**Lab Assignment Week 2**

**I. Objectives**: To examine how *simple* regression coefficients are estimated and interpreted, and to examine the symptoms, diagnosis, and remedies for influential data points.

**II. Datasets**: *CEOSAL1.dta*, *accidents50.csv*

**III. Packages**: *foreign*, *car*

**IV. Preparation**

1) Open RStudio by double-clicking the icon or selecting RStudio from the Windows Start menu.

2) Clear any data in memory by typing rm(list=ls()) as in line 2

3) Download datasets *CEOSAL1.dta* and *accidents50.csv* and place both in a working directory

4) Download R script “Lab02.R” and place it in a working directory.

5) Open the R script by typing *Ctrl*+*O* or by clicking on File in the upper-left corner, using the dropdown menu, and navigating to the script in your working directory.

6) Run lines 3–5 in the R script to install and load two packages that you will need

**V. Instructions for Lab Week 2**

The first dataset that you will use in this lab contains data on CEO salaries. These data are used in chapter 2 of the Wooldridge textbook (see page 6 of the worksheet). The main question that we will explore is whether a CEO’s salary reflects the Return on Equity, which is defined as the company’s net income as a percentage of common equity (averaged over three years).

Begin by loading the dataset, either by running line 7 or typing the following code, changing the directory if needed:

> CEOdta<-read.dta("C:/Users/Scott/CEOSAL1.DTA")

If you failed to load the foreign package earlier, then R will not load the Stata dataset.

If you pay close attention to the **Console** window, you might see a warning, indicating that R cannot load labels for Stata 5 datasets. The variables that we use in this lab are a CEO’s salary (*salary*) in thousands of dollars, the natural logarithm of a CEO’s salary (*lsalary*), and a firm’s return on equity (*roe*). The dataset also contains dummy variables indicating whether the company is industrial, financial, consumer products, or a utility.

Run line 8 to create a new data frame in which data are sorted by the company’s return on equity; this will help with some of the figures later.

> sorted <- CEOdta[order(CEOdta$roe),]

***A. Estimating and Interpreting Regression Coefficients***

Run lines 10–11 to examine the bivariate relationship between the CEO’s salary and his/her company’s return on equity; the second line of code adds a trend line. Does it appear that CEOs are generally rewarded when roe is higher? What is unusual about this figure?

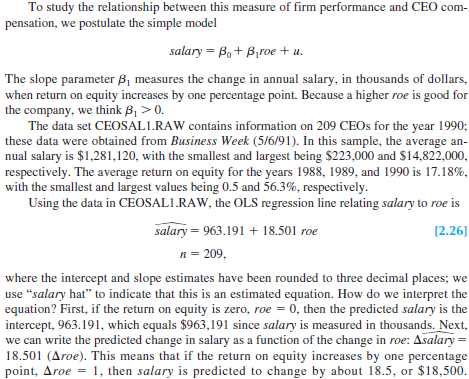
> plot(sorted$roe, sorted$salary, pch=19)

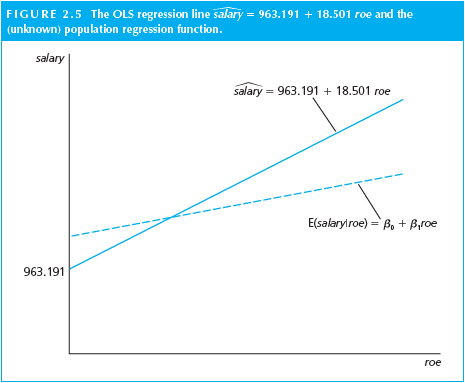
> abline(lm(sorted$salary~sorted$roe), col="red")

Run line 12 to regress salary on roe (equation [2.26] in Wooldridge) and display the estimates:

> reg1<-lm(sorted$salary~sorted$roe); summary(reg1)

The code shown above replicates Wooldridge’s Example 2.3. For your reference, here is the relevant page from Wooldridge (Fifth Edition):





(Ignore the dashed blue line, which represents an unknown population regression function.)

Write the intercept ( = \_\_\_\_\_\_\_) and the slope (= \_\_\_\_\_\_\_) of the simple regression line.

