heteroskedasticity and autocorrelation consistent* errors). To motivate this correction, recall that in univariate case

$$\hat{\beta}_1 = \beta_1 + \sum \frac{x_i}{\sum x_i^2}\varepsilon_i = \beta_1 + \sum w_i\varepsilon_i\,.$$

Taking into account the equality $\operatorname{var}\sum_{t=1}^{T} Z_t = \sum_{t,s=1}^{T} \operatorname{cov}(Z_t, Z_s)$, we get

$$\operatorname{var}\hat{\beta}_1 = \sum_{t=1}^{T} w_t^2 \operatorname{var}\varepsilon_t \cdot \left(1 + \frac{\sum_{t\neq s} w_t w_s \operatorname{cov}(\varepsilon_t, \varepsilon_s)}{\sum_{t=1}^{T} w_t^2 \operatorname{var}\varepsilon_t}\right), \qquad (4.12)$$

from which, upon replacing cov and var by their estimates, we get the HAC $\widehat{\operatorname{var}}\hat{\beta}_1$. Do not be disturbed if you see slightly different HAC standard errors in different statistical programs – there are many variants of (4.12). In GRETL, go to Model→Ordinary Least Squares and check the Robust standard errors box:

```
Dependent variable: cons
HAC standard errors, bandwidth 2 (Bartlett kernel)

             coefficient   std. error    t-ratio   p-value
  ------------------------------------------------------------
```

```
const        0.197315     0.299594       0.6586    0.5159
income       0.00330776   0.00118427     2.793     0.0097    ***
price       -1.04441      0.876164      -1.192     0.2440
temp         0.00345843   0.000410546    8.424     6.63e-09  ***
```

Compared with Model 1, the only changes are in `std. error`, `t-ratio`, and `p-value`, but all they are not essential.

3b. If we reject the WN hypothesis, the second variant of our behavior is to transform the original equation so that the new errors become WN. This approach not only corrects standard errors but also change coefficients and make the estimators efficient. We shall consider two procedures – the Cochrane-Orcutt (CORC) and Hildreth-Lu.

- The iterative *CORC procedure* is described as follows. Take the simplest model

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t \qquad (4.13)$$

where $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$, multiply the lagged equation $Y_{t-1} = \beta_0 + \beta_1 X_{t-1} + \varepsilon_{t-1}$ by ρ, and subtract it from the first equation in (4.13). We get

$$Y_t - \rho Y_{t-1} = \beta_0(1-\rho) + \beta_1(X_t - \rho X_{t-1}) + u_t \qquad (4.14)$$

or

$$Y_t^* = \beta_0^* + \beta_1 X_t^* + u_t \qquad (4.15)$$

where $Y_t^* = Y_t - \rho Y_{t-1}$ and $X_t^* = X_t - \rho X_{t-1}$ are *quasi-differenced* variables; since u_t are serially independent with a constant variance, the OLS estimators in (4.15) will produce the BLUE&C estimators $\hat{\beta}_1$ and $\hat{\beta}_0 = \hat{\beta}_0^* / (1-\rho)$. The only problem is with ρ - as a rule, we do not know it, therefore we shall estimate it with the help of the following iterative process:

1. Estimate (4.13) with OLS; use its residuals $\hat{\varepsilon}_t^{(1)}$ to estimate $\rho^{(1)} = \dfrac{\sum \hat{\varepsilon}_t^{(1)} \hat{\varepsilon}_{t-1}^{(1)}}{\sum \hat{\varepsilon}_t^{(1)\,2}}$.

2. Substitute $\rho^{(1)}$ to (4.14) and estimate (4.15) with OLS; denote its estimated coefficients by $\beta_0^{*(1)}$ and $\beta_1^{(1)}$; substitute $\beta_0^{(1)} = \beta_0^{*(1)} / (1-\rho^{(1)})$ and $\beta_1^{(1)}$ to (4.13) and calculate $\hat{\varepsilon}_t^{(2)} = Y_t - \beta_0^{(1)} - \beta_1^{(1)} X_t$, then estimate $\rho^{(2)} = \dfrac{\sum \hat{\varepsilon}_t^{(2)} \hat{\varepsilon}_{t-1}^{(2)}}{\sum \hat{\varepsilon}_t^{(2)\,2}}$.

3. Substitute $\rho^{(2)}$ to (4.14) etc; this iterative procedure will be stopped when the estimates of ρ from two succesive iterations differ no more than some preselected value, such as 0.001. The final $\hat{\rho}$ is then used to get the CORC (or *FGLS*) estimates of both original β_1 in (4.13), ρ, and $\beta_0 = \beta_0^* / (1-\rho)$. These FGSL estimators are not unbiased but they are consistent and asymptotically efficient.

Note that if there are lagged values of Y as explanatory variables, one should not use the CORC procedure (the standard errors of $\hat{\beta}_m$ from (4.15) are not correct even asymptotically). Another complication with the CORC procedure is that minimizing of RSS in (4.15) can produce multiple solutions for ρ. In this case, the CORC procedure might give a local minimum. Hence it is better to use a grid-search procedure.

- The grid-search *Hildreth-Lu procedure* is as follows. Calculate quasi-differenced Y_t^* and X_t^* for different values of ρ at intervals of 0.1 in the range $-1 \le \rho \le 1$. Estimate the regression of Y_t^* on X_t^* and calculate the RSS in each case. Choose the value of ρ for which the RSS is minimum. Again repeat the procedure for smaller intervals of ρ around this value. For instance, if the value of ρ for which RSS is minimum is -0.6, repeat this search procedure for values of ρ at intervals of 0.01 in the range $-0.7 < \rho < -0.5$.

4.6 example. We continue analyzing the icecream.dat data. First, recall the OLS model:

```
OLS -  using observations 1-30
Dependent variable: cons

              coefficient   std. error    t-ratio   p-value
  ---------------------------------------------------------
  const       -0.113195     0.108280      -1.045    0.3051
  income       0.00353017   0.00116996     3.017    0.0055   ***
  temp         0.00354331   0.000444956    7.963    1.47e-08 ***

rho                  0.391242   Durbin-Watson        1.003337
```

Now, go to Model→Time series→ Cochrane-Orcurtt...:

```
Cochrane-Orcutt, using observations 2-30 (T = 29)
Dependent variable: cons
rho = 0.377096

              coefficient   std. error    t-ratio   p-value
  ---------------------------------------------------------
  const       -0.130557     0.137246      -0.9513   0.3502
  income       0.00362842   0.00148571     2.442    0.0217   **
  temp         0.00367472   0.000537402    6.838    2.94e-07 ***
```

or, if you go to Model→Time series→Hildreth-Lu...,:

```
Fine-tune rho using the CORC procedure...

                ITER       RHO        ESS
                  1      0.38000   0.0266723
                  2      0.37761   0.0266721
                  3      0.37708   0.0266721
Hildreth-Lu, using observations 2-30 (T = 29)
Dependent variable: cons
rho = 0.377084

              coefficient   std. error    t-ratio   p-value
  ---------------------------------------------------------
```

```
const       -0.130561     0.137244      -0.9513    0.3502
income       0.00362847   0.00148570     2.442     0.0217   **
temp         0.00367472   0.000537396    6.838     2.94e-07 ***
```

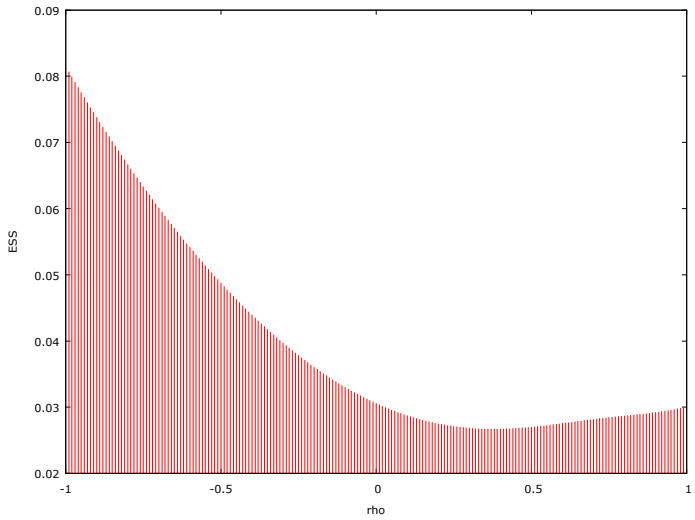


Figure 4.6. The Hildreth-Lu graph of RSS as a function of ρ

All three models are similar and can be expressed as $cons_t = -0.1306 + 0.0036 \cdot income_t + 0.0037\, temp_t + \varepsilon_t$ where $\varepsilon_t + 0.377\varepsilon_{t-1} + u_t$. ◀◀

****

Again, as in heteroskedasticity case, we have to chose between two procedures: CORC or similar and HAC errors. However, currently another method is generally preferred by applied econometricians. To explain it, consider a simple model $\begin{cases} Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t \\ \varepsilon_t = \rho\varepsilon_{t-1} + u_t \end{cases}$ which can be transformed to $Y_t = \beta_0(1-\rho) + \beta_1 X_t + \rho Y_{t-1} - \rho\beta_1 X_{t-1} + u_t$ which can be generalized to $Y_t = \gamma_0 + \theta_1 Y_{t-1} + \gamma_1 X_t + \gamma_2 X_{t-1} + u_t$. This model is called autoregressive distributed lag model and is studied in *Practical Econometrics.II* course.

****

We have considered two cases of GLS, namely, heteroskedastic and autoregressive errors and also two methods to correct the deviations of errors from the iid case – WLS and CORC methods, respectively. While these methods are still in use, an alternative approach has found increasing favor: that is, use OLS but compute robust standard errors (or more generally, covariance matrices). This is typically combined with an emphasis on using large datasets – large enough that the researcher can place some reliance on the (asymptotic) consistency property of OLS. This approach has been enabled by the availability of cheap computing power. The computation of robust standard errors and the handling of very large datasets were daunting tasks at one time, but now they are unproblematic. The other point favoring the newer methodology is that while FGLS offers an efficiency advantage in principle, it often involves making additional statistical assumptions which may or may not be justified, which may not be easy to test rigorously, and which may threaten the consistency of the estimator.

****

Sometimes model's residuals demonstrate certain kind of persistency, for example, retain the same sign for some time – the natural guess then is that errors are AR(1). On the other hand, such a behavior of residuals may just indicate that our model is misspecified. For example, Models 1 and 2 below are described, respectively, as $Y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \varepsilon_t$ or $Y_t = \beta_0 + \beta_1 \log t + \varepsilon_t$, thus, both have curvilinear trends. If we have mistakenly estimated them via linear trends, residuals will show spurious autocorrelation. To notice our mistake is easy in univariate case but in the case, where we have many variables, we need some formal tests.

