Statistics 512: Applied Linear Models

Topic 3

Topic Overview

This topic will cover

- thinking in terms of matrices
- regression on multiple predictor variables
- case study: CS majors
- Text Example (NKNW 241)

Chapter 5: Linear Regression in Matrix Form

The SLR Model in Scalar Form

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad \text{where} \quad \epsilon_i \sim^{iid} N(0, \sigma^2)$$

Consider now writing an equation for each observation:

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_1 + \epsilon_1 \\ Y_2 &= \beta_0 + \beta_1 X_2 + \epsilon_2 \\ \vdots &\ \vdots\ \ \vdots \\ Y_n &= \beta_0 + \beta_1 X_n + \epsilon_n \end{aligned}$$

The SLR Model in Matrix Form

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 X_1 \\ \beta_0 + \beta_1 X_2 \\ \vdots \\ \beta_0 + \beta_1 X_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

(I will try to use **bold** symbols for matrices. At first, I will also indicate the dimensions as a subscript to the symbol.)

- $\mathbf{X}$ is called the design matrix.
- β is the vector of parameters
- ϵ is the error vector
- $\mathbf{Y}$ is the response vector

The Design Matrix

$$\mathbf{X}_{n\times 2} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}$$

Vector of Parameters

$$\beta_{2\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

Vector of Error Terms

$$\epsilon_{n\times 1} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Vector of Responses

$$\mathbf{Y}_{n\times 1} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

Thus,

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\beta + \epsilon \\ \mathbf{Y}_{n\times 1} &= \mathbf{X}_{n\times 2}\beta_{2\times 1} + \epsilon_{n\times 1} \end{aligned}$$

Variance-Covariance Matrix

In general, for any set of variables $U_1, U_2, \ldots, U_n$, their *variance-covariance matrix* is defined to be

$$\sigma^2\{\mathbf{U}\} = \begin{bmatrix} \sigma^2\{U_1\} & \sigma\{U_1, U_2\} & \cdots & \sigma\{U_1, U_n\} \\ \sigma\{U_2, U_1\} & \sigma^2\{U_2\} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \sigma\{U_{n-1}, U_n\} \\ \sigma\{U_n, U_1\} & \cdots & \sigma\{U_n, U_{n-1}\} & \sigma^2\{U_n\} \end{bmatrix}$$

where $\sigma^2\{U_i\}$ is the variance of U_i, and $\sigma\{U_i, U_j\}$ is the covariance of U_i and U_j.
When variables are uncorrelated, that means their covariance is 0. The variance-covariance matrix of uncorrelated variables will be a *diagonal* matrix, since all the covariances are 0.

Note: Variables that are independent will also be uncorrelated. So when variables are correlated, they are automatically dependent. However, it is possible to have variables that are dependent but uncorrelated, since correlation only measures *linear* dependence. A nice thing about normally distributed RV's is that they are a convenient special case: if they are uncorrelated, they are also independent.

Covariance Matrix of ϵ

$$\sigma^2\{\epsilon\}_{n\times n} = Cov \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} = \sigma^2\mathbf{I}_{n\times n} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

Covariance Matrix of Y

$$\sigma^2\{\mathbf{Y}\}_{n\times n} = Cov \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \sigma^2\mathbf{I}_{n\times n}$$

Distributional Assumptions in Matrix Form

$\epsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I})$
$\mathbf{I}$ is an $n \times n$ identity matrix.

- Ones in the diagonal elements specify that the variance of each ϵ_i is 1 times σ^2.
- Zeros in the off-diagonal elements specify that the covariance between different ϵ_i is zero.
- This implies that the correlations are zero.

Parameter Estimation

Least Squares

Residuals are $\epsilon = \mathbf{Y} - \mathbf{X}\beta$. Want to minimize sum of squared residuals.

$$\sum \epsilon_i^2 = [\epsilon_1\, \epsilon_2\, \cdots \epsilon_n] \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} = \epsilon'\epsilon$$

We want to minimize $\epsilon'\epsilon = (\mathbf{Y}-\mathbf{X}\beta)'(\mathbf{Y}-\mathbf{X}\beta)$, where the "prime" $()'$ denotes the transpose of the matrix (exchange the rows and columns).
We take the derivative with respect to the vector β. This is like a quadratic function: think "$(\mathbf{Y} - \mathbf{X}\beta)^2$".
The derivative works out to 2 times the derivative of $(\mathbf{Y} - \mathbf{X}\beta)'$ with respect to β.
That is, $\frac{d}{d\beta}\left((\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)\right) = -2\mathbf{X}'(\mathbf{Y} - \mathbf{X}\beta)$. We set this equal to $\mathbf{0}$ (a vector of zeros), and solve for β.
So, $-2\mathbf{X}'(\mathbf{Y} - \mathbf{X}\beta) = \mathbf{0}$. Or, $\mathbf{X}'\mathbf{Y} = \mathbf{X}'\mathbf{X}\beta$ (the "normal" equations).

Normal Equations

$$\mathbf{X}'\mathbf{Y} = (\mathbf{X}'\mathbf{X})\beta$$

Solving this equation for β gives the least squares solution for $\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$.
Multiply on the left by the inverse of the matrix $\mathbf{X}'\mathbf{X}$. (Notice that the matrix $\mathbf{X}'\mathbf{X}$ is a 2×2 square matrix for SLR.)