The intercept informs you the expected salary when return on equity equals zero. Recall that *salary* is coded in thousands of dollars. What is the expected salary for a CEO when *roe* = 0? (The actual range of *roe* in the dataset is 0.5% to 56.3%, so a slight extrapolation is required.)

The slope informs you the impact of a one-unit change in *roe* (in percent) on *salary* (in thousands). A one unit increase in *roe* is expected to (increase? decrease?) \_\_\_\_\_\_\_\_\_ the CEO’s salary by \_\_\_\_\_\_\_\_\_ dollars.

Let us use this example to demonstrate a few key points regarding regression lines. To see the mean values of the independent variable and the dependent variable, run line 14 to see the summary statistics. Write the following values for later reference: = \_\_\_\_\_ ; = \_\_\_\_\_.

If you run line 16 then R will use the regression coefficients to generate the linear prediction of the CEO’s salary when *roe* is set at 0.

> coef(reg1)[1]+coef(reg1)[2]\*0

What value does R provide for the CEO’s salary, conditional on *roe* equaling 0? \_\_\_\_\_\_

If you run line 17 then R will use the regression coefficients to generate the linear prediction of the CEO’s salary when *roe* is set at its mean value.

> coef(reg1)[1]+coef(reg1)[2]\* mean(sorted$roe)

What value does R provide for the CEO’s salary, conditional on *roe* equaling ? \_\_\_\_\_\_

Compare this *predicted* value to the *actual* mean salary in the dataset ().

In lecture we are working on estimating simple regression coefficients using just four values: the mean of *x*, the mean of *y*, the variance of *x*, and the covariance of *x* and *y*. Run lines 19–20 to find the slope and then the intercept (using the slope):

> bhat = cov(sorted$salary,sorted$roe)/var(sorted$roe)

> ahat = mean(sorted$salary) - bhat\*mean(sorted$roe)

In lecture, we have also discussed two key assumptions about the residuals of simple regression models; these assumptions are the source of the equations for the estimated slope and estimated intercept. Specifically, those assumptions (which you are about to check) are:

i) the average residual equals zero (line 22):

> mean(reg1$residuals)

ii) the correlation between the residual and the covariate equal zero (line 23):

> cor(reg1$residuals, sorted$roe)

***B. Influential data points.***

Economists often transform data involving counts, times, and money by taking the natural logarithm. You can examine histograms of salary and the natural log of salary by typing:

> hist(sorted$salary, breaks=15)

> hist(sorted$lsalary, breaks=15)

Which of these variables (salary or the natural log of salary) appears to have a more ‘normal’ distribution?

In the remainder of section B, we use the natural log of salary as our dependent variable.

Run line 25 to re-estimate the regression model with the log of salary as the dependent variable and display the regression model’s intercept, slope, etc.:

> reg2<-lm(sorted$lsalary~sorted$roe)

> summary(reg2)

Obviously the estimated coefficients are very different when we use the natural log of salary as the dependent variable, in place of salary in (thousands of) dollars.[[1]](#footnote-1) Later in the semester, we will discuss how to interpret the coefficients of a “log-linear” model (like this one) correctly and precisely. In the meantime…

Run lines 27–28 to plot the data points and the regression line:

> plot(sorted$roe, sorted$lsalary, pch=19)

> abline(reg2, col="red")

If you also run lines 29–31, you will graph a confidence interval around the fitted line.

In searching for influential data, a first step is to examine the distribution of the residuals:

> reg2.res<-resid(reg2)

> summary(reg2.res)

Line 34 in the R script gives the residual a name; you can examine the residuals using summary statistics (line 35), a stem-and-leaf plot (line 36) or a histogram (line 37).

The mean residual equals 0, however there are a few residuals that far from zero. A residual with magnitude exceeding 3× **Residual standard error** should be observed just once for every 333 observations. In this case, be concerned about residuals that are greater than 3 × 0.5553 = 1.666

Lines 38–39 plot the residuals versus the values of the explanatory variable:

> plot(CEOdta$roe, reg2.res, ylab="Residuals")

> abline(0,0)

Fortunately, what this figure shows is that the cases with large vertical deviations are centrally located along the horizontal dimension. According to Fox, such observations’ effects on the estimated intercept are greater than the effects on the estimated slope.