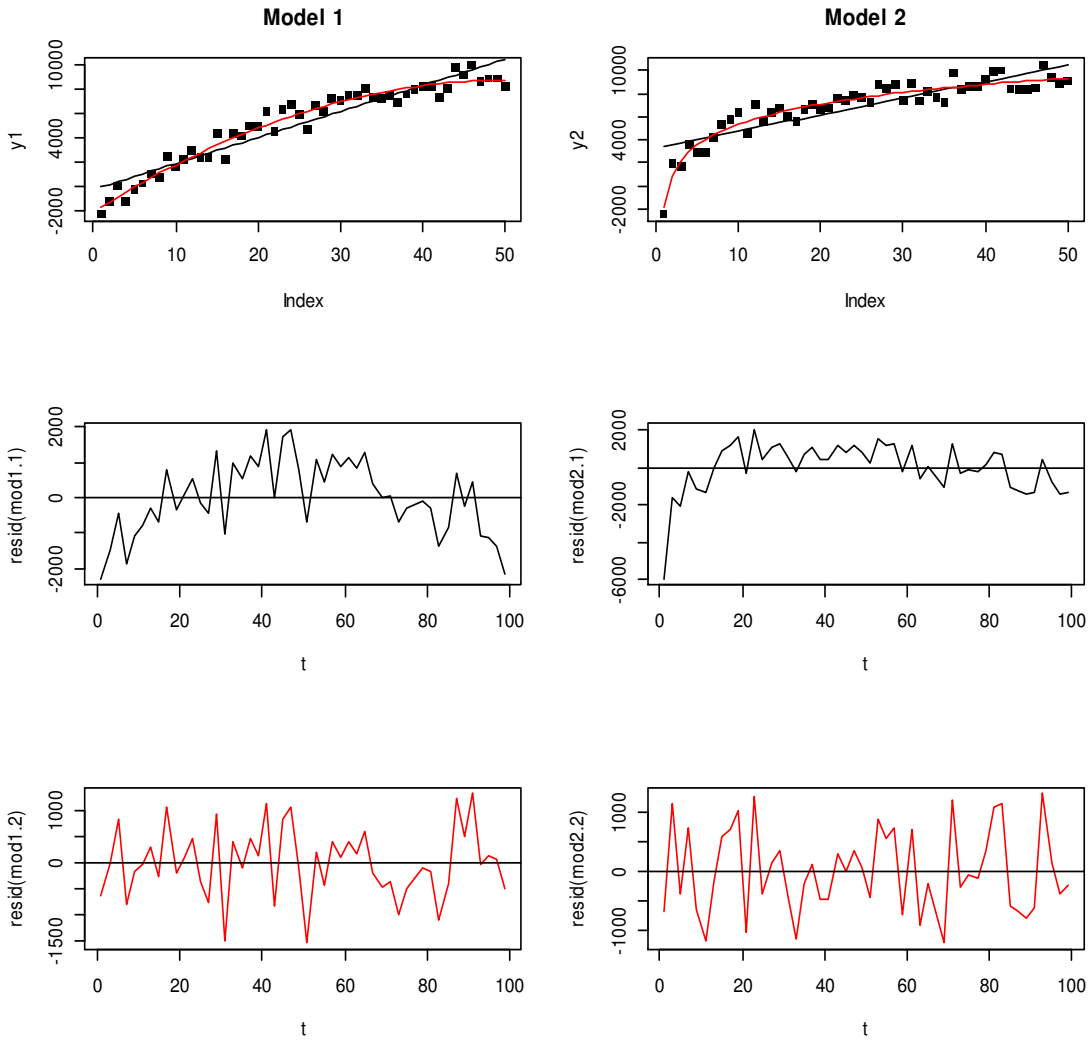


Figure 4.7. Residuals in the second line correspond to the linear models (most probably residuals will not reject AR(1) hypothesis); residuals in the third line correspond to true models, they are very much like WN

Thus, the autocorrelated residuals are more frequently the result of misspecified regression equation rather than genuine autocorrelation. The next section is devoted to misspecification issues.

4.10. Regression Model Specification Tests

In reading applied work, you will often encounter regression equations where the dependent variable appears in logarithmic form, for example, $\log(wage) = \beta_0 + \beta_1 educ + \varepsilon$. Recall that this model gives an (approximately) constant percentage effect as opposed to linear model $wage = \beta_0 + \beta_1 educ + \varepsilon$ where β_1 describes a constant units effect in response to 1 year change in *educ* (the former model is more realistic). On the other hand, if hourly wage's DGP is determined by $\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \varepsilon$, but we omit the squared experience term, $exper^2$, then we are committing a functional form misspecification which generally leads to biased[27] estimators of $\beta_0, \beta_1,$ and β_2. Fortunately, in many cases, using logarithms of certain variables and adding quadratics is sufficient for detecting many important nonlinear relationships in economics.

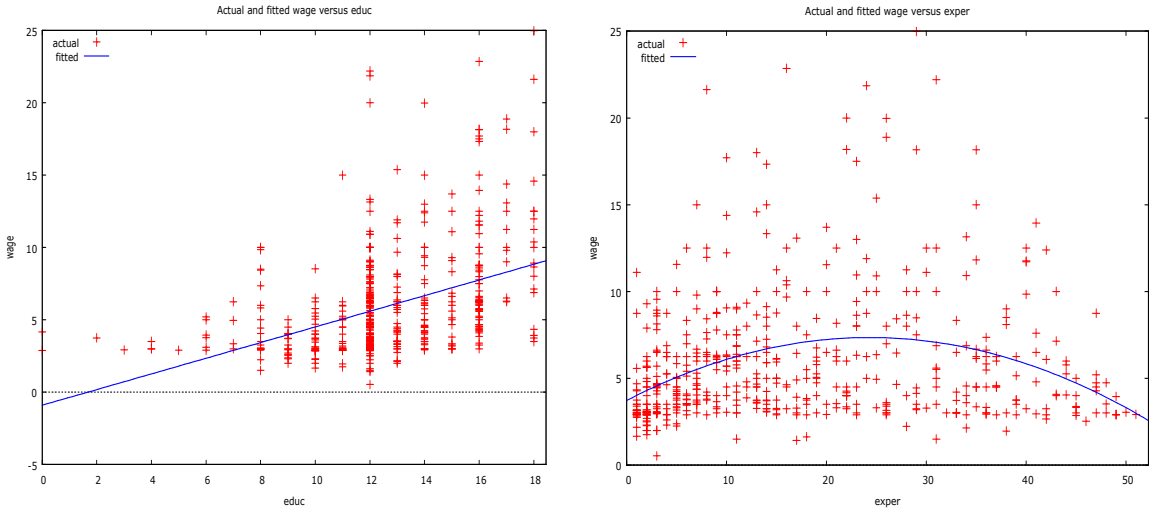


Figure 4.8. We use the WAGE1.txt data set; the linear dependence of *wage* on *educ* (left) is unsatisfactory (probably, *wage* is better described as exp($\beta_0 + \beta_1 educ$)); the parabolic dependence of *wage* on *exper* (right) seems resonable

- **Should we include square and, maybe, cubic terms into the model?**

We shall discuss the RESET (Regression Specification Error Test) test here. If the DGP is correctly described by the model $Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \varepsilon$, then no nonlinear functions of the explanatory variables should be significant when added to the model. Clearly, we could add quadratic terms $X_1^2, ..., X_k^2$ and, probably, mixed terms $X_1 X_2, ..., X_{k-1} X_k$, and then use the $F-$test to test their collective significance but, if k is large, we would loose many degrees of freedom and, thus, accuracy. To use alternative approach, let $\hat{Y}$ denote the OLS fit from our original equation. Consider the expanded equation

[27] Because of the omitted variable bias.

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \delta_2 \hat{Y}^2 + \delta_3 \hat{Y}^3 + \varepsilon\,, \qquad (4.16)$$

where, for example, the term $\hat{Y}^2$, in fact, includes quadratic terms into the model. The null hypothesis now is that the original equation is correctly specified. Thus, RESET is the $F-$ statistic for testing[28] $H_0 : \delta_2 = \delta_3 = 0$ in (4.16). The distribution of $F-$statistic is approximately $F_{2,N-k-3}$ in large samples under the null (and, of course, Gauss-Markov assumptions). An LM version (do you remember it?) is also available (and the respective chi-square distribution will have two df's). The general philosophy of the test is: if we can significantly improve the model by artificially including powers of the predictions of the model, then the original model must have been inadequate.

4.7 example. Let us again consider the hprice.txt data set. In 4.3 example we found that the model $price = \beta_0 + \beta_1 bdrms + \beta_2 lotsize + \beta_3 sqrft + \varepsilon$ is heteroskedastic which was, probably, detected because of the misspecification of the model. In the model window, go to Tests→ Ramsey's RESET→squares only:

```
Auxiliary regression for RESET specification test
OLS, using observations 1-88
Dependent variable: price

             coefficient     std. error    t-ratio   p-value
  ---------------------------------------------------------
  const      237.893        89.2875         2.664    0.0093  ***
  bdrms       -5.40693      10.6454        -0.5079   0.6129
  lotsize     -0.00104059    0.00118493    -0.8782   0.3824
  sqrft       -0.0515997     0.0582962     -0.8851   0.3786
  yhat^2       0.00202011    0.000659320    3.064    0.0029  ***

Test statistic: F = 9.387619,
with p-value = P(F(1,83) > 9.38762) = 0.00295
```

The model is exactly the same as

```
Dependent variable: price

             coefficient     std. error    t-ratio   p-value
  ---------------------------------------------------------
  const      237.893        89.2875         2.664    0.0093  ***
  bdrms       -5.40693      10.6454        -0.5079   0.6129
  lotsize     -0.00104059    0.00118493    -0.8782   0.3824
  sqrft       -0.0515997     0.0582962     -0.8851   0.3786
  sq_yhat1     0.00202011    0.000659320    3.064    0.0029  ***

Log-likelihood      -478.1628   Akaike criterion     966.3255
Schwarz criterion    978.7122   Hannan-Quinn         971.3158
```

which means that the original equations lacks square terms. To be more specific, we shall add square and mixed terms to the model (`bl=bdrms*lotsize` etc):

```
Dependent variable: price
```