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

REMEMBER THIS.

Reality Break:

This is just to convince you that we have done nothing new nor magic – all we are doing is writing the same old formulas for b_0 and b_1 in matrix format. Do NOT worry if you cannot reproduce the following algebra, but you SHOULD try to follow it so that you believe me that this is really not a new formula.
Recall in Topic 1, we had

$$\begin{aligned} b_1 &= \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \equiv \frac{SS_{XY}}{SS_X} \\ b_0 &= \bar{Y} - b_1\bar{X} \end{aligned}$$

Now let's look at the pieces of the new formula:

$$\begin{aligned}
\mathbf{X'X} &= \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_1 & X_2 & \cdots & X_n \end{bmatrix} \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} = \begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix} \\
(\mathbf{X'X})^{-1} &= \frac{1}{n\sum X_i^2 - (\sum X_i)^2} \begin{bmatrix} \sum X_i^2 & -\sum X_i \\ -\sum X_i & n \end{bmatrix} = \frac{1}{nSS_X} \begin{bmatrix} \sum X_i^2 & -\sum X_i \\ -\sum X_i & n \end{bmatrix} \\
\mathbf{X'Y} &= \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_1 & X_2 & \cdots & X_n \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \sum Y_i \\ \sum X_i Y_i \end{bmatrix}
\end{aligned}$$

Plug these into the equation for $\mathbf{b}$:

$$\begin{aligned}
\mathbf{b} &= (\mathbf{X'X})^{-1}\mathbf{X'Y} = \frac{1}{nSS_X} \begin{bmatrix} \sum X_i^2 & -\sum X_i \\ -\sum X_i & n \end{bmatrix} \begin{bmatrix} \sum Y_i \\ \sum X_i Y_i \end{bmatrix} \\
&= \frac{1}{nSS_X} \begin{bmatrix} (\sum X_i^2)(\sum Y_i) - (\sum X_i)(\sum X_i Y_i) \\ -(\sum X_i)(\sum Y_i) + n\sum X_i Y_i \end{bmatrix} \\
&= \frac{1}{SS_X} \begin{bmatrix} \bar{Y}(\sum X_i^2) - \bar{X}\sum X_i Y_i \\ \sum X_i Y_i - n\bar{X}\bar{Y} \end{bmatrix} \\
&= \frac{1}{SS_X} \begin{bmatrix} \bar{Y}(\sum X_i^2) - \bar{Y}(n\bar{X}^2) + \bar{X}(n\bar{X}\bar{Y}) - \bar{X}\sum X_i Y_i \\ SP_{XY} \end{bmatrix} \\
&= \frac{1}{SS_X} \begin{bmatrix} \bar{Y}SS_X - SP_{XY}\bar{X} \\ SP_{XY} \end{bmatrix} = \begin{bmatrix} \bar{Y} - \frac{SP_{XY}}{SS_X}\bar{X} \\ \frac{SP_{XY}}{SS_X} \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix},
\end{aligned}$$

where

$$\begin{aligned}
SS_X &= \sum X_i^2 - n\bar{X}^2 = \sum (X_i - \bar{X})^2 \\
SP_{XY} &= \sum X_i Y_i - n\bar{X}\bar{Y} = \sum (X_i - \bar{X})(Y_i - \bar{Y})
\end{aligned}$$

All we have done is to write the same old formulas for b_0 and b_1 in a fancy new format. See NKNW page 200 for details. Why have we bothered to do this? The cool part is that the same approach works for multiple regression. All we do is make $\mathbf{X}$ and $\mathbf{b}$ into bigger matrices, and use exactly the same formula.

Other Quantities in Matrix Form

Fitted Values

$$\hat{\mathbf{Y}} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix} = \begin{bmatrix} b_0 + b_1 X_1 \\ b_0 + b_1 X_2 \\ \vdots \\ b_0 + b_1 X_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \mathbf{Xb}$$

Hat Matrix

$$\begin{aligned}\hat{\mathbf{Y}} &= \mathbf{Xb} \\ \hat{\mathbf{Y}} &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ \hat{\mathbf{Y}} &= \mathbf{HY}\end{aligned}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. We call this the "hat matrix" because is turns Y's into $\hat{Y}$'s.

Estimated Covariance Matrix of b

This matrix $\mathbf{b}$ is a linear combination of the elements of $\mathbf{Y}$.
These estimates are normal if $\mathbf{Y}$ is normal.
These estimates will be approximately normal in general.

A Useful Multivariate Theorem

> Suppose $\mathbf{U} \sim N(\mu, \Sigma)$, a multivariate normal vector, and $\mathbf{V} = \mathbf{c} + \mathbf{DU}$, a linear transformation of $\mathbf{U}$ where $\mathbf{c}$ is a vector and $\mathbf{D}$ is a matrix. Then $\mathbf{V} \sim N(\mathbf{c} + \mathbf{D}\mu, \mathbf{D}\Sigma\mathbf{D}')$.

