Nelson A. Rockefeller College of Public Affairs and Policy

Rockefeller College
University at Albany

Outliers and DFBETA

Outliers can sometimes cause problems with regression results. Outliers are defined by Gujarati (p. 540-541) as an observation with a large residual – a larger vertical distance between the observation and the predicted line than is generally true for the rest of the data. Such observations may have high “leverage” if they are disproportionately far away from the rest of the data points. High leverage observations may also be “influential”: that is, they pull the regression line toward that observation. Not every high leverage observation is an influential one. Below I reproduce a set of three graphs from Gujarati (p. 541). Only in graph (b) does the outlier (the asterisk in the box) really matter – it causes the regression line to be “pulled” down from what otherwise would be found using the rest of the data (the open circles). In (c), the outlier has leverage, but it falls in line with the other observations so the slope is unaffected. In (a) the outlier is has a large residual, but is not far from the mean of the independent. Thus only the intercept is really affected; the slope is unaffected.

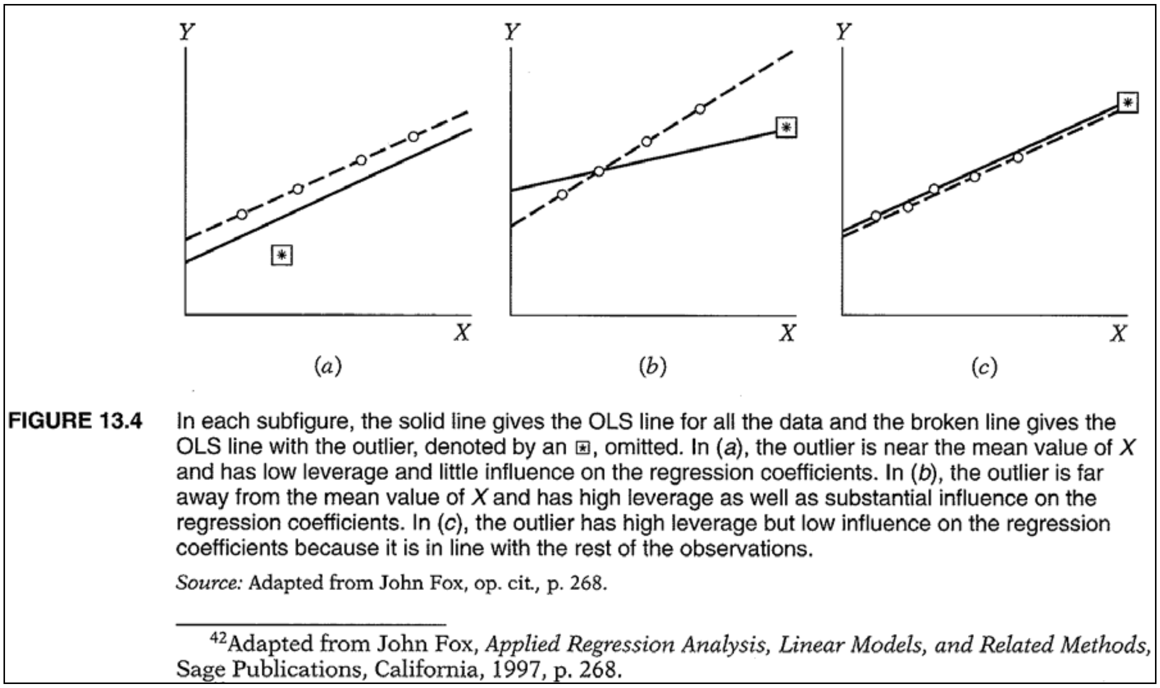

FIGURE 13.4 In each subfigure, the solid line gives the OLS line for all the data and the broken line gives the OLS line with the outlier, denoted by an ⧆, omitted. In (*a*), the outlier is near the mean value of *X* and has low leverage and little influence on the regression coefficients. In (*b*), the outlier is far away from the mean value of *X* and has high leverage as well as substantial influence on the regression coefficients. In (*c*), the outlier has high leverage but low influence on the regression coefficients because it is in line with the rest of the observations.

Source: Adapted from John Fox, op. cit., p. 268.

[42] Adapted from John Fox, *Applied Regression Analysis, Linear Models, and Related Methods*, Sage Publications, California, 1997, p. 268.

Dectecting Outliers

There are multiple methods for detecting outliers (see the section in the Stata manual title `regress postestimation – Postestimation tools for regress`). Probably the most popular tools is DFBETA. DFBETA is a measure found for each observation in a dataset. The DFBETA for a particular observation is the difference between the regression coefficient for an included variable (say age, or education in our well-worn salary example) calculated for all of the data and the regression coefficient calculated with the observation deleted, scaled by the standard error calculated with the observation deleted. The cut-off value for DFBETAs is 2/sqrt(n), where n is the number of observations. However, another cut-off is to look for observations with a value greater than 1.00. Here cutoff means, "this observation could be overly influential on the estimated coefficient."

The Stata command for DFBETA is `dfbeta`. If you want to know the DFBETA for a single variable, use the following command:

```
dfbeta age
```

The results are, by default, put into a variable with the name `DFage`.

To examine the results, use the following commands

```
list salary age DFage if abs(DFage) > 2/sqrt(950), divider
list salary age DFage if abs(DFage) > 1, divider
```

The first command uses the 2/sqrt(n) criteria to identify high-leverage, high influence observations; the second uses the 1.00 cutoff.

A second option is to get a measure of DFBETA for all included regressors:

```
dfbeta
```

This results in a set of new variables being created, where each has the name DF <var name>.

In the example below first I ran:

```
reg salary age age2 educ fem
```

Then:

```
dfbeta
```

The command created variables named `DFage`, `DFage2`, `DFeduc`, and `DFfem`.

To inspect the results I used the following commands:

```
list salary age DFage DFage2 DFeduc DFfem if (abs(DFage) > 1 |
abs(DFage2) > 1 | abs(DFeduc)> 1 | abs(DFfem) > 1), divider

list salary age DFage DFage2 DFeduc DFfem if (abs(DFage) > 2/sqrt(950)
| abs(DFage2) > 2/sqrt(950) | abs(DFeduc)> 2/sqrt(950) | abs(DFfem) >
2/sqrt(950)), divider

list salary age DFage DFage2 DFeduc DFfem if (abs(DFage) > .2 |
abs(DFage2) > .2 | abs(DFeduc)> .2 | abs(DFfem) > .2), divider
```

DFBETA is a rule of thumb: drop observations that have too much influence on the regression line, where "too much" is defined by the 2/sqrt(n) or 1.00 cutoffs. Choices about outliers are controversial. Some statisticians would suggest that dropping valid observations (i.e., ones that conform to the design of the survey instrument) is never a good idea: the sample is what it is. Bias may result, or the range over which the regression is applicable may change. For instance, in the `gender.dta` dataset, many of the high-leverage, high-influence observations are for people who are very old or very young. One strategy may be to restrict the nature of the study – to those individuals between 18 and 65, for instance. Yet we lose something in the process.

Others would suggest that leaving in outliers can lead to the "crack-using granny" problem I have discussed before in class: one observation is allowed to have too much influence over the regression (and any research or policy conclusions that flow from it).

One solution is to report findings with and without outliers so that fair readers can make up their own minds.

Once again, regression analysis is both an art and a science. Here, we are closer to the art end of the spectrum.