[28] How will you test the null H_0 : *the original model is correct* if you include only square term in (4.16)?

```
              coefficient      std. error    t-ratio    p-value
  ------------------------------------------------------------
  const         239.504         90.0438       2.660     0.0095  ***
  bdrms        -104.861         43.0718      -2.435     0.0172  **
  lotsize        -0.0112265      0.00504570  -2.225     0.0290  **
  sqrft           0.149445       0.0686030    2.178     0.0324  **
  sq_bdrms       13.2465         6.00980      2.204     0.0305  **
  sq_lotsize     -1.28981e-07    3.66779e-08  -3.517     0.0007  ***
  sq_sqrft       -1.43239e-06    1.45693e-05  -0.09832   0.9219
  bl              0.00634235     0.00199654    3.177     0.0021  ***
  bs             -0.0167016      0.0132036    -1.265     0.2097
  ls             -3.43006e-07    2.19292e-06  -0.1564    0.8761

Log-likelihood      -457.6175   Akaike criterion     935.2350
Schwarz criterion    960.0084   Hannan-Quinn         945.2156
```

and simplify it (in Model window, go to Tests→Omit variables where check „Sequential elimination of ...“):

```
Model 4, Dependent variable: price

              coefficient     std. error    t-ratio   p-value
  ------------------------------------------------------------
  const        261.163         84.8316       3.079    0.0028   ***
  bdrms        -81.9941        36.5447      -2.244    0.0276   **
  lotsize       -0.00822570     0.00446652  -1.842    0.0692   *
  sqrft          0.0764229      0.0120744    6.329    1.28e-08 ***
  sq_bdrms       7.06968        4.05795      1.742    0.0853   *
  sq_lotsize    -1.17193e-07    1.94210e-08  -6.034    4.57e-08 ***
  bl             0.00515397     0.00105853    4.869    5.46e-06 ***

R-squared            0.810691   Adjusted R-squared   0.796668
F(6, 81)            57.81192    P-value(F)           3.11e-27
Log-likelihood     -458.7423    Akaike criterion     931.4845
Schwarz criterion   948.8259    Hannan-Quinn         938.4709
```

This „final“ model has the smallest Akaike statistic and all significant terms (at 10% significance level). The Koenker test does not show any heteroskedasticity but the model has severe multicollinearity problem which can explain some (which?) „strange“ signs[29] of the coefficients:

```
Variance Inflation Factors
Minimum possible value = 1.0
Values > 10.0 may indicate a collinearity problem

          bdrms    38.345
        lotsize    83.752
          sqrft     1.970
       sq_bdrms    31.547
     sq_lotsize    12.874
             bl    83.659
```

[29] For example, `bdrms` has a negative coefficient, thus, if the number of bedrooms increases, the price (according to our model) decreases. This is because `bdrms` is strongly correlated with other variables which „compensate“ this „inaccuracy“. Also, it is quite possible (because of multicollinearity) that the addition or removal of a few observations will change the sign and value of the coefficient at `bdrms`. Note that the predictive power of the model is quite high – `R-squared` equals 0.81.

Can we cure some of the problems by passing to logarithms? If we apply RESET test to the model

```
Model 5, Dependent variable: lprice

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const      -1.29704      0.651284     -1.992    0.0497    **
  bdrms       0.0369583    0.0275313     1.342    0.1831
  llotsize    0.167967     0.0382811     4.388    3.31e-05  ***
  lsqrft      0.700232     0.0928652     7.540    5.01e-011 ***
Log-likelihood       25.86066   Akaike criterion    -43.72132
Schwarz criterion   -33.81197   Hannan-Quinn        -39.72909
```

we get

```
Auxiliary regression for RESET specification test
OLS, using observations 1-88
Dependent variable: lprice

             coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const       87.8849      240.974       0.3647   0.7163
  bdrms       -0.925329      2.76975    -0.3341   0.7392
  llotsize    -4.18098      12.5952     -0.3319   0.7408
  lsqrft     -17.3491       52.4899     -0.3305   0.7418
  yhat^2       3.91024      13.0143      0.3005   0.7646
  yhat^3      -0.192763      0.752080   -0.2563   0.7984

Test statistic: F = 2.565042,
with p-value = P(F(2,82) > 2.56504) = 0.0831
```

which means that we do not need squared and cubed terms in our Model 5.

- **Which model to choose: log(*Y*) =… or *Y* =…?**

The final question is: which model to choose, the „final“ one for `price` or the one for `lprice`? We can compare directly these two models neither by R^2 nor by AIC (because their lhs's differ) but we can repeat the procedure from Section 3.8: `genr R2 = corr(price, exp(yhat5+sigma_5^2/2))^2` (=0.74)[30] is smaller than $R^2 = 0.81$, but taking into account the fact that the model for `price` has more variables and also the multicollinearity of the model, we stick to the model in logs.

Here we have presented one solution to the question: how to compare two models? When the choice is between the linear and linear-log model, or among the log-linear and double-log specification, things are easy because we have the same dependent variable in each of the two models. So, we can estimate both models and choose the functional form that yields the lower AIC. However, in cases where the dependent variable is not the same, as, for example, in the linear form $Y = \beta_0 + \beta_1 U + \varepsilon$ and log-linear $\log(Y) = \beta_0 + \beta_1 V + \varepsilon$, we cannot directly compare these two models by AIC (in our linear model it equals `931.4845` and in log-log model `-43.72132`).

[30] This scalar is saved in session icon view.

We shall present one more method which allows us to compare such models. The model $\log(price) = \beta_0 + \beta_1 bdrms + \beta_2 \log(lotsize) + \beta_3 \log(sqrft) + \varepsilon$ is equivalent to $price = \exp(\beta_0 + \beta_1 bdrms + \beta_2 \log(lotsize) + \beta_3 \log(sqrft)) \cdot \exp(\varepsilon)$ which is close to the nonlinear model $price = B + B_0 * \exp(\beta_1 bdrms + \beta_2 \log(lotsize) + \beta_3 \log(sqrft)) + \varepsilon$. Nonlinear models are solved through iterative numeric procedures and a usual problem with exponential models is to find proper starting values. It appears that often the coefficients of the log-log model can serve this purpose.

```
ols lprice 0 bdrms llotsize lsqrft
genr B = 100
genr B0 = exp($coeff(0))
genr beta1 = $coeff(bdrms)
genr beta2 = $coeff(llotsize)
genr beta3 = $coeff(lsqrft)
nls price = B + B0*exp(beta1*bdrms+beta2*llotsize+beta3*lsqrft)
    params B B0 beta1 beta2 beta3
end nls
series price_f = $yhat # fitted values
```

```
Convergence achieved after 622 iterations
price = B + B0*exp(beta1*bdrms+beta2*llotsize+beta3*lsqrft)

               estimate       std. error    t-ratio   p-value
  ------------------------------------------------------------
  B           181.268         27.5975        6.568     4.18e-09 ***
  B0            9.47395e-06    3.11154e-05   0.3045    0.7615
  beta1         0.100894       0.0470406     2.145     0.0349   **
  beta2         0.414635       0.0866969     4.783     7.41e-06 ***
  beta3         1.58991        0.331795      4.792     7.15e-06 ***

R-squared            0.769668   Adjusted R-squared   0.758568
Log-likelihood      -467.3725   Akaike criterion     944.7449
Schwarz criterion    957.1316   Hannan-Quinn         949.7352
```

When comparing the final version of AIC, namely 944.7449, with the AIC of the linear model 931.4845, we see again that the linear model is better (in the AIC sense). Thus, if our purpose is to use the model for prediction, both models are of similar quality, but if we want to explain the influence of each variable on `price`, the constant elasticity log-log Model 5 is more transparent and well-defined. ◀◀

More direct methods to choose between $\log Y = \beta_0 + \beta_1 \log X_1 + \beta_2 \log X_2 + u$ and $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + v$ are described in [TH, p.344] or [AH, p.165].

- **Are the model disturbances normally distributed?**

Recall that assumption M4 stated that the disturbancies had to be normally distributed about their zero mean. The assumption is necessary if the inferential aspects of classical regression ($t-$test, $F-$test etc) are to be valid in small samples (for large samples, because of the central limit theorem and law of large numbers, tests will be asymptotically valid even if the disturbances are not normally distributed).

There are many tests for normality[31] of errors. In GRETL, in Model 4 window go to Tests→Normality of residual, where you will see (Fig. 4.10, left) the histogram of residuals and also the χ^2_2-statistic (=17.415) of the Doornik-Hansen test of H_0: *residuals are normal* together with its $p-$value (=0.0002). We can perform three more tests of the H_0 through Variable→Normality test (see the output below) but all they reject normality. However, hprice.txt contains many (88) observations and residuals are rather symmetric (see Fig. 4.9, left) which makes us trust all the significance tests in $H_0: \beta_m = 0$.

```
Test for normality of uhat4:

 Doornik-Hansen test = 17.4152, with p-value 0.000165325
 Shapiro-Wilk W = 0.943844, with p-value 0.000830343
 Lilliefors test = 0.104038, with p-value ~= 0.02
 Jarque-Bera test = 54.0978, with p-value 1.78985e-012
```

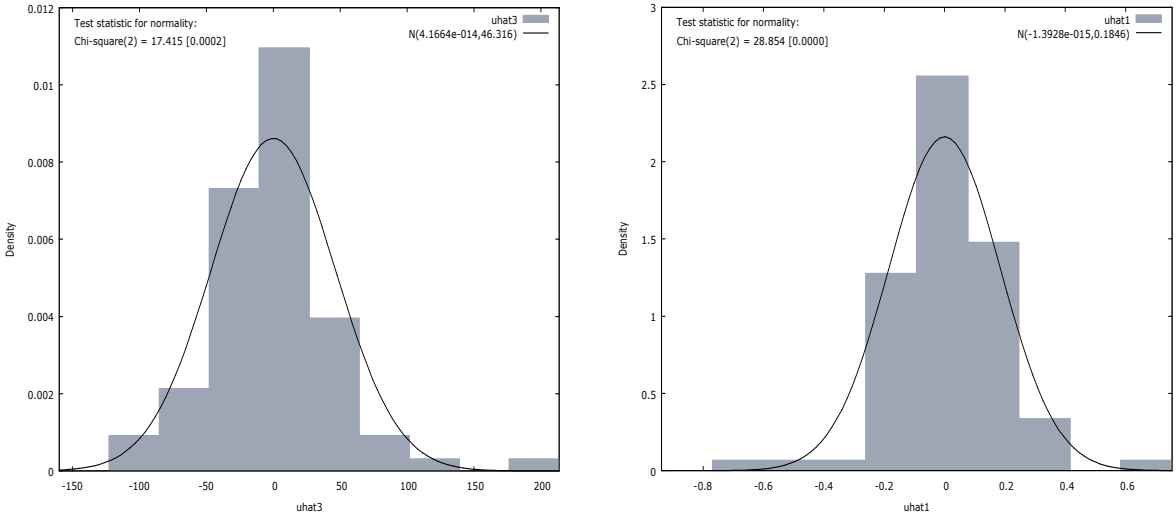


Figure 4.9. Histograms of the residuals of the linear Model 4 (left) and log-log Model 5 (right)

Similar conclusions hold for nonnormality of the residuals in the log-log Model 5 (see Fig. 4.10,right, and the output below).

```
Test for normality of uhat5:

 Doornik-Hansen test = 28.854, with p-value 5.42542e-007
 Shapiro-Wilk W = 0.95184, with p-value 0.00250267
 Lilliefors test = 0.0664989, with p-value ~= 0.43
 Jarque-Bera test = 34.8895, with p-value 2.6536e-008
```

[31] We do not know errors, therefore, as always, we use residuals to test any null hypothesis about errors.