Recall: $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\,\mathbf{Y}$ and $\mathbf{Y} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I})$.
Now apply theorem to $\mathbf{b}$ using

$$\mathbf{U} = \mathbf{Y}, \mu = \mathbf{X}\beta, \Sigma = \sigma^2\mathbf{I}$$
$$\mathbf{V} = \mathbf{b}, \mathbf{c} = \mathbf{0}, \text{ and } \mathbf{D} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

The theorem tells us the vector $\mathbf{b}$ is normally distributed with mean

$$(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\beta = \beta$$

and covariance matrix

$$\begin{aligned}\sigma^2\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\mathbf{I}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)' &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\left((\mathbf{X}'\mathbf{X})^{-1}\right)' \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

using the fact that both $\mathbf{X}'\mathbf{X}$ and its inverse are symmetric, so $((\mathbf{X}'\mathbf{X})^{-1})' = (\mathbf{X}'\mathbf{X})^{-1}$.

Next we will use this framework to do multiple regression where we have more than one explanatory variable (i.e., add another column to the design matrix and additional beta parameters).

Multiple Regression

Data for Multiple Regression

- Y_i is the response variable (as usual)

- $X_{i,1}, X_{i,2}, \ldots, X_{i,p-1}$ are the $p-1$ explanatory variables for cases $i = 1$ to n.
- Example – In Homework #1 you considered modeling GPA as a function of entrance exam score. But we could also consider intelligence test scores and high school GPA as potential predictors. This would be 3 variables, so $p = 4$.
- *Potential problem to remember!!!* These predictor variables are likely to be themselves correlated. We always want to be careful of using variables that are themselves strongly correlated as predictors together in the same model.

The Multiple Regression Model

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_{p-1} X_{i,p-1} + \epsilon_i \text{ for } i = 1, 2, \ldots, n$$

where

- Y_i is the value of the response variable for the ith case.
- $\epsilon_i \sim^{iid} N(0, \sigma^2)$ (exactly as before!)
- β_0 is the intercept (think multidimensionally).
- $\beta_1, \beta_2, \ldots, \beta_{p-1}$ are the regression coefficients for the explanatory variables.
- $X_{i,k}$ is the value of the kth explanatory variable for the ith case.
- Parameters as usual include all of the β's as well as σ^2. These need to be estimated from the data.

Interesting Special Cases

- Polynomial model:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \ldots + \beta_{p-1} X_i^{p-1} + \epsilon_i$$

- X's can be indicator or dummy variables with $X = 0$ or 1 (or any other two distinct numbers) as possible values (e.g. ANOVA model). Interactions between explanatory variables are then expressed as a product of the X's:

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,1} X_{i,2} + \epsilon_i$$

Model in Matrix Form

$$\begin{aligned} \mathbf{Y}_{n\times 1} &= \mathbf{X}_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1} \\ \epsilon &\sim N(\mathbf{0}, \sigma^2\mathbf{I}_{n\times n}) \\ \mathbf{Y} &\sim N(\mathbf{X}\beta, \sigma^2\mathbf{I}) \end{aligned}$$

Design Matrix $\mathbf{X}$:

$$\mathbf{X} = \begin{bmatrix} 1 & X_{1,1} & X_{1,2} & \cdots & X_{1,p-1} \\ 1 & X_{2,1} & X_{2,2} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & X_{n,2} & \cdots & X_{n,p-1} \end{bmatrix}$$

Coefficient matrix β:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$$

Parameter Estimation

Least Squares

Find $\mathbf{b}$ to minimize $SSE = (\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb})$
Obtain normal equations as before: $\mathbf{X'Xb} = \mathbf{X'Y}$

Least Squares Solution

$$\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$$

Fitted (predicted) values for the mean of Y are

$$\hat{\mathbf{Y}} = \mathbf{Xb} = \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'Y} = \mathbf{HY},$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}$.

Residuals

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{HY} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

Notice that the matrices $\mathbf{H}$ and $(\mathbf{I} - \mathbf{H})$ have two special properties. They are

- *Symmetric:* $\mathbf{H} = \mathbf{H'}$ and $(\mathbf{I} - \mathbf{H})' = (\mathbf{I} - \mathbf{H})$.
- *Idempotent:* $\mathbf{H^2} = \mathbf{H}$ and $(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H}) = (\mathbf{I} - \mathbf{H})$

Covariance Matrix of Residuals

$$\begin{aligned} Cov(\mathbf{e}) &= \sigma^2(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H})' = \sigma^2(\mathbf{I}-\mathbf{H}) \\ Var(e_i) &= \sigma^2(1-h_{i,i}), \end{aligned}$$

where $h_{i,i}$ is the ith diagonal element of $\mathbf{H}$.

Note: $h_{i,i} = \mathbf{X'}_i(\mathbf{X'X})^{-1}\mathbf{X}_i$ where $\mathbf{X}'_i = [1\, X_{i,1} \cdots X_{i,p-1}]$.

Residuals e_i are usually somewhat correlated: $cov(e_i, e_j) = -\sigma^2 h_{i,j}$; this is not unexpected, since they sum to 0.