Another method for diagnosing influential data points is to examine numerical values. We will focus on three. First, line 41 creates ‘hat values’ and line 42 plots hat values against the residuals. You should be concerned when one observation has a ‘hat value’ greater than 2 times average leverage. Given the number of parameters estimated (2) and the sample size (*n* = 209) , average leverage is 0.0095. So be concerned if an observation’s hat value is greater than .0191.

Second, line 43 creates the studentized residuals and displays them using the stem() command. You should be concerned whenever an absolute value of discrepancy exceeds the *t* distribution’s critical value for 95% confidence and *n*–1 degrees of freedom; given the sample size (*n* = 209), *t*\* is about 1.96. Line 44 plots the studentized residuals against the hat values.

Third, line 45 creates the Cook’s distance values for each observation and displays them using the stem() command. Note the warning: **The decimal point is 2 digit(s) to the left of |**

You should be concerned that some cases have Cook’s distance greater than 0.0193 (= 4/(*n* – *p*)).

Line 46 plots the studentized residuals against the hat values, with the sizes of the circles proportional to the Cook’s distance values of each case. This is a neat plot, which should show you that cases with high leverage tend to have small studentized residuals and Cook’s distances, while cases with large studentized residuals tend to have little leverage.

We are about to move to another dataset. It is not a bad idea to clear the data you already used, the models you estimated, and any quantities you generated from R’s memory by typing rm(list=ls()) . The **Environment** window should be empty once you do this.

***C. Influential data points in state-level data***

The second dataset is based on chapter 1 of Tufte’s *Data Analysis for Politics and Policy*. Tufte examines the relationship between population density (population per square mile in 1967) and average death rates from automobile accidents (per 100,000 population, in 1966 through 1968).

Run line 48 to load the data, and run line 49 to plot the values of Deaths.per.100k against Density.

> data <- read.csv("C:/Users/Scott/accidents50.csv")

> plot(data$Density, data$Deaths.per.100k, pch=19)

For reasons that we will discuss later in the semester, we will be using the natural log of both variables, estimating what Wooldridge refers to as a log-log model (see Table 2.3). Run line 50 to plot the natural logarithm of Deaths.per.100k against the natural logarithm of Density.

> plot(log(data$Density), log(data$Deaths.per.100k), pch=19)

You should observe that most of the observations follow a roughly linear pattern, with lower levels of population density being associated with higher levels of deaths from auto accidents. However, one data point (Alaska) has extremely low population density and auto accident deaths.

Run line 51 to estimate a regression using all fifty observations and to examine the results.

> reg3<-lm(log(data$Deaths.per.100k) ~ log(data$Density))

> summary(reg3)

In log-log models, we interpret the slope coefficient as follows: a one ***percent*** change in the value of the independent variable yields a ***percent*** change in the value of the dependent variable. The estimated slope coefficient is = \_\_\_\_\_\_; this indicates that when population density increases by 1%, the number of vehicular fatalities will \_\_\_\_\_\_\_\_ (increase? decrease?) by \_\_\_\_\_\_ %.

Run line 53 to add the regression line to the plot, or simply type abline(reg3, col="red")

Does it appear to you that the regression line is being pulled away from the main mass of data due to the inclusion of one unusual observation (Alaska)?

There are multiple numerical measures that we can use to explore whether the one unusual case is an influential observation; here are four:

* line 55 creates ‘hat values’ to measure leverage;
* line 56 creates the studentized residuals to measure discrepancy;
* line 57 creates the Cook’s distance values for each observation;
* line 58 creates the DfFITS values for each observation; and

Line 59 creates a matrix of the first four measures, with a numerical code for the state (1 = Alabama, 2 = Alaska, etc.). Line 60 sorts the data in the new matrix by the hat value so you can focus on states with greatest leverage.