4.11. Instrumental Variables

In this section you will learn how to use instrumental variables to obtain consistent estimates of a model parameters when its independent variables are correlated with the model errors.

Until now, it was assumed that the error terms in the linear regression model $\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\varepsilon}$ satisfy the M2 and M3 conditions, that is, $E(\varepsilon_i \mid \mathbf{X}) = 0$ and, respectively,

$$\text{var}(\vec{\varepsilon} \mid \mathbf{X}) = \text{var}(\vec{\varepsilon}) = \sigma^2 I_N \ . \qquad (4.19)$$

Recall that the term $E(\varepsilon_i \mid \mathbf{X})$ in $E(\vec{\varepsilon} \mid \mathbf{X}) = (E(\varepsilon_1 \mid \mathbf{X}),...,E(\varepsilon_N \mid \mathbf{X}))$ is called a conditional expectation of ε_i given (or assuming that we know all the values in) the design matrix $\mathbf{X}$ (or the whole information contained in $\mathbf{X}$). In our case, Y_i depends on two random variables, observable $\vec{X}_i$ and nonobservable ε_i, thus this conditional expectation describes the relationship between $\vec{X}_i$ and ε_i in some sense (in expectation sense, we do not mention distributions of these two variables anywhere). Conditional expectations have basically the same properties as ordinary expectation, for example, $E(\vec{X}_i + \vec{X}_j \mid \mathbf{X}) = E(\vec{X}_i \mid \mathbf{X}) + E(\vec{X}_j \mid \mathbf{X}) = \vec{X}_i + \vec{X}_j$ (since we know $\mathbf{X}$, we also know $\vec{X}_i$, therefore we can treat it as a known number, that is, $E(\vec{X}_i \mid \mathbf{X}) = \vec{X}_i$). Some new properties of conditional expectations are: 1) if two random variables, for example, ε_i and $\vec{X}_j$ are independent for any i and j or, in other words, the vector $\vec{\varepsilon}$ and $\mathbf{X}$ are independent, then $E(\vec{\varepsilon} \mid \mathbf{X}) = E(\vec{\varepsilon})$ (recall that if the random events *A* and *B* are independent, then $P(A \mid B) = P(A)$), 2) whatever is the random variable U, $E(U) = E\big(E(U \mid \mathbf{X})\big)$ (the double or total expectation rule), and 3) $E(f(\mathbf{X})\,U \mid \mathbf{X}) = f(\mathbf{X})E(U \mid \mathbf{X})$ (once $\mathbf{X}$ is known, $f(\mathbf{X})$ can be treated as a constant). Here are three examples of their application (provided M2, i.e., $E(\varepsilon \mid \mathbf{X}) = 0$, holds):

1) $E(\vec{Y} \mid \mathbf{X}) = E(\mathbf{X}\vec{\beta} \mid \mathbf{X}) + E(\vec{\varepsilon} \mid \mathbf{X}) = \mathbf{X}\vec{\beta}$

2) In univariate case, $E(\hat{\beta}_1 \mid \mathbf{X}) = E(\beta_1 \mid \mathbf{X}) + E\left(\sum \frac{x_i}{\sum x_i^2}\varepsilon_i \mid \mathbf{X}\right) = \beta_1 + \sum \frac{x_i}{\sum x_i^2} E\left(\varepsilon_i \mid \mathbf{X}\right) = \beta_1$,

thus to prove unbiasedness of $\hat{\beta}_m^{OLS}$ it sufficies to require M2 (if we add M3, the OLS estimators of β_m are BLUE&C). If (4.19) is violated (i.e., $\text{var}(\vec{\varepsilon}) = \mathbf{V} \neq \sigma^2 I_N$), then there exist a BLUE&C modification of the OLS method called GLS (Generalized Least Squares) where now the formulas of $\hat{\beta}_m^{GLS}$ contain $\mathbf{V}$ [32](we have already had two examples of $\mathbf{V}$ for the cases

[32] Prior to performing OLS, multiply from the left both sides of the equation $\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\varepsilon}$ by the matrix $\mathbf{P}$ from $\mathbf{P}'\mathbf{P} = \mathbf{V}$; the transformed errors become uncorrelated and homoskedastic.

of heteroskedastic[33] and autoregressive errors). If the matrix $\mathbf{V}$ is unknown and must be estimated, the method is called an FGLS (Feasible GLS).

3) Let $Y = X_1 + X_2 + \varepsilon$ and $EX_1 = EX_2 = E(\varepsilon \mid X_1, X_2) = 0$. Then $EY = 0$, $E(Y \mid X_1) = X_1 + E(X_2 \mid X_1)$, $E(Y \mid X_1, X_2) = X_1 + X_2$ (can you prove all these claims?), and also $E(E(Y \mid X_1, X_2)) = E(X_1 + X_2) = 0$ $(= EY)$.

Let us return to the case where some of the explanatory variables correlate with ε. To simplify matters, consider a univariate case $Y = \beta_0 + \beta_1 X + \varepsilon$. It is easy to verify that $E(\varepsilon \mid X) = 0$ implies $\text{cov}(\varepsilon, X) = 0$. Indeed, $\text{cov}(\varepsilon, X) = E\big((\varepsilon - E\varepsilon)\cdot(X - EX)\big) = E\big(\varepsilon X\big) - EX \cdot E\varepsilon = E\big(XE(\varepsilon \mid X)\big) = 0$. An important claim can be derived from this equality: if ε correlates with X (such an X is called *endogeneous* in the model; an explanatory variable which does not correlate with ε is *exogenous*), then $E(\varepsilon \mid X) \neq 0$ and $E(\hat{\beta}_1 \mid \mathbf{X})$ in 2) above is no longer β_1 (thus $\hat{\beta}_1^{OLS}$ will be biased). Even more so, the bias will not disappear in large samples (thus $\hat{\beta}_1^{OLS}$ is inconsistent). Indeed, as it follows from (3.5), $\hat{\beta}_1 = \beta_1 + \sum \frac{x_i \varepsilon_i}{\sum x_i^2} = \beta_1 + \frac{\widehat{\text{cov}}(X, \varepsilon)}{\widehat{\text{var}} X}$ $\not\to \beta_1$. The implications are very serious: if (in multivariate case) ε correlates with any X_m, then all OLS estimators of the coefficients are no longer BLUE&C. In addition, none of the usual hypothesis testing or interval estimation procedures are valid.

To give an intuitive explanation to the above , consider univariate regression $Y = (\beta_0 + \beta_1 X + \varepsilon =)\ 2 + 0.3\, X + \varepsilon$ where X and ε are positively correlated: $(X, \varepsilon) \sim N(0,0; 3^2, 1^2, (\rho =)\ 0.7)$.

```
library(MASS); set.seed(2);N=100;ro=0.7
Sigma=matrix(c(3^2,3*1*ro,1*3*ro,1^2),2,2)
Sigma
Xeps=mvrnorm(N,c(0,0),Sigma)
X=Xeps[,1]; eps=Xeps[,2]
Y=2+0.3*X+eps  # DGP
plot(X,Y)
mod=lm(Y~X);summary(mod)
abline(2,0.3); abline(mod,lty=2)
legend(-6.5,6,c("true","OLS estimated"),lty=c(1,2))
```

In words: Y depends not only on X, but also on many other variables which reside in ε; if X is positively correlated with any of them, then a unit increase in X will also cause some increase in ε (thus, $\hat{\beta}^{OLS}$ could be, say, twice the true value of β which is to measure the individual effect of X on Y).

[33] If errors are heteroskedastic, $\mathbf{V} = diag(\sigma_1^2, ..., \sigma_N^2)$ and the estimators $\hat{\beta}_m^{GLS}$ are in fact the estimators for weighted variables.

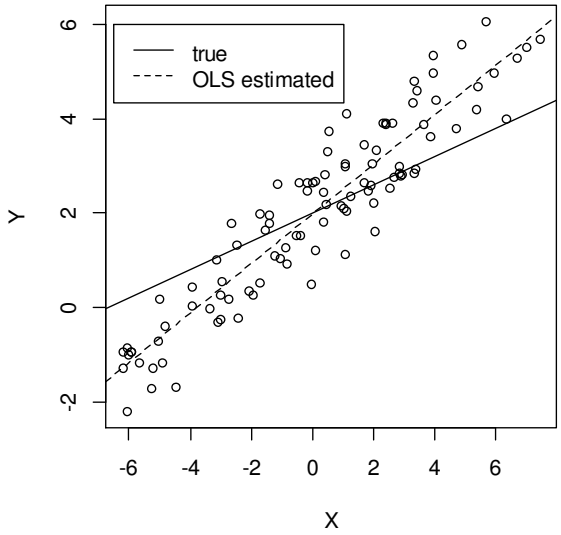


Figure 4.10. Explanatory variable X is correlated with the error ε; the estimated OLS regression line does not go close to the true regression line $Y = \beta_0 + \beta_1 X$.

Since X is correlated with ε, the error term increases together with X and Y goes further and further from the true regression line. On the other hand, the OLS estimates β_1 according to its formulas, therefore, the estimated regression line does not coincide with the true one (the OLS estimator of β_1 is biased and inconsistent). In Computer Labs, 3.9 Example, we explain how to correct the problem.

Now we shall present several cases where X and ε are correlated.

- **Measurement error**

The errors-in-variables problem occurs when an explanatory variable is measured with error. To demonstrate that the variable correlates with the disturbance in this case, consider the following example. Let us assume that an individual's personal saving is based on their "permanent" or long-run income.

A theory of consumer spending which states that people will spend money at a level consistent with their expected long term average income. The level of expected long term income then becomes thought of as the level of "permanent" income that can be safely spent. A worker will save only if his or her current income is higher than the anticipated level of permanent income, in order to guard against future declines in income.

Let Y_i be the annual savings, X_i the permanent annual income of the *i*th worker, and $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ a simple model to represent this relationship. Since the permanent income X is unobservable, we replace it with a *proxy* observable variable X^* = current income,

$X^* = X + u$, $u \sim (0, \sigma_u^2)$. Now $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i = \beta_0 + \beta_1 X_i^* + \varepsilon_i^*$, where $\varepsilon_i^* = (\varepsilon_i - \beta_1 u_i)$, but since $\text{cov}(X_i^*, \varepsilon_i^*) = E((X_i + u_i) \cdot (\varepsilon_i - \beta_1 u_i)) = -\beta_1 \sigma_u^2 \neq 0$ the OLS estimator of β_1 in our observable equation is inconsistent. Later, we shall discuss another, not the OLS, method to estimate β_1 in such a situation.