Estimation of σ

Since we have estimated p parameters, $SSE = \mathbf{e'e}$ has $df_E = n-p$. The estimate for σ^2 is the usual estimate:

$$\begin{aligned} s^2 &= \frac{\mathbf{e'e}}{n-p} = \frac{(\mathbf{Y}-\mathbf{Xb})'(\mathbf{Y}-\mathbf{Xb})}{n-p} = \frac{SSE}{df_E} = MSE \\ s &= \sqrt{s^2} = \texttt{Root MSE} \end{aligned}$$

Distribution of b

We know that $\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$. The only RV involved is Y, so the distribution of $\mathbf{b}$ is based on the distribution of $\mathbf{Y}$.

Since $\mathbf{Y} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I})$, and using the multivariate theorem from earlier (if you like, go through the details on your own), we have

$$\begin{aligned} \mathrm{E}(\mathbf{b}) &= \left((\mathbf{X'X})^{-1}\mathbf{X'}\right)\mathbf{X}\beta = \beta \\ \sigma^2\{\mathbf{b}\} &= \mathrm{Cov}(\mathbf{b}) = \sigma^2(\mathbf{X'X})^{-1} \end{aligned}$$

Since σ^2 is estimated by the MSE s^2, $\sigma^2\{\mathbf{b}\}$ is estimated by $s^2(\mathbf{X'X})^{-1}$.

ANOVA Table

Sources of variation are

- Model (SAS) or Regression (NKNW)
- Error (Residual)
- Total

SS and *df* add as before

$$\begin{aligned} SSM + SSE &= SST \\ df_M + df_E &= df_{Total} \end{aligned}$$

but their values are different from SLR.

Sum of Squares

$$\begin{aligned} SSM &= \sum(\hat{Y}_i - \bar{Y})^2 \\ SSE &= \sum(Y_i - \hat{Y}_i)^2 \\ SSTO &= \sum(Y_i - \bar{Y})^2 \end{aligned}$$

Degrees of Freedom

$$\begin{aligned} df_M &= p-1 \\ df_E &= n-p \\ df_{Total} &= n-1 \end{aligned}$$

The total degrees have not changed from SLR, but the model *df* has increased from 1 to $p-1$, i.e., the number of X variables. Correspondingly, the error *df* has decreased from $n-2$ to $n-p$.

Mean Squares

$$\begin{aligned} MSM &= \frac{SSM}{df_M} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{p-1} \\ MSE &= \frac{SSE}{df_E} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-p} \\ MST &= \frac{SSTO}{df_{Total}} = \frac{\sum(Y_i - \bar{Y})^2}{n-1} \end{aligned}$$

ANOVA Table

Source	df	SS	MSE	F
Model	$df_M = p-1$	SSM	MSM	$\frac{MSM}{MSE}$
Error	$df_E = n-p$	SSE	MSE	
Total	$df_T = n-1$	SST		

***F*-test**

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$ (all regression coefficients are zero)
$H_A : \beta_k \neq 0$, for at least one $k = 1, \ldots, p-1$; at least of the β's is non-zero (or, not all the β's are zero).
$F = MSM/MSE$
Under H_0, $F \sim F_{p-1,n-p}$
Reject H_0 if F is larger than critical value; if using SAS, reject H_0 if p-value $< \alpha = 0.05$.

What do we conclude?

If H_0 is rejected, we conclude that at least one of the regression coefficients is non-zero; hence at least one of the X variables is useful in predicting Y. (Doesn't say which one(s) though). If H_0 is not rejected, then we cannot conclude that *any* of the X variables is useful in predicting Y.

p-value of F-test

The p-value for the F significance test tell us one of the following:

- there is no evidence to conclude that *any* of our explanatory variables can help us to model the response variable using this kind of model ($p \geq 0.05$).
- *one or more* of the explanatory variables in our model *is* potentially useful for predicting the response in a linear model ($p \leq 0.05$).

R^2

The squared multiple regression correlation (R^2) gives the proportion of variation in the response variable explained by the explanatory variables.
It is sometimes called the *coefficient of multiple determination* (NKNW, page 230).
$R^2 = SSM/SST$ (the proportion of variation explained by the model)
$R^2 = 1 - (SSE/SST)$ ($1 -$ the proportion not explained by the model)
F and R^2 are related:

$$F = \frac{R^2/(p-1)}{(1-R^2)/(n-p)}$$

Inference for Individual Regression Coefficients

Confidence Interval for β_k

We know that $\mathbf{b} \sim N(\beta, \sigma^2(\mathbf{X'X})^{-1})$
Define

$$\mathbf{s^2\{b\}}_{p\times p} = MSE \times (\mathbf{X'X})^{-1}$$

$$s^2\{b_k\} = \left[\mathbf{s^2\{b\}}\right]_{k,k}, \text{ the } k\text{th diagonal element}$$

CI for β_k: $b_k \pm t^c s\{b_k\}$, where $t^c = t_{n-p}(0.975)$.

Significance Test for β_k

$H_0 : \beta_k = 0$
Same test statistic $t^* = b_k/s\{b_k\}$
Still use df_E which now is equal to $n - p$
p-value computed from t_{n-p} distribution.
This tests the significance of a variable *given that the other variables are already in the model* (i.e., fitted last). Unlike in SLR, the t-tests for β are *different* from the F-test.

Multiple Regression – Case Study

Example: Study of CS Students

Problem: Computer science majors at Purdue have a large drop-out rate.
Potential Solution: Can we find predictors of success? Predictors must be available at time of entry into program.

