05. Linear Regression Exercises

Pierre Hoonhout

Exercise 1

We consider the following exercise, taken from the book by Jeffrey Wooldridge, called *Introductory Econometrics: A Modern Approach, Fifth Edition*.

[A Log Wage Equation]

Using the same data as in Example 2.4, but using log(*wage*) as the dependent variable, we obtain the following relationship:

$$\widehat{\log(wage)} = 0.584 + 0.083\ educ \qquad \boxed{2.44}$$

$$n = 526,\ R^2 = 0.186.$$

The coefficient on *educ* has a percentage interpretation when it is multiplied by 100: $\widehat{wage}$ increases by 8.3% for every additional year of education. This is what economists mean when they refer to the "return to another year of education."

It is important to remember that the main reason for using the log of *wage* in (2.42) is to impose a constant percentage effect of education on *wage*. Once equation (2.42) is obtained, the natural log of *wage* is rarely mentioned. In particular, it is *not* correct to say that another year of education increases log(*wage*) by 8.3%.

The intercept in (2.42) is not very meaningful, because it gives the predicted log(*wage*), when *educ* = 0. The *R*-squared shows that *educ* explains about 18.6% of the variation in log(*wage*) (*not wage*). Finally, equation (2.44) might not capture all of the nonlinearity in the relationship between wage and schooling. If there are "diploma effects," then the twelfth year of education—graduation from high school—could be worth much more than the eleventh year. We will learn how to allow for this kind of nonlinearity in Chapter 7.

The data can be obtained by installing the wooldridge package, which contains all the datasets used in this book.

1a) Use the wooldridge package (and in particlar, the dataset wage1) to reproduce the estimation results given above.

Solution:

```
library(wooldridge)
data("wage1")
result <- lm(log(wage) ~ educ, data=wage1)
summary(result)
```

```
##
## Call:
## lm(formula = log(wage) ~ educ, data = wage1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.21158 -0.36393 -0.07263  0.29712  1.52339
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.583773   0.097336   5.998 3.74e-09 ***
## educ        0.082744   0.007567  10.935  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4801 on 524 degrees of freedom
## Multiple R-squared:  0.1858, Adjusted R-squared:  0.1843
## F-statistic: 119.6 on 1 and 524 DF,  p-value: < 2.2e-16
```

1b) Give a 95% confidence interval for β_{educ}. Is β_{educ} significantly different from zero?

As $\hat{\beta}_{educ}^{ols} \sim N\left(\beta_{educ}, \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right)$, the pivotal quantity is equal to:

$$T = \frac{\hat{\beta}_{educ}^{ols} - \beta_{educ}}{\sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\sum_{i=1}^n (x_i-\bar{x})^2}}} \sim t_{n-2}.$$

If we are not willing to assume that ε has a Normal distribution, the RSD of T becomes approximately $N(0,1)$. This can be shown using the central limit theorem and the so-called Slutsky-theorem, but this is beyond the scope of this course.

Using the assumption that the errors have a Normal distribution, we obtain the quantiles as follows:

```
qt(0.975, df=524)
```

```
## [1] 1.964502
```

We can now make the following probability statement:

$$P\left(-1.964502 < \frac{\hat{\beta}_{educ}^{ols} - \beta_{educ}}{\sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\sum_{i=1}^n (x_i-\bar{x})^2}}} < 1.964502\right) = 0.95$$

Rewriting this statement gives

$$P\left(\hat{\beta}_{educ}^{ols} - 1.964502 \times \sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\sum_{i=1}^n (x_i-\bar{x})^2}} < \beta_{educ} < \hat{\beta}_{educ}^{ols} + 1.964502 \times \sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\sum_{i=1}^n (x_i-\bar{x})^2}}\right) = 0.95$$

With $\hat{\beta}_{educ} = 0.082744$ and $\sqrt{\frac{\hat{\sigma^2}_\varepsilon}{\sum_{i=1}^n (x_i-\bar{x})^2}} = 0.007567$, we obtain

$$P\left(0.06787861 < \beta_{educ} < 0.09760939\right) = 0.95.$$

As 0 is not included in this interval, we reject the null hypothesis that $\beta_{educ} = 0$. In other words: β_{educ} is significantly different from 0. Or: We have statistical evidence that *educ* has a non-zero effect on *wage*.

1c) Test the null-hypothesis $H_0 : \beta_{educ} = 0$ with significance level $\alpha = 5\%$.