- **Simultaneous equation bias**

Recall that in a competetive market, the prices and quantities of goods are determined jointly by the forces of supply and demand, i.e., as a solution of the simultaneous system of two equations, one equation for the supply curve and the other equation for the demand curve:

$$\begin{cases} Q_t = (\beta_0^D + \varepsilon_t^D) + \beta_1^D P_t \\ Q_t = (\beta_0^S + \varepsilon_t^S) + \beta_1^S P_t \end{cases}$$

Take a look at the first, demand, equation and assume that ε_2^D is bigger than in previous moment – this implies that the demand curve will be lifted upwards and that now both coordi-

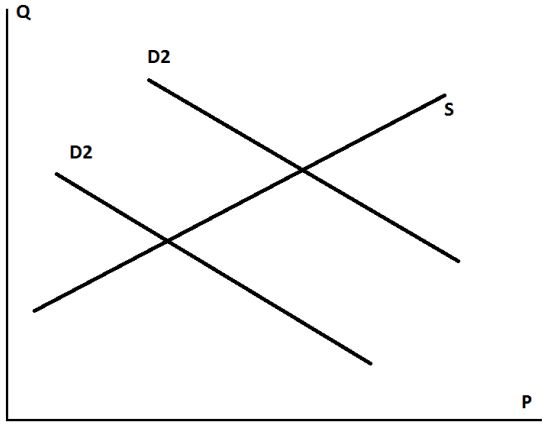


Figure 4.11. The crossing point of the supply and demand curves defines the equilibrium price and quantity

nates of the new point of equilibrium, price and quantity, will increase. This means that the error ε^D and explanatory variable P are positively correlated, the OLS procedure will fail if applied to the equation because of the endogeneity problem; the resulting bias (and inconsistency) is called the *simultaneous equation bias*.

- **Omitted variable**

When an omitted variable is correlated with an included explanatory variable, then the regression error will be correlated with this explanatory variable, making it endogenous. The classic example is from labor economics. A person's wage is determined by in part his or her level of education. Let us specify a log-linear regression model explaining observed hourly wage as[34]

$$\log(WAGE) = \beta_0 + \beta_1 EDUC + \beta_2 EXPER + \beta_3 EXPER^2 + \varepsilon .$$

What else affects wages? Labor economists are most concerned about the omission of a variable measuring ability which may affect the quality of their work and their wage. This variable is a component of the error term, since we usually have no measure for it. The problem is that not only might ability affect wages, but more able individuals may also spend more years in school, causing a positive correlation between ε and the education variable *EDUC*, so that $\text{cov}(EDUC, \varepsilon) > 0$.

****

In the simple linear regression model $Y = \beta_0 + \beta_1 X + \varepsilon$ where X is random and $\text{cov}(X, \varepsilon) = EX\varepsilon \neq 0$, the OLS estimators $\hat{\beta}_m^{OLS}$ are biased and inconsistent. When faced with such a situation, we must consider alternative estimation procedurs. Recall that if the U1-U3 assumptions hold, $\hat{\beta}_m^{OLS}$ can be obtained via the method of moments (see Sect. 3.3):

$$\begin{cases} (0 = E\varepsilon = \bar{\hat{\varepsilon}} =) & (1/N)\sum (Y_i - (b_0 + b_1 X_i)) = 0 \\ (0 = \text{cov}(X, \varepsilon) = \widehat{\text{cov}}(X, \hat{\varepsilon}) =) & (1/N)\sum X_i (Y_i - (b_0 + b_1 X_i)) = 0 \end{cases} \qquad (4.20)$$

Now $0 \neq \text{cov}(X, \varepsilon)$, therefore we cannot use the above system. Suppose, however, that there is another variable Z, called an instrument, such that

i) Z does not have a direct effect on Y (it does not belong to the rhs of the model).
ii) Z is not correlated with ε, it is <u>exogenous</u>.
iii) Z is strongly (or at least not weakly[35]) correlated with X, the endogenous explanatory variable.

In the second equation of (4.20), replace X by Z and solve the system: you will get

$$\begin{cases} \hat{\beta}_1^{IV} = \dfrac{\sum (Z_i - \bar{Z})(Y_i - \bar{Y})}{\sum (Z_i - \bar{Z})(X_i - \bar{X})} \\ \hat{\beta}_0^{IV} = \bar{Y} - \hat{\beta}_1^{IV} \bar{X} \end{cases}$$

[34] Logaritmic transformations are often used for variables that are monetary values such as wages, salaries, income, prices, sales, and expenditures, and in general for variables that measure the „size" of something. These variables have the characteristic that they are positive and often have densities that are positively skewed, with a long tail to the right. Logs of such variables are usually closer to normal.
[35] See a footnote in p. 4-54.

These new, *instrumental variables (IV)*, *estimators* have the following properties:

- They are consistent (very good!)
- However, if the X is exogeneous but we replace it with Z, the variance of the instrumental variables estimator will always be larger than the variance of the OLS estimator (if instrument is weak, *IV* estimation is not reliable).

Another, equivallent to the *IV*, the *two-stage least squares (2SLS)* method can be described as follows: assume that the variable X in $Y = \beta_0 + \beta_1 X + \varepsilon$ is endogenous and Z is its instrument, i.e, $\rho_{XZ} = cor(X, Z) \neq 0$.

Stage 1. Use the OLS method in $X = \delta_0 + \delta_1 Z + u$ and save the fitted values[36] $\hat{X}$ $(= \hat{\delta}_0^{OLS} + \hat{\delta}_1^{OLS} Z)$.

Stage 2. In the original regression equation, replace X by $\hat{X}$ and find ordinary OLS estimators – these are denoted as $\hat{\beta}_1^{2SLS}$. Note that they are exactly the same as $\hat{\beta}_1^{IV}$.

In principle, it would be enough to only use the IV method, however there is a complication – what to do if there were two or more instruments $Z^{(1)}, Z^{(2)}, ...$? In the system of equations

$$\begin{cases} \sum (Y_i - (b_0 + b_1 X_i)) = 0 \\ \sum Z_i^{(1)} (Y_i - (b_0 + b_1 X_i)) = 0 \\ \sum Z_i^{(2)} (Y_i - (b_0 + b_1 X_i)) = 0 \\ ..... \end{cases}$$

we will have two unknowns (b_0 and b_1) and at least three equations, therefore, most probably, the system will be inconsistent. We could remove some redundant equations with Z's but, generally, discarding information is not attractive. It appears that it is *2SLS* which gives us the right solution (in Stage 1, regress X on $Z^{(1)}, Z^{(2)}$ and so on). In general multivariate case, if in

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_l X_l + \gamma_1 X_1^* + ... + \gamma_n X_n^* + \varepsilon \qquad (4.21)$$

$X_1, ..., X_l$ stand for egzogenous variables, $X_1^*, ..., X_n^*$ for endogenous, and $Z_1, ..., Z_p$ (where $p \geq n$ (!)) for instruments, in Stage 1, save the OLS fits $\hat{X}_m^* = \hat{\delta}_0 + \hat{\delta}_1 X_1 + ... + \hat{\delta}_l X_l + \hat{\delta}_{l+1} Z^{(1)} + ... + \hat{\delta}_{l+p} Z^{(p)}$, $m = 1, ..., n$, then, in Stage 2, substitute these fits to (4.21) for X_m^*, $m = 1, ..., n$, and use OLS again. If $p = n$ (one instrument for one endogenous variable), we say

[36] Since $\rho_{XZ} \neq 0$, δ_1 also $\neq 0$ (why?).

that the equation is (exactly) *identified*, if $p > n$ *overidentified*. In both these cases parameters of the model will be consistently estimated.

4.8 example. We use the data on married women in the file mroz.txt to estimate the wage model

$$\log(WAGE) = \beta_0 + \beta_1 EDUC + \beta_2 EXPER + \beta_3 EXPER^2 + \varepsilon .$$

`lfp`	dummy variable = 1 if woman worked in 1975, else 0
`educ`	Wife's educational attainment, in years
`wage`	Wife's 1975 average hourly earnings, in 1975 dollars
`mothereduc`	wife's mother's education level
`fathereduc`	wife's father's education level
`exper`	Actual years of wife's previous labor market experience

Using the N=428 women (out of 753) in the subsample who are in the labor force (Sample→Restrict, based on criterion…→`lfp=1`), the OLS estimates and their standard errors are given by the following regression output table:

```
Dependent variable: l_wage

             coefficient    std. error    t-ratio    p-value
  ---------------------------------------------------------
  const      -0.522041      0.198632      -2.628    0.0089    ***
  educ        0.107490      0.0141465      7.598    1.94e-013 ***
  exper       0.0415665     0.0131752      3.155    0.0017    ***
  sq_exper   -0.000811193   0.000393242   -2.063    0.0397    **
```

We estimate that an additional year of education increases wages approximately 10.75%, ceteris paribus. However, if an unobservable ability, living in ε, has a positive effect on `wage` (i.e., if ε correlates with `educ`), then this estimate is overstated, as the contribution of ability is attributed to the education variable. Therefore, probably, it would be better for the government to redirect the state investment in education and spend tax dollars on, say, social security instead of schools.

We shall use instrumental variables to correctly estimate the influence of `educ` on `l_wage`. A mother's education `mothereduc` does itself not belong in the daughter's wage equation, and it is reasonable to propose that more educated mothers are more likely to have more educated daughters (thus `educ` and `mothereduc` correlate). Another question is whether a woman's ability is correlated with her mother's education (to be a valid instrument, these variables must be uncorrelated). To test the assumption, we shall use the Hausman test later, but, for a while, we assume it to be true.

Stage 1.

```
Dependent variable: educ
```

	coefficient	std. error	t-ratio	p-value	
const	9.77510	0.423889	23.06	7.57e-077	***
exper	0.0488615	0.0416693	1.173	0.2416	
sq_exper	-0.00128106	0.00124491	-1.029	0.3040	
mothereduc	0.267691	0.0311298	8.599	1.57e-016	***

Note that the coefficient of `mothereduc` is very significant, with a $t-$value greater than[37] 3.16. Now we save `educ_hat` and perform

Stage 2.

Dependent variable: l_wage

	coefficient	std. error	t-ratio	p-value	
const	0.198186	0.493343	0.4017	0.6881	
exper	0.0448558	0.0141644	3.167	0.0017	***
sq_exper	-0.000922076	0.000423969	-2.175	0.0302	**
educ_hat	0.0492630	0.0390562	1.261	0.2079	

Note that the coefficient 0.049 is in fact the *IV* coefficient for `educ` (not `educ_hat`) and, surprisingly, not significant. Also, both stages can be performed in one step with Model→Instrumental variables→Two-Stage Least Squares… (see Fig. 4.12).