Data Available

Grade point average (GPA) after three semesters (Y_i, the response variable)
Five potential predictors ($p = 6$)

- X_1 = High school math grades (HSM)
- X_2 = High school science grades (HSS)
- X_3 = High school English grades (HSE)
- X_4 = SAT Math (SATM)
- X_5 = SAT Verbal (SATV)
- Gender (1 = male, 2 = female) (we will ignore this one right now, since it is not a continuous variable).

We have $n = 224$ observations, so if all five variables are included, the design matrix $\mathbf{X}$ is 224×6. The SAS program used to generate output for this is `cs.sas`.

Look at the individual variables

Our first goal should be to take a look at the variables to see...

- Is there anything that sticks out as unusual for any of the variables?
- How are these variables related to each other (pairwise)? *If two predictor variables are strongly correlated, we wouldn't want to use them in the same model!*

We do this by looking at statistics and plots.

```
data cs;
   infile 'H:\System\Desktop\csdata.dat';
   input id gpa hsm hss hse satm satv genderm1;
```

Descriptive Statistics: `proc means`

```
proc means data=cs maxdec=2;
   var gpa hsm hss hse satm satv;
```

The option `maxdec = 2` sets the number of decimal places in the output to 2 (just showing you how).

Output from `proc means`

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
gpa	224	2.64	0.78	0.12	4.00
hsm	224	8.32	1.64	2.00	10.00
hss	224	8.09	1.70	3.00	10.00
hse	224	8.09	1.51	3.00	10.00
satm	224	595.29	86.40	300.00	800.00
satv	224	504.55	92.61	285.00	760.00

Descriptive Statistics

Note that `proc univariate` also provides lots of other information, not shown.

```
proc univariate data=cs noprint;
   var gpa hsm hss hse satm satv;
   histogram gpa hsm hss hse satm satv /normal;
```

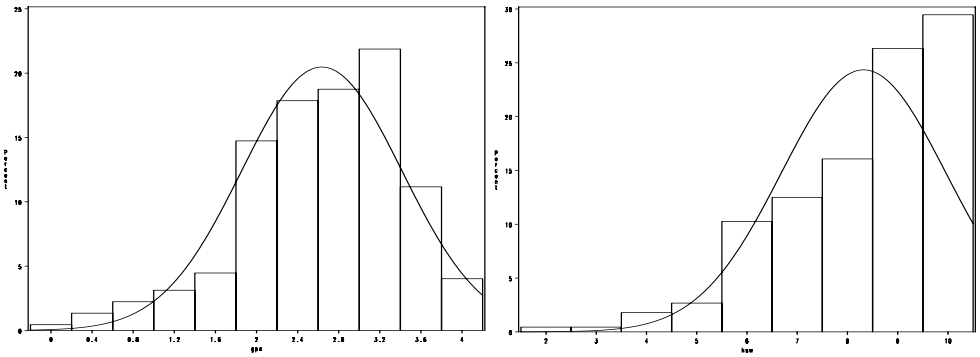

Figure 1: Graph of GPA (left) and High School Math (right)

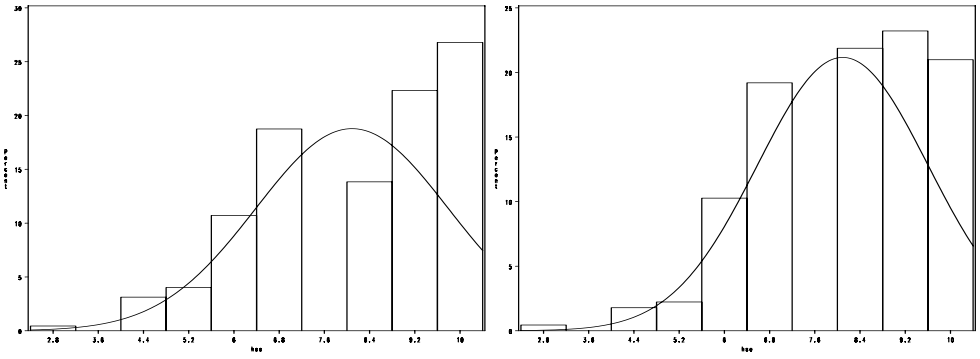

Figure 2: Graph of High School Science (left) and High School English (right)

NOTE: If you want the plots (e.g., histogram, qqplot) and not the copious output from `proc univariate`, use a `noprint` statement

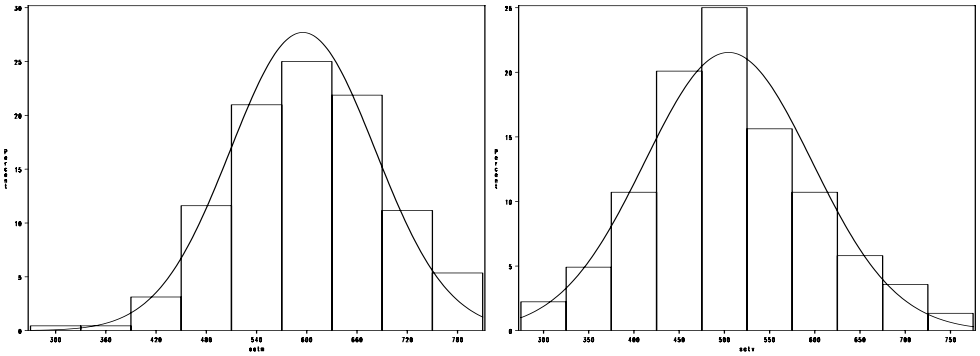

Figure 3: Graph of SAT Math (left) and SAT Verbal (right)

```
proc univariate data = cs noprint;
     histogram gpa / normal;
```

Interactive Data Analysis

Read in the dataset as usual
From the menu bar, select

```
Solutions -> analysis -> interactive data analysis
```

Obtain SAS/Insight window

- Open library `work`
- Click on Data Set `CS` and click "`open`".