Solution:

As $\hat{\beta}^{ols}_{educ} \sim N\left(\beta_{educ}, \frac{\sigma^2_\varepsilon}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$, the pivotal statistic is equal to:

$$T = \frac{\hat{\beta}^{ols}_{educ} - 0}{\sqrt{\frac{\hat{\sigma}^2_\varepsilon}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \underset{H_0}{\sim} t_{n-2}.$$

If we are not willing to assume that ε has a Normal distribution, the RSD of T becomes approximately $N(0, 1)$. Using the assumption that the errors have a Normal distribution: as $T_{obs} = \frac{0.082744 - 0}{0.007567} = 10.935 > 1.964502$, we reject the null-hypothesis that educ has no effect on wage with $\alpha = 5\%$.

```
qt(0.975, df=524)
```

```
## [1] 1.964502
```

```
1 - pt(10.935, df=524) + pt(-10.935, df=524)
```

```
## [1] 1.640305e-25
```

We already knew this from the 95% confidence interval. Note that $T_{obs} = 10.935$ is given in the output of the lm command. As lm reports the p-value, we can read off the test-result directly from the output, no matter the value of α.

Exercise 2

Example 2.3 in the Wooldridge book is as follows:

EXAMPLE 2.3 CEO SALARY AND RETURN ON EQUITY

For the population of chief executive officers, let y be annual salary (*salary*) in thousands of dollars. Thus, $y = 856.3$ indicates an annual salary of \$856,300, and $y = 1{,}452.6$ indicates a salary of \$1,452,600. Let x be the average return on equity (*roe*) for the CEO's firm for the previous three years. (Return on equity is defined in terms of net income as a percentage of common equity.) For example, if $roe = 10$, then average return on equity is 10%.

To study the relationship between this measure of firm performance and CEO compensation, we postulate the simple model

$$salary = \beta_0 + \beta_1 roe + u.$$

The slope parameter β_1 measures the change in annual salary, in thousands of dollars, when return on equity increases by one percentage point. Because a higher *roe* is good for the company, we think $\beta_1 > 0$.

The data set CEOSAL1.RAW contains information on 209 CEOs for the year 1990; these data were obtained from *Business Week* (5/6/91). In this sample, the average annual salary is \$1,281,120, with the smallest and largest being \$223,000 and \$14,822,000, respectively. The average return on equity for the years 1988, 1989, and 1990 is 17.18%, with the smallest and largest values being 0.5 and 56.3%, respectively.

Using the data in CEOSAL1.RAW, the OLS regression line relating *salary* to *roe* is

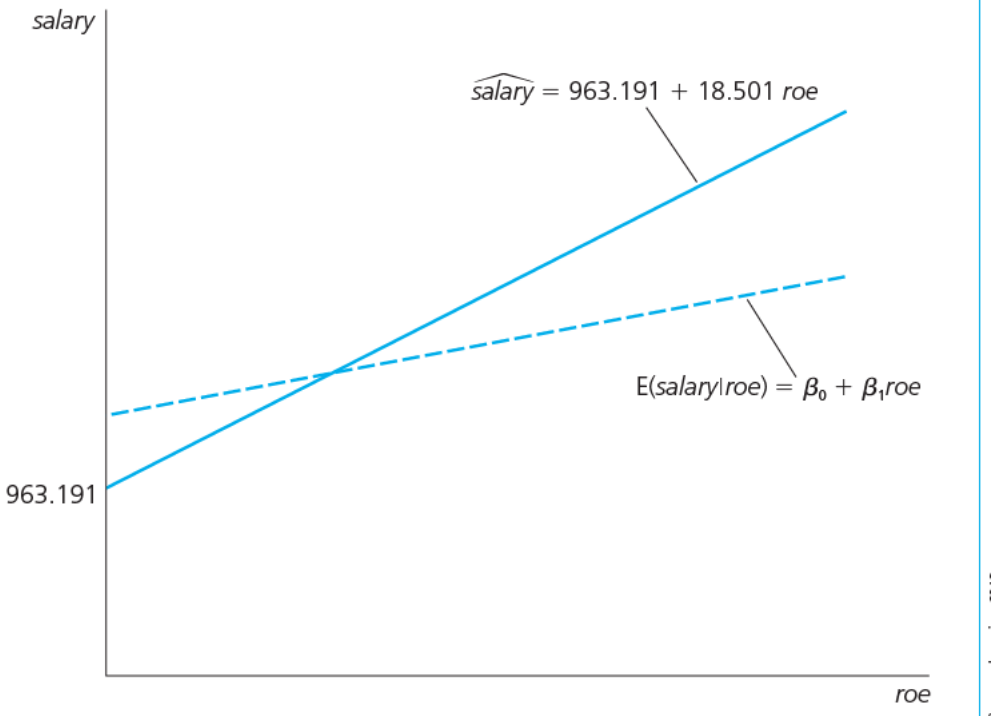
$$\widehat{salary} = 963.191 + 18.501\, roe \qquad [2.26]$$
$$n = 209,$$