A few words about the Hausman test whose $p-$value is presented in the model's printout below. Note that, in order to test that, in equation on p. 4-53, `educ` correlates with unobservable ε (this is where the `mothereduc` resides), it makes no sense to calculate the sample correlation between `educ` and observable $\hat{\varepsilon}$ (because, according to the OLS procedure (see (3.8)), this correlation is always 0). This is why the Hausman test takes another approach. The null hypothesis $H_0: X$ *is egzogenous* in $Y = \beta_0 + \beta_1 X + \varepsilon$ (thus, in our case, $H_0:$ `educ` *is egzogenous*) is equivalent to $H_0:$ *the OLS estimate of* β_1 *is consistent* against the alternative $H_1: X$ *is endogenous* (in this latter case we should look for an instrumental variable). The idea of the test is to compare the performance of the LS estimator to an IV estimator. Under the null and alternative hypotheses, we know the following:

- If the null hypothesis is true, both the least squares estimator $\hat{\beta}^{OLS}$ and the instrumental variables estimator $\hat{\beta}^{IV}$ are consistent. Thus, in large samples the difference between them converges to zero. Naturally, if the null hypothesis is true, use the more efficient estimator, which is the least squares estimator.
- If the null hypothesis is false, the least squares estimator is not consistent, and the instrumental variables estimator is consistent. Consequently, the difference between them does not converge to zero in large samples. If the null hypothesis is not true, use the instrumental variables estimator, which is consistent.

[37] The rule of thumb says that if the $t-$value of the candidate to the instrumental variable is greater than $3.16=\sqrt{10}$ (or respective $F-$statistic >10), we can rely on that instrument, it is *strong*.

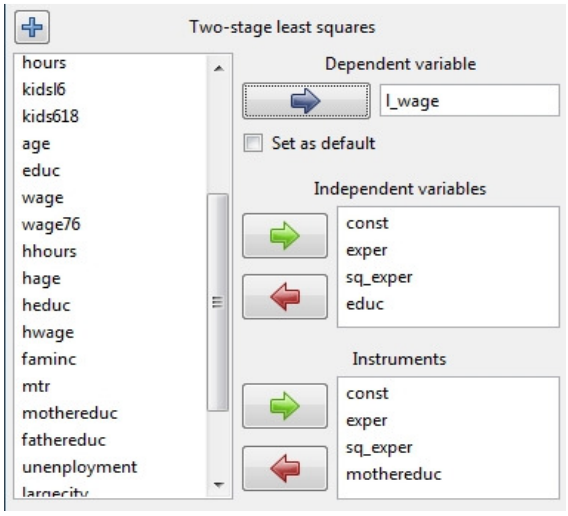


Figure 4.12. Fill in the Instruments box with all the egzogeneous variables plus instruments

```
Dependent variable: l_wage
Instrumented: educ
Instruments: const exper sq_exper mothereduc


             coefficient    std. error       z      p-value
  ---------------------------------------------------------
  const        0.198186      0.472877       0.4191   0.6751
  exper        0.0448558     0.0135768      3.304    0.0010   ***
  sq_exper    -0.000922076   0.000406381   -2.269    0.0233   **
  educ         0.0492630     0.0374360      1.316    0.1882


Hausman test -
  Null hypothesis: OLS estimates are consistent
  Asymptotic test statistic: Chi-square(1) = 3.00338
  with p-value = 0.0830908

Weak instrument test -
  First-stage F-statistic (1, 424) = 73.9459
```

The weak instrument test gives the value of $F-$statistic equal to 8.599^2=73.9459 which is more than 10, thus, `mothereduc` is a very strong instrument. On the other hand, is `educ` correlated with ε? (If it is not, $\hat{\beta}_m^{OLS}$ are consistent and there is no need for instruments.) The $p-$value of the relevant Hausman test is 0.083 and this implies that, in fact, we do not need instruments (at least, with 5% significance). However, note two changes as compared to the original OLS estimates. First, the estimated return to education is 4.93%, which is lower than the OLS estimate of 10.75%. This is consistent with the fact the least squares estimator tends to overestimate the effect of education if `educ` is positively correlated with the omitted factors in the error term. Also notice that the standard error on the coefficient of education (0.0374) is over 2.5 times larger than the standard error reported with the OLS estimates (0.0141). This reflects the fact that even with a good instrumental variable, the IV estimator is not efficient. How can we improve the efficiency of the IV estimator? One of the possibilities is to add more and stronger instruments, namely, we shall add `fathereduc`. Respective model is

```
Dependent variable: l_wage
Instrumented: educ
Instruments: const exper sq_exper mothereduc fathereduc
           coefficient std.error   z       p-value
-----------------------------------------------------------
const       0.0481003  0.400328   0.1202  0.9044
exper       0.0441704  0.0134325  3.288   0.0010 ***
sq_exper   -0.00089897 0.0004017 -2.238   0.0252 **
educ        0.0613966  0.0314367  1.953   0.0508 *

Hausman test -
Null hypothesis: OLS estimates are consistent
Asymptotic test statistic: Chi-square(1) = 2.8256
with p-value = 0.0927721

Sargan over-identification test -
Null hypothesis: all instruments are valid
Test statistic: LM = 0.378071
with p-value = P(Chi-square(1) > 0.378071) = 0.538637

Weak instrument test -
First-stage F-statistic (2, 423) = 55.4003
```

This output means that at least one instrument is strong (F >10), but their usefulness is doubtful. The Sargan test claims that both instruments are valid (if the test rejects, the specification of the model is rejected in the sense that the sample evidence is inconsistent with joint validity of all (in our case, two) moment conditions; without additional information it is not possible to determine which of the population moments is incorrect, i.e., which of the instruments are invalid).

Compare the model with the previous where only `mothereduc` was used as an instrument: the estimate of the return to education increased to 6.14% and the standard error has slightly reduced; note that `educ` is now statistically significant.

4.12. Simultaneous Equations Models

5. DISCRETE RESPONSE MODELS

In this course, we have primarily focused on econometric models in which dependent variable was continuous – quantities, prices, and industrial outputs are example of such variables. However, microeconomics is a general theory of choice, and many of the choices that individuals or firms make cannot be measured by a continuous response variable. If the choice is of "either-or" nature, it can be represented by a binary variable that takes the value 1 in the case of "success" and 0 otherwise. Examples include the following:

- An economic model explaining why some college students decide to study medicine and others do not.
- One can use the mroz.txt data to estimate the labor force participation model.
- The consumer loan default predicting model etc

We shall introduce shortly two models, logit and probit, to describe these cases (we assume that the probability of positive outcome, that is $P(Y=1)$, depends on explanatory variables and our purpose will be to model the dependence). Other examples of discrete response are models for count data, for example,

- The number of children in a household.
- The number of trips to a physician a person makes during a year etc

The response variable in these cases takes the values 0, 1, 2,… and the variable is often satisfactory desribed via the Poisson distribution where its parameter λ may depend on explanatory variables (Poisson regression model).

5.1. Maximum Likelihood Estimation

Strictly speaking, (one-dimensional) population is a distribution function $F(\cdot,\vec{\theta})$ depending on several, usually unknown, parameters $\vec{\theta}$ (for example, normal population with unknown mean μ and variance σ^2, $\vec{\theta}=(\mu,\sigma^2)$, is described by the bell-shaped density function $\varphi(y,\vec{\theta})=\left(1/\sigma\sqrt{2\pi}\right)\exp\left(-(y-\mu)^2/2\sigma^2\right)$, $x\in R$). A collection of N independent r.v. (Y_1, ..., Y_N), each having the same distribution, is called a random sample from respective population; if we take a concrete realization of these r.v., namely, the numbers $(y_1,...,y_N)$, this collection is called a concrete sample (we have already agreed to always use the uppercase letters and to distinguish the samples by a context). Mathematical statistics strives to find functions of a sample (namely, estimators or estimates, denoted as $\hat{\vec{\theta}}$) such that they are "close" in some sense to $\vec{\theta}$. For example, the *method of moments* (MM) suggests to estimate the normal population moments, i.e., the first moment μ and the second central moment σ^2, by its sampling moments:

$$\hat{\mu}^{MM}=\bar{Y}=\sum Y_i/N,\qquad \widehat{\sigma^2}^{MM}=\sum\left(Y_i-\bar{Y}\right)^2/(N-1).$$

The *maximum likelihood* (ML) method is more complicated: it takes the probability (more specifically, density) of the sample $\varphi(Y_1,...,Y_N;\vec{\theta}) = \varphi(Y_1,\vec{\theta})\cdot...\cdot\varphi(Y_N,\vec{\theta})$, treats it as a function of $\vec{\theta}$ and searches for the value of $\vec{\theta}$ which maximizes the *likelihood function* $L(\vec{\theta};Y_1,...,Y_N) = \varphi(Y_1,...,Y_N;\vec{\theta})$ (the maximizing value $\hat{\vec{\theta}}$ is called the ML estimator or estimate). For example, in the normal case

$$L(\mu,\sigma^2;Y_1,...,Y_N) = \left(1/\sigma\sqrt{2\pi}\right)^N \exp\left(-\sum(Y_i-\mu)^2/2\sigma^2\right);$$

to find its maximum, we have to differentiate L with respect to μ and σ^2. However, as a rule, it is easier to differentiate the <u>*logarithmic*</u> *likelihood function* $l = \log L$ which, in normal case, equals to

$$l(\mu,\sigma^2;Y_1,...,Y_N) = -(N/2)\log\sigma^2 - N\log\sqrt{2\pi} - \sum(Y_i-\mu)^2/2\sigma^2 .$$

One can readily verify that in this, normal, case ML estimators (almost) coincide with the MM estimators:

$$\hat{\mu}^{ML} = \bar{Y} = \sum Y_i / N$$
$$\widehat{\sigma^2}^{ML} = \sum\left(Y_i-\bar{Y}\right)^2 / N .$$

More complicated example describes the case where the mean value of Y depends on some external variable X, for example, $Y = \beta_0 + \beta_1 X + \varepsilon, \varepsilon \sim N(0,\sigma^2)$ (this is a univariate regression model and our purpose is to estimate (using the sample $\left((X_1,Y_1),...,(X_N,Y_N)\right)$) three parameters, β_0, β_1 and σ^2). The estimation may be complicated in the case where X is a random variable, but we know that if the model satisfies conditions U1 – U4, then the method of OLS gives the BLUE&C estimators. It is easy to show that OLS <u>in normal case</u> (i.e., where $\vec{Y} \mid \mathbf{X} \sim N(\mathbf{X}\vec{\beta},\sigma_\varepsilon^2\mathbf{I})$) is exactly the same as the method of ML. Indeed, since

$$l(\beta_0,\beta_1,\sigma^2;Y_1,...,Y_N,X_1,...,X_N) = -(N/2)\log\sigma^2 - N\log\sqrt{2\pi} - \sum(Y_i-\beta_0-\beta_1X_i)^2/2\sigma^2 ,$$

to maximize l is the same as minimize $\sum(Y_i-\beta_0-\beta_1X_i)^2$.

More to the point is the example of the Bernoulli r.v. which takes two values:

$$Y = \begin{cases} 1 & \text{with probability } p, \\ 0 & \text{with probability } q\ (=1-p). \end{cases}$$

Relevant random sample is $(Y_1,...,Y_N)$ where each Y_i is the Bernoulli r.v. (concrete sample is a collection of 0's and 1's and obviously, if p is closer to 1, the sample contains more unities).