Getting a Scatter Plot Matrix

(`CTRL`) Click on `GPA`, `SATM`, `SATV`
Go to menu `Analyze`
Choose option `Scatterplot(Y X)`
You can, while in this window, use `Edit -> Copy` to copy this plot to another program such as Word: (See Figure 4.)
This graph – once you get used to it – can be useful in getting an overall feel for the relationships among the variables. Try other variables and some other options from the `Analyze` menu to see what happens.

Correlations

SAS will give us the r (correlation) value between pairs of random variables in a data set using `proc corr`.

```
proc corr data=cs;
   var hsm hss hse;
```

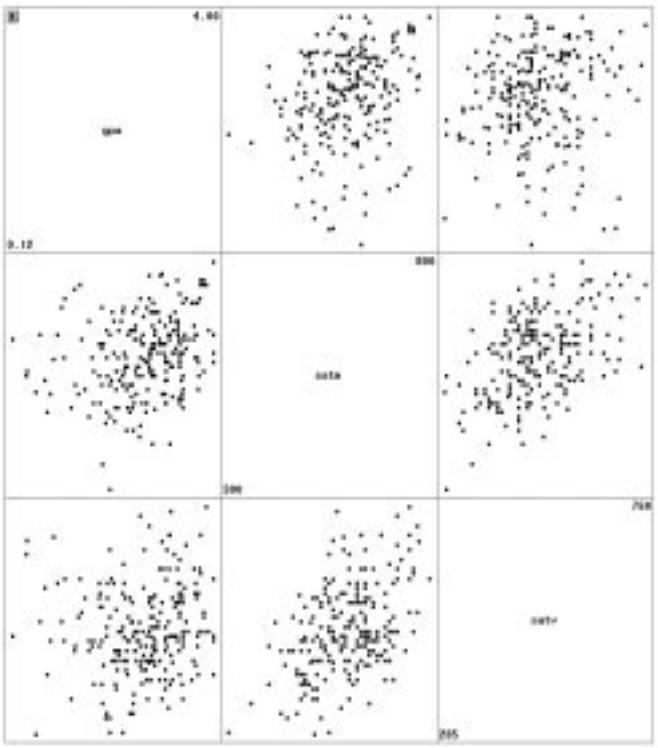

Figure 4: Scatterplot Matrix

	hsm	hss	hse
hsm	1.00000	0.57569	0.44689
		<.0001	<.0001
hss	0.57569	1.00000	0.57937
	<.0001		<.0001
hse	0.44689	0.57937	1.00000
	<.0001	<.0001	

To get rid of those p-values, which make it difficult to read, use a `noprob` statement. Here are also a few different ways to call `proc corr`:

```
proc corr data=cs noprob;
   var satm satv;
```

	satm	satv
satm	1.00000	0.46394
satv	0.46394	1.00000

```
proc corr data=cs noprob;
   var hsm hss hse;
   with satm satv;
```

	hsm	hss	hse
satm	0.45351	0.24048	0.10828
satv	0.22112	0.26170	0.24371

```
proc corr data=cs noprob;
   var hsm hss hse satm satv;
   with gpa;
```

	hsm	hss	hse	satm	satv
gpa	0.43650	0.32943	0.28900	0.25171	0.11449

Notice that not only do the X's correlate with Y (this is good), but the X's correlate with *each other* (this is bad). That means that some of the X's may be *redundant* in predicting Y.

Use High School Grades to Predict GPA

```
proc reg data=cs;
   model gpa=hsm hss hse;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	27.71233	9.23744	18.86	<.0001
Error	220	107.75046	0.48977		
Corrected Total	223	135.46279			

Root MSE	0.69984	R-Square	0.2046
Dependent Mean	2.63522	Adj R-Sq	0.1937
Coeff Var	26.55711		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.58988	0.29424	2.00	0.0462
hsm	1	0.16857	0.03549	4.75	<.0001
hss	1	0.03432	0.03756	0.91	0.3619
hse	1	0.04510	0.03870	1.17	0.2451

Remove HSS

```
proc reg data=cs;
   model gpa=hsm hse;
```

R-Square	0.2016	Adj R-Sq	0.1943

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.62423	0.29172	2.14	0.0335
hsm	1	0.18265	0.03196	5.72	<.0001
hse	1	0.06067	0.03473	1.75	0.0820

Rerun with HSM only