If you type the code in line 61, then R will show you the values of hat values, studentized residuals, Cook’s distance, and DfFits in the Console window. (Alternatively, if you click on the small image of a spreadsheet to the right of ⮊diag in the **Data** section of the **Environment** window, then you can examine the data for all states.)

Observation 2 (Alaska) is extreme in terms of both leverage and discrepancy. Be concerned whenever one observation has:

* a ‘hat value’ greater than 2 times the average leverage (which equals 0.04, given the number of parameters estimated (2) and the sample size (*n* = 50));
* a ‘studentized residual’ whose absolute value exceeds the critical value of the *t* distribution for 95% confidence and *n*–1 degrees of freedom; given the sample size (*n* = 50), *t*\* is about 2; for later reference, write the studentized residual for Alaska: \_\_\_\_\_\_\_
* Cook’s distance greater than 0.08 (= 4/*n*).

Maybe the best measure of a data point’s influence is how the model estimates changed due to including it in the sample. Line 63 creates the DfBETA values (effect on estimated intercept, then effect on estimated slope) and then displays the first six entries. Write down the values for case 2 (Alaska) below. Because these tell you how the estimates calculated with Alaska included were influenced by Alaska, you would compute the “new” values that would have been estimated if Alaska had been omitted by **subtracting** these values from the “old” estimates of reg3:

old intercept \_\_\_\_\_\_; *influence on* intercept \_\_\_\_\_\_\_; new intercept w/o Alaska \_\_\_\_\_\_\_

old slope \_\_\_\_\_\_\_\_\_; *influence on* slope \_\_\_\_\_\_\_\_\_\_; new slope w/o Alaska \_\_\_\_\_\_\_\_\_

Line 65 plots the *absolute value of* studentized residuals on the vertical axis against leverage on the horizontal axis. A vertical line is included at 0.08, indicating the threshold for concern about leverage; a horizontal line is included at 2 indicating the threshold for concern about discrepancy. You should observe that there is one case that exceeds both thresholds.

There are two simple ways to address the problem of an influential observation, once it has been diagnosed. The first method creates a dummy variable for Alaska (line 67) and adds that to the model specification (line 68); inspect the results by running line 69.

> dummy<-ifelse(data$State=="Alaska",c(1),c(0))

> reg4<-lm(log(data$Deaths.per.100k) ~ log(data$Density) + dummy)

> summary(reg4)

Run line 70; you can inspect the results by plotting the new regression line (in blue) against the old regression line (in black).

If you examine the estimated coefficients (line 69), what is the *t*-statistic for the Alaska? \_\_\_\_\_\_ Does this number look familiar? (It **should** equal the studentized residual that we found before.)

The second method is to re-estimate the regression model without Alaska; lines 71–72 provide code for doing this. Examine the regression results (line 73), which indicate that the estimated slope coefficient is \_\_\_\_\_. For comparison, the slope coefficient way back in line 51 was \_\_\_\_\_ . This new slope coefficient **should** equal the old coefficient **minus** the slope DfBETA.

A consequence of dropping Alaska is to raise our estimate of the impact of population density on deaths from accidents; the difference between the old coefficient and the new coefficients equals \_\_\_\_\_. For comparison, look back at the DFBETA value for Alaska for the slope coefficient.

Line 74 superimposes the regression line from the modified regression that omitted Alaska (green); compare it to the original regression that included Alaska (in black).

Another consequence should be that the model’s fit greatly improves when we drop Alaska from the analysis or include a dummy for Alaska. The original model yielded R2 = \_\_\_\_\_ , compared to R2 = \_\_\_\_\_ in the modified model. Over the next few weeks, as our attention turns to omitted variable bias and then to model fit, we will discuss how adding a dummy variable to control for one case or one factor has effects on coefficients, standard errors, R2, and so on.

To clear the **Environment**, type rm(list=ls()) or click on the broom icon.

To clear the **Console** window, type *l* while holding down the CTRL key.

1. The residual standard error is also much smaller, just as the standard deviation of *y* would be much smaller because you took the natural logarithm. This creates difficulties that you need to be aware of if you try to compare regression models with *y* as the DV to models with *ln y* as the DV. [↑](#footnote-ref-1)