To estimate the probability p $(= P(Y = 1))$, we can use either MM or ML. Since $p = EY$, the MM proposes $\hat{p}^{MM} = \overline{Y}$ ($\overline{Y}$ is the *relative frequency* of successes). The ML method is based on the equality $P(Y) = p^Y (1-p)^{1-Y}$ which implies that the probability of the sample equals $L(p; Y_1, ..., Y_N) = p^{S_N} (1-p)^{N-S_N}$ where $S_N = Y_1 + ... + Y_N$. Thus, to maximize $l(p; Y_1, ..., Y_N) = S_N \log p + (N - S_N) \log(1-p)$, we differentiate l with respect to p:

$$l'_p = \frac{S_N}{p} - \frac{N - S_N}{1-p} = 0$$

and obtain

$$\hat{p}^{ML} = \frac{S_N}{N} \left(= \overline{Y}\right)^1$$

(note that this ML estimator of p is in no way connected with least squares). In general, ML estimators have some nice properties (in „regular" cases they are asymptotically unbiased, effective and consistent), therefore they are widely used.

5.2. Binary Response Variable

So far, to describe the dependence of Y on explanatory variables we mostly used a linear regression model $Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \varepsilon$. However, in the case where Y attains only two values, 0 and 1, the model is unsatisfactory. To explain this, consider the date set coke.txt which describes a population of customers who bought either coke or pepsi:

Y	= coke	=1 if coke chosen, =0 if pepsi chosen
X_1	= pratio	price coke relative to price pepsi
X_2	= pr_pepsi	price of 2 liter bottle of pepsi
X_3	= pr_coke	price of 2 liter bottle of coke
X_4	= disp_pepsi	= 1 if pepsi is displayed at time of purchase, otherwise = 0
X_5	= disp_coke	= 1 if coke is displayed at time of purchase, otherwise = 0

Our purpose is to establish how the chances (that is, probability) to buy coke depends on the available explanatory variables or, in other words, we want to create a model $p = P(Y = 1) = f(X_1, ..., X_k)$. For simplicity, we analyse the univariate model first, assume that the conditions U1-U3 hold true and consider the model

$$\texttt{coke} = \beta_0 + \beta_1 \texttt{pratio} + \varepsilon,$$

[1] Once again, MM estimator coincides with the ML estimator, but, in general, this is rarely true.

where, taking into account the fact that `coke` equals 1 or 0,

$$\varepsilon = \begin{cases} 1-(\beta_0+\beta_1\texttt{pratio}) & \text{with probability } \beta_0+\beta_1\texttt{pratio}, \\ -(\beta_0+\beta_1\texttt{pratio}) & \text{with probability } 1-(\beta_0+\beta_1\texttt{pratio}). \end{cases}$$

Note that $E(\varepsilon\,|\,\texttt{pratio}) \equiv 0$ and $\big(E(\texttt{coke}\,|\,\texttt{pratio}) =\big)\ P(\texttt{coke}=1\,|\,\texttt{pratio}) = \beta_0 + \beta_1\ \texttt{pratio}$. The problem with this *linear probability model* is that the variance of ε depends on `pratio`:

$$\text{var}(\varepsilon) = (\beta_0+\beta_1\texttt{pratio})\cdot(1-(\beta_0+\beta_1\texttt{pratio}))\,,$$

thus the model is heteroskedastic. Still bigger problem is that $\beta_0+\beta_1\ \texttt{pratio}$ can be less than 0 or bigger than 1 what contradicts the properties of probabilities. Thus we have to look for another model and, clearly, any distribution function F will serve our purpose well:

$$\texttt{coke} = F(\beta_0+\beta_1\ \texttt{pratio})+\varepsilon.$$

The most popular distribution functions in this context are logistic $F(x)=\Lambda(x)=\exp(x)/(\exp(x)+1)$, $x\in R$, and normal $F(x)=\Phi(x)=(2\pi)^{-1/2}\int_{-\infty}^{x}\exp\left(-z^2/2\right)dz$, $x\in R$. Both functions have similar shapes (see Fig. 5.1, left), therefore we shall study only the logistic function in more detail. The rhs of the equation $\texttt{coke}=\Lambda(\beta_0+\beta_1\ \texttt{pratio})+\varepsilon$ is now always between 0 and 1, however, the errors are not normal and, also, this nonlinear[2] regression model is heteroskedastic again, therefore, $\hat\beta_0^{NLS}$ and $\hat\beta_1^{NLS}$ are ineffective (recall that weighting could help us here). On the other hand, the ML method is in many aspects best, so we shall apply it. The likelihood function now is

$$L(\beta_0,\beta_1;Y_1,...,Y_N,X_1,...,X_N)=\prod\Big(\Lambda(\beta_0+\beta_1X_i)\Big)^{Y_i}\cdot\big(1-\Lambda(\beta_0+\beta_1X_i)\big)^{1-Y_i};$$

the parameters β_0 and β_1 are estimated by maximizing this expression, which is highly nonlinear in the parameters and cannot be estimated by conventional regression programs (but both GRETL and R have the necessary tools). The probability model

$$p=\Lambda(\beta_0+\beta_1X)=\exp(\beta_0+\beta_1X)/\big(1+\exp(\beta_0+\beta_1X)\big)$$

is called the *logit* model or, if one replaces Λ by Φ, *probit* model. The logit model can be also rewritten in a linear form

$$\log\frac{p}{1-p}=\beta_0+\beta_1X$$

[2] We call the respective model `nls-w`.

where the *link* function $\log p/(1-p)$ is again called a logit function (can you draw its graph for $0<p<1$?) or the *log-odds ratio*[3]. The expression

$$\hat{p} = \hat{P}(\texttt{coke}=1) = \Lambda(\hat{\beta}_0^{ML} + \hat{\beta}_1^{ML}\,\texttt{pratio})$$

allows us to estimate the probability that a shopper will choose to buy Coke for a given `pratio`.

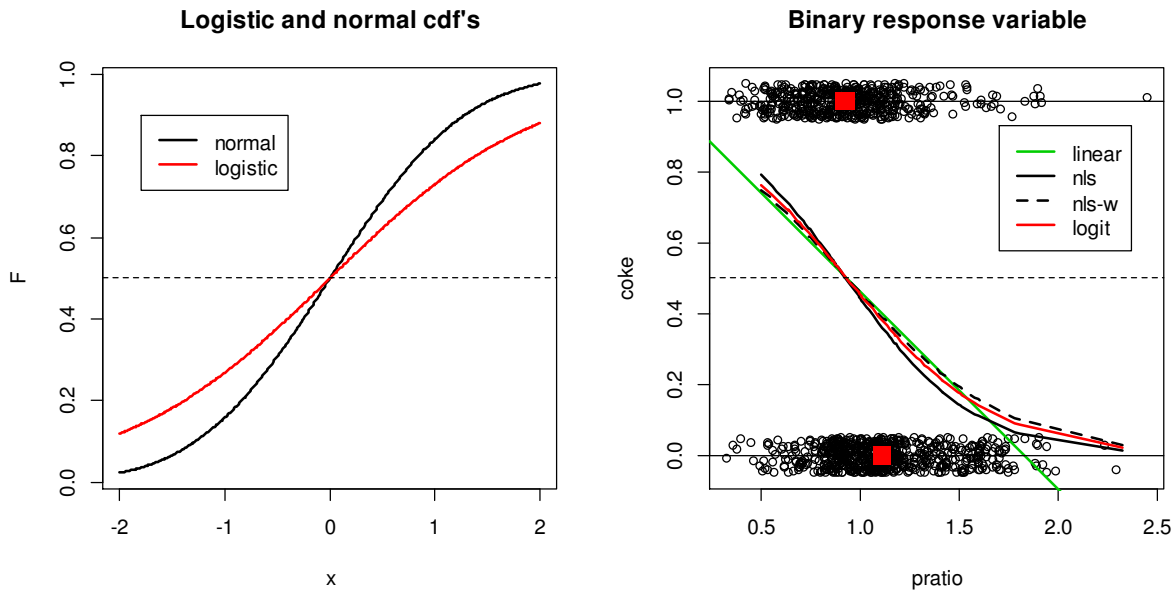


Figure 5.1. Standard normal and logistic distribution functions (left) and different probability models for `coke` (right) (the script is in Computer Labs)

The variance of the (standard) logistic distribution function equals $\pi^2/3$ as compared to 1 in standard normal case, therefore the estimates of β_m obtained from the logit model have to be multiplied by $\sqrt{3}/\pi$ to be comparable with to the estimates obtained from the probit model (usually both models give almost the same coefficients after this correction). The graphs in Fig. 5.1, right, are those of different models of $\hat{P}(\texttt{coke}=1)$. The linear model (green line) takes negative values for sufficiently big values of `pratio`, therefore it is unsatisfactory. On the other hand, the weighted logistic model `nls-w` is almost as good as logit model but the latter one is easier to apply.

In fact, the curves themselves are not of big interest, more important is the prediction rule of the response value (for what value of `pratio` the customer will buy Coke?) and also the quality of the forecast. One of the goodness-of-fit measures is the *percent correctly predicted*. Define a binary predictor of `coke` to be one (we predict, a shopper will buy coke but not pepsi) if the predicted probability is at least 0.5 and zero otherwise. There are four possible out-

[3] For example, if the probability of a household possessing an automobile is $p=0.75$ then the odds ratio, i.e., the ratio $p/(1-p)$, is 0.75/0.25 = 3/1, or odds of three to one that a car is possessed. What is the odds ratio if $p=1/10$?

comes on each pair $(Y_i, \hat{Y}_i)$: when both are zero or both are one, we make the correct prediction. The „Number of cases ‚correctly predicted' " is the percentage of times that $Y_i = \hat{Y}_i$ (in our example it is 66.2%)

```
Model 1: Logit, using observations 1-1140
Dependent variable: coke

             coefficient   std. error      z        p-value      slope
  ------------------------------------------------------------------------
  const        2.52508      0.271574      9.298    1.43e-020
  pratio      -2.71081      0.266631    -10.17     2.79e-024    -0.666411

McFadden R-squared    0.082656   Adjusted R-squared    0.080105
Log-likelihood       -719.0694   Akaike criterion      1442.139
Schwarz criterion     1452.216   Hannan-Quinn          1445.945

Number of cases 'correctly predicted' = 755 (66.2%)
f(beta'x) at mean of independent vars = 0.246
Likelihood ratio test: Chi-square(1) = 129.582 [0.0000]

            Predicted
                0     1
  Actual 0    508   122
         1    263   247
```

The threshold probability of 0.5 is rather arbitrary (for example, a bank issues a loan only if the predicted probability of default is less than 0.05). Interestingly, all four probability curves in Fig. 5.1 cross the 0.5 level at the same point of `pratio`=0.928, thus all they have the same percent correctly predicted and the same „Actual-Predicted" table.