```
proc reg data=cs;
   model gpa=hsm;
```

```
R-Square     0.1905    Adj R-Sq     0.1869
                      Parameter       Standard
Variable     DF       Estimate          Error    t Value    Pr > |t|
Intercept     1        0.90768        0.24355       3.73      0.0002
hsm           1        0.20760        0.02872       7.23      <.0001
```

The last two models (HSM only and HSM, HSE) appear to be pretty good. Notice that R^2 go down a little with the deletion of HSE , so HSE does provide a little information. This is a judgment call: do you prefer a slightly better-fitting model, or a simpler one?

Now look at SAT scores

```
proc reg data=cs;
   model gpa=satm satv;
```

```
R-Square     0.0634    Adj R-Sq     0.0549
                      Parameter       Standard
Variable     DF       Estimate          Error    t Value    Pr > |t|
Intercept     1        1.28868        0.37604       3.43      0.0007
satm          1        0.00228     0.00066291       3.44      0.0007
satv          1    -0.00002456     0.00061847      -0.04      0.9684
```

Here we see that SATM is a significant explanatory variable, but the overall fit of the model is poor. SATV does not appear to be useful at all.

Now try our two most promising candidates: HSM and SATM

```
proc reg data=cs;
   model gpa=hsm satm;
```

```
R-Square     0.1942    Adj R-Sq     0.1869
                      Parameter       Standard
Variable     DF       Estimate          Error    t Value    Pr > |t|
Intercept     1        0.66574        0.34349       1.94      0.0539
hsm           1        0.19300        0.03222       5.99      <.0001
satm          1     0.00061047     0.00061117       1.00      0.3190
```

In the presence of HSM, SATM does not appear to be at all useful. Note that adjusted R^2 is the same as with HSM only, and that the p-value for SATM is now no longer significant.

General Linear Test Approach: HS and SAT's

```
proc reg data=cs;
   model gpa=satm satv hsm hss hse;
   *Do general linear test;
   * test H0: beta1 = beta2 = 0;
   sat: test satm, satv;
   * test H0: beta3=beta4=beta5=0;
   hs: test hsm, hss, hse;
```

```
R-Square     0.2115    Adj R-Sq     0.1934
                      Parameter       Standard
Variable     DF       Estimate          Error    t Value    Pr > |t|
Intercept     1        0.32672        0.40000       0.82      0.4149
satm          1     0.00094359     0.00068566       1.38      0.1702
satv          1    -0.00040785     0.00059189      -0.69      0.4915
hsm           1        0.14596        0.03926       3.72      0.0003
hss           1        0.03591        0.03780       0.95      0.3432
hse           1        0.05529        0.03957       1.40      0.1637
```

The first `test` statement tests the full vs reduced models (H_0 : no satm or satv, H_a : full model)

```
      Test sat Results for Dependent Variable gpa
                                Mean
Source             DF         Square    F Value    Pr > F
Numerator           2        0.46566       0.95    0.3882
Denominator       218        0.49000
```

We do not reject the H_0 that $\beta_1 = \beta_2 = 0$. Probably okay to throw out SAT scores.

The second `test` statement tests the full vs reduced models (H_0 : no hsm, hss, or hse, H_a : all three in the model)

```
      Test hs Results for Dependent Variable gpa
                                Mean
Source             DF         Square    F Value    Pr > F
Numerator           3        6.68660      13.65    <.0001
Denominator       218        0.49000
```

We reject the H_0 that $\beta_3 = \beta_4 = \beta_5 = 0$. CANNOT throw out all high school grades.

Can use the `test` statement to test any set of coefficients equal to zero. Related to extra sums of squares (later).

Best Model?

Likely the one with just HSM. Could argue that HSM and HSE is marginally better. We'll discuss comparison methods in Chapters 7 and 8.

Key ideas from case study

- First, look at graphical and numerical summaries for one variable at a time.
- Then, look at relationships between pairs of variables with graphical and numerical summaries.
- Use plots and correlations to understand relationships.
- The relationship between a response variable and an explanatory variable depends on what other explanatory variables are in the model.

- A variable can be a significant ($p < 0.05$) predictor alone and not significant ($p > 0.05$) when other X's are in the model.
- Regression coefficients, standard errors, and the results of significance tests depend on what other explanatory variables are in the model.
- Significance tests (p-values) do not tell the whole story.
- Squared multiple correlations (R^2) give the proportion of variation in the response variable explained by the explanatory variables, and can give a different view; see also adjusted R^2.
- You can fully understand the theory in terms of $\mathbf{Y} = \mathbf{X}\beta + \epsilon$.
- To effectively use this methodology in practice, you need to understand how the data were collected, the nature of the variables, and how they relate to each other.

Other things that should be considered

- *Diagnostics:* Do the models meet the assumptions? You might take one of our final two potential models and check these out on your own.
- Confidence intervals/bands, Prediction intervals/bands?

Example II (NKNW page 241)

Dwaine Studios, Inc. operates portrait studios in $n = 21$ cities of medium size.
Program used to generate output for confidence intervals for means and prediction intervals is `nknw241.sas`.

- Y_i is sales in city i
- X_1 : population aged 16 and under
- X_2 : per capita disposable income

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \epsilon_i$$

Read in the data

```
data a1;
   infile 'H:\System\Desktop\CH06FI05.DAT';
   input young income sales;
proc print data=a1;
```

```
Obs    young    income    sales
  1     68.5     16.7     174.4
  2     45.2     16.8     164.4
  3     91.3     18.2     244.2
  4     47.8     16.3     154.6
...
```

```
proc reg data=a1;
   model sales=young income
```

```
                                  Sum of           Mean
Source                     DF       Squares         Square    F Value    Pr > F
Model                       2         24015          12008      99.10    <.0001
Error                      18    2180.92741      121.16263
Corrected Total            20         26196

Root MSE              11.00739    R-Square     0.9167
Dependent Mean       181.90476    Adj R-Sq     0.9075
Coeff Var              6.05118

                    Parameter      Standard
Variable     DF      Estimate         Error   t Value   Pr > |t|
Intercept     1     -68.85707      60.01695     -1.15     0.2663
young         1       1.45456       0.21178      6.87     <.0001
income        1       9.36550       4.06396      2.30     0.0333
```

`clb` option: Confidence Intervals for the β's

```
proc reg data=a1;
   model sales=young income/clb;
```

```
    95% Confidence Limits
Intercept    -194.94801      57.23387
young           1.00962       1.89950
income          0.82744      17.90356
```

Estimation of $\mathbf{E}(Y_h)$

$\mathbf{X}_h$ is now a vector of values. (Y_h is still just a number.)
$(1, X_{h,1}, X_{h,2}, \ldots, X_{h,p-1})' = \mathbf{X}_h'$: this is row h of the design matrix. We want a point estimate and a confidence interval for the subpopulation mean corresponding to the set of explanatory variables $\mathbf{X}_h$.

Theory for $\mathbf{E}(Y)_h$

$$\begin{aligned} \mathrm{E}(Y_h) &= \mu_j = \mathbf{X'}_h\beta \\ \hat{\mu}_h &= \mathbf{X}_h'\mathbf{b} \\ s^2\{\hat{\mu}_h\} &= \mathbf{X}_h'\mathbf{s}^2\{\mathbf{b}\}\mathbf{X}_h = s^2\mathbf{X}_h'(\mathbf{X'X})^{-1}\mathbf{X}_h \end{aligned}$$

$95\%CI : \hat{\mu}_h \pm s\{\hat{\mu}_h\}t_{n-p}(0.975)$

`clm` option: Confidence Intervals for the Mean

```
proc reg data=a1;
   model sales=young income/clm;
   id young income;
```

Obs	young	income	Dep Var sales	Predicted Value	Std Error Mean Predict	95% CL Mean	
1	68.5	16.7	174.4000	187.1841	3.8409	179.1146	195.2536
2	45.2	16.8	164.4000	154.2294	3.5558	146.7591	161.6998
3	91.3	18.2	244.2000	234.3963	4.5882	224.7569	244.0358
4	47.8	16.3	154.6000	153.3285	3.2331	146.5361	160.1210

Prediction of $Y_{h(new)}$

Predict a new observation Y_h with X values equal to $\mathbf{X}_h$.
We want a prediction of Y_h based on a set of predictor values with an interval that expresses the uncertainty in our prediction. As in SLR, this interval is centered at $\hat{Y}_h$ and is wider than the interval for the mean.

Theory for Y_h

$$\begin{aligned} Y_h &= \mathbf{X'}_h\beta + \epsilon \\ \hat{Y}_h &= \hat{\mu}_h = \mathbf{X'}_h\beta \\ s^2\{pred\} &= \mathrm{Var}(\hat{Y}_h + \epsilon) = \mathrm{Var}(\hat{Y}_h) + \mathrm{Var}(\epsilon) \\ &= s^2\left(1 + \mathbf{X'}_h(\mathbf{X'X})^{-1}\mathbf{X}_h\right) \end{aligned}$$

CI for $Y_{h(new)}$: $\hat{Y}_h \pm s\{pred\}t_{n-p}(0.975)$

`cli` option: Confidence Interval for an Individual Observation

```
proc reg data=a1;
   model sales=young income/cli;
   id young income;
```

Obs	young	income	Dep Var sales	Predicted Value	Std Error Mean Predict	95% CL Predict	
1	68.5	16.7	174.4000	187.1841	3.8409	162.6910	211.6772
2	45.2	16.8	164.4000	154.2294	3.5558	129.9271	178.5317
3	91.3	18.2	244.2000	234.3963	4.5882	209.3421	259.4506
4	47.8	16.3	154.6000	153.3285	3.2331	129.2260	177.4311

Diagnostics

Look at the distribution of each variable to gain insight.
Look at the relationship between pairs of variables. `proc corr` is useful here. BUT note that relationships between variables can be more complicated than just pairwise: correlations are NOT the entire story.

- Plot the residuals vs...
  - the predicted/fitted values
  - each explanatory variable
  - time (if available)
- Are the relationships linear?
  - Look at Y vs each X_i
  - May have to transform some X's.
- Are the residuals approximately normal?
  - Look at a histogram
  - Normal quantile plot
- Is the variance constant?
  - Check all the residual plots.

Remedies

Similar remedies to simple regression (but more complicated to decide, though).
Additionally may eliminate some of the X's (this is call *variable selection*).
Transformations such as Box-Cox
Analyze with/without outliers
More detail in NKNW Ch 9 and 10