There are also variuos *pseudo R-squared* measures for binary response. McFadden suggests the measure $1 - l_{UR} / l_0$, where l_{UR} is the log-likelihood function for the estimated (unrestricted) model, and l_0 is the log-likelihood function in the model with only an intercept. If the explanatory variables have no explanatory power, then pseudo R-squared is zero, just as the usual R-squared ($1 - l_{UR} / l_0$ cannot reach unity but, anyway, the more the better). The usual considerations for Akaike and Schwarz criteria also hold.

Before passing to a more complex model, we shall comment other estimates in the above table. In the usual linear model $E(Y \mid X) = \beta_0 + \beta_1 X$, the meaning of the coefficient β_1 is the slope of the regression line or the marginal effect of X on Y: $dE(Y \mid X) / dX = \beta_1$ (note that the effect is the same for any X; what is the meaning of β_1?). Now the *slope* is estimated as $dE(Y \mid X) / dX = \lambda(\beta_0 + \beta_1 X) \cdot \beta_1$ where $\lambda(\cdot) = \Lambda'(\cdot)$ is the density[4] of the logistic distribution function (clearly, now the slope varies with the values of X). The factor β_1 has no particular meaning here but $\lambda(\beta_0 + \beta_1 X) \cdot \beta_1$ estimates the change of $P(Y = 1)$ when X increases to $X + 1$. In interpreting the estimated model, it is useful to calculate this value at, say, the mean of the regressors. For convenience, it is also worth noting that the same *scale factor* (i.e.,

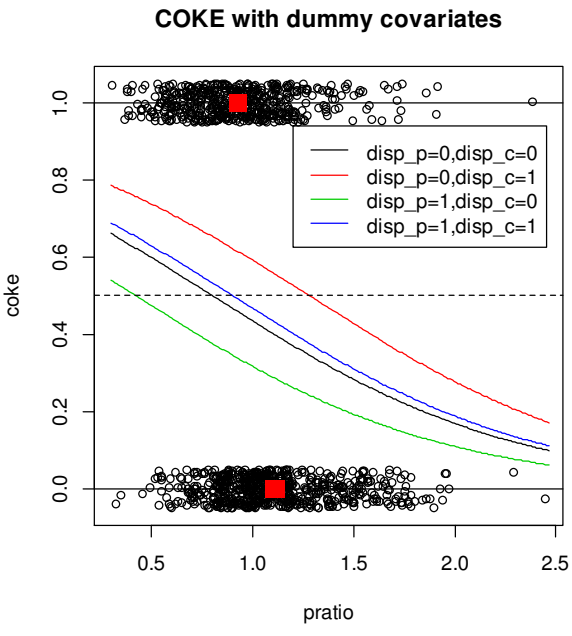
[4] Can you calculate $\lambda(\beta_0 + \beta_1 X)$?

f(beta'x)=) $\lambda(\beta_0+\beta_1 X_1+...)$ applies to all the slopes in the multivariate model (see below).

Now we shall upgrade the model by including all the explanatory variables. After removing insignificant pr_coke, we arrive at the model

```
Model 3: Logit, using observations 1-1140
Dependent variable: coke
Standard errors based on Hessian

              coefficient   std. error      z      p-value    slope
  ------------------------------------------------------------------
  const         -0.326933     0.724709   -0.4511   0.6519
  disp_pepsi    -0.504248     0.179239   -2.813    0.0049   -0.122407
  disp_coke      0.633881     0.179631    3.529    0.0004    0.156073
  pratio        -1.32550      0.365406   -3.627    0.0003   -0.326440
  pr_pepsi       1.15291      0.335294    3.439    0.0006    0.283936

Mean dependent var   0.447368   S.D. dependent var   0.497440
McFadden R-squared   0.102630   Adjusted R-squared   0.096251
Log-likelihood      -703.4127   Akaike criterion     1416.825
Schwarz criterion    1442.019   Hannan-Quinn         1426.340

Number of cases 'correctly predicted' = 767 (67.3%)
f(beta'x) at mean of independent vars = 0.246
Likelihood ratio test: Chi-square(4) = 160.895 [0.0000]

            Predicted
               0     1
  Actual 0   505   125
         1   248   262
```

The computation of the derivatives $\lambda(\beta_0+\beta_1 X_1+...)\cdot\beta_i$ of the conditional mean function is useful when the variable in question is continuous and, also, often produces a reasonable approximation for a dummy variable (our model contains two dummy variables, disp_pepsi and disp_coke). Another way to analyze the effect of a dummy variable on the whole distribution is to compute $P(Y=1)$ over the range of $\hat{\beta}_0+\hat{\beta}_1 X_1+...$ and with the two values of the binary variable. Using the coefficients from our table, we have the following probabilities as a function of pratio, at the mean of pr_pepsi:



disp_p=0, disp_c=0:

$\hat{P}(\text{coke}=1)=\Lambda$ (-0.3269-1.3255*pratio+ 1.1529*mean(pr_pepsi))

disp_p=0, disp_c=1:

$\hat{P}(\text{coke}=1)=\Lambda$ (-0.3269-1.3255*pratio+ 1.1529*mean(pr_pepsi)+0.6339)

etc. The marginal effect of any dummy variable is the difference between two respective curves. For example, if disp_p=0 and disp_c=1, the

probability $P(Y=1)$ is described by the red curve. Similarly, if disp_p=1 and disp_c=0, the probability is depicted by the green curve. The difference between the curves around the mean value of `pratio` (=1.027) is close to 0.27 thus, more specifically, in the first case 58% of all the customers will buy coke whereas in the second only 31%. If the store charges the wholesaler for exposing coke, the wholesaler can estimate whether it is worthwhile to do this.

5.3. Generalizations

5.3.1. Multinomial Logit

In probit and logit models, the decision maker chooses between two alternatives. Clearly we are often faced with choices involving more than two alternatives. These are called multinomial choice situations. Examples include the following:

1. If you are shopping for a laundry detergent, which one do you choose? Tide, Ariel, Rex, and so on. The consumer is faced with a wide array of alternatives. Marketing researchers relate these choices to prices of the alternatives, advertising, and product characteristics.
2. If you enroll in the Faculty of mathematics and informatics, will you major in econometrics, pure or applied mathematics, bioinformatics, or programming systems?

In each of these cases, we wish to relate the observed choice to a set of explanatory variables. More specifically, as in probit and logit models, we wish to explain and predict the probability that an individual with a certain set of characteristics chooses one of the alternatives. The estimation and interpretation of such models is, in principle, similar to that in logit and probit models. The models themselves go under the names *multinomial logit* or *conditional logit*.

5.3.2. Ordered Choice Models

The choice options in multinomial and conditional logit models have no natural ordering or arrangement. However, in some cases choices are ordered in a specific way. Here are some examples:

1. Results of opinion surveys in which responses can be strongly in disagreement, in disagreement, neutral, in agreement, or strongly in agreement.
2. Assignment of grades or work performance ratings. Students receive grades A, B, C, D, and F, which are ordered on the basis of a teacher's evaluation of their performance. Employees are often given evaluations on scales such as Outstanding, Very Good, Good, Fair, and Poor, which are similar in spirit.
3. Standard and Poor's rates bonds as AAA, AA, A, BBB, and so on, as a judgment about the credit worthiness of the company or country issuing a bond, and how risky the investment might be.

When modeling these types of outcomes, numerical values are assigned to the outcomes, but the numerical values are ordinal and reflect only the ranking of the outcomes. In the first example, we might assign a dependent variable y the values

$$y = \begin{cases} 1 \text{ strongly disagree} \\ 2 \text{ disagree} \\ 3 \text{ neutral} \\ 4 \text{ agree} \\ 5 \text{ strongly agree} \end{cases}$$

The usual linear regression model is not appropriate for such data, because in regression we would treat the y values as having some numerical meaning when they do not. Estimation, as with previous choice models, is by maximum likelihood and the model itself is called *ordered probit* or *logit model*.

5.3.3. Models for Count Data

When the dependent variable in a regression model is a count of the number of occurrences of an event, the outcome variable is $y = 0, 1, 2, 3, \ldots$ These numbers are actual counts, and thus different from the ordinal numbers of the previous section. Examples include the following:

1. The number of trips to a physician a person makes during a year (it could depend, for example, on the age of the person and general condition).
2. The number of children in a household (it could depend on whether the family lives in town or countryside, household income, house or apartment etc).
3. The number of automobile accidents at a particular intersection during a month (it could depend on the week day).
4. The number of awards earned by a student at one high school. Predictors of the number of awards earned include the type of program in which the student was enrolled (e.g., general or academic), the score on final exam in math, town or countryside etc.

While we are again interested in explaining and predicting probabilities, such as the probability that an individual will take two or more trips to the doctor during a year, the probability distribution we use as a foundation is the Poisson, not the normal or the logistic. If Y is a Poisson random variable, then its probability function is

$$P(Y = y) = \frac{\mathrm{e}^{-y}\,\lambda^{y}}{y!},\ y = 0,1,2,...,\ \lambda > 0.$$

Recall that $EY = \operatorname{var} Y = \lambda$. In a regression model, we try to explain the behavior of EY as a function of some explanatory variables. We do the same here, keeping the value of $EY \geq 0$ by defining $E(Y \mid X) = \lambda = \exp(\beta_0 + \beta_1 X)$ (≥ 0). This choice defines the Poisson regression model for count data. More specifically, assume that we have a sample $\{(Y_i, X_i), Y_i \in \{0,1,2,...\}, i = 1,...,N\}$, $P(Y_i = y \mid X_i; \vec{\beta}) = e^{-E(Y_i|X_i)} \cdot (E(Y_i \mid X_i))^{y} / y!$, $y = 0,1,2,...$, and the likelihood function equals

$$L(\beta_0, \beta_1 \mid \vec{Y}, \vec{X}) = \prod_{i=1}^{N} e^{-\exp(\beta_0+\beta_1 X_i)} \left(e^{\beta_0+\beta_1 X_i}\right)^{Y_i} / Y_i!$$

By the method of maximum likelihood, we wish to find the set of parameters $\vec{\beta}$ that makes this probability as large as posible. A formula of the above type is difficult to work with therefore, as usual, we take logarithms etc. One worked example of the Poisson regression can be found in http://www.ats.ucla.edu/stat/r/dae/poissonreg.htm.

REFERENCES

[AH] Asteriou D., Hall S. Applied Econometrics (Revised Ed.), Palgrave Macmillan, 2007

[HGL] Hill R.C., Griffiths W.E., Lim G.C. Principles of Econometrics, 4th Ed., Wiley, 2012

[M] Maddala G.S. Introduction to Econometrics, 3rd Ed., 2005

[L] Lapinskas R. A Very Short Introduction to Statistics with gretl, http://uosis.mif.vu.lt/~rlapinskas/ShortStatGRETL/

[T] Thomas R.L. Modern Econometrics, An Introduction. Prentice Hall, 1997