where the intercept and slope estimates have been rounded to three decimal places; we use "*salary* hat" to indicate that this is an estimated equation. How do we interpret the equation? First, if the return on equity is zero, $roe = 0$, then the predicted *salary* is the intercept, 963.191, which equals \$963,191 since *salary* is measured in thousands. Next, we can write the predicted change in salary as a function of the change in *roe*: $\Delta\widehat{salary} = 18.501\,(\Delta roe)$. This means that if the return on equity increases by one percentage point, $\Delta roe = 1$, then *salary* is predicted to change by about 18.5, or \$18,500. Because (2.26) is a linear equation, this is the estimated change regardless of the initial salary.

We can easily use (2.26) to compare predicted salaries at different values of *roe*. Suppose $roe = 30$. Then $\widehat{salary} = 963.191 + 18.501(30) = 1{,}518{,}221$, which is just over \$1.5 million. However, this does *not* mean that a particular CEO whose firm had a $roe = 30$ earns \$1,518,221. Many other factors affect salary. This is just our prediction from the OLS regression line (2.26). The estimated line is graphed in Figure 2.5, along with the population regression function E(*salary*|*roe*). We will never know the PRF, so we cannot tell how close the SRF is to the PRF. Another sample of data will give a different regression line, which may or may not be closer to the population regression line.

The difference between the conditional mean function (for some β_0 and β_1) and the estimated regression line is given in the following graph:

FIGURE 2.5 The OLS regression line $\widehat{salary} = 963.191 + 18.501\ roe$ and the (unknown) population regression function.



2a) Use the wooldridge package (and in particlar, the dataset ceosal1) to reproduce the estimation results given above.

Solution:

The estimation results for the model in levels are:

```
library(wooldridge)
result1 <- lm(salary ~ roe, data=ceosal1)
summary(result1)
```

```
##
## Call:
## lm(formula = salary ~ roe, data = ceosal1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1160.2  -526.0  -254.0   138.8 13499.9
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   963.19     213.24   4.517 1.05e-05 ***
## roe            18.50      11.12   1.663   0.0978 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1367 on 207 degrees of freedom
## Multiple R-squared:  0.01319,    Adjusted R-squared:  0.008421
## F-statistic: 2.767 on 1 and 207 DF,  p-value: 0.09777
```

2b) Give a 99% confidence interval for β_{roe}. Is β_{roe} significantly different from zero using $\alpha = 1\%$?

Solution:

As $\hat{\beta}_{roe}^{ols} \sim N\left(\beta_{roe}, \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$, the pivotal quantity is equal to:

$$T = \frac{\hat{\beta}_{roe}^{ols} - \beta_{roe}}{\sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim t_{n-2}.$$

If we are not willing to assume that ε has a Normal distribution, the RSD of T becomes approximately $N(0, 1)$. Using the assumption that the errors have a Normal distribution, we obtain the quantiles as follows:

```
qt(0.995, df=207)
```

```
## [1] 2.599788
```

We can make the following probability statement:

$$P\left(-2.599788 < \frac{\hat{\beta}_{roe}^{ols} - \beta_{roe}}{\sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} < 2.599788\right) = 0.99.$$

Rewriting this statement gives

$$P\left(\hat{\beta}_{roe}^{ols} - 2.599788 \times \sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} < \beta_{roe} < \hat{\beta}_{roe}^{ols} + 2.599788 \times \sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}\right) = 0.99.$$

With $\hat{\beta}_{roe} = 18.50$ and $\sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} = 11.12$, we obtain

```
18.50 - 2.599788*11.12
```

```
## [1] -10.40964
```

```
18.50 + 2.599788*11.12
```

```
## [1] 47.40964
```

Hence,

$$P\left(-10.40964 < \beta_{roe} < 47.40964\right) = 0.99.$$

As 0 is included in this interval, we do not reject the null hypothesis that $\beta_{educ} = 0$ at the $\alpha = 1\%$ level. In other words: β_{roe} is not significantly different from 0. We have no statistical evidence that *roe* has an effect on *wage*.

2c) Test the null-hypothesis $H_0 : \beta_{roe} = 0$ using $\alpha = 5\%$.

Solution:

As $\hat{\beta}_{roe}^{ols} \sim N\left(\beta_{roe}, \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$, the pivotal statistic is equal to:

$$T = \frac{\hat{\beta}_{roe}^{ols} - 0}{\sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \underset{H_0}{\sim} t_{n-2}.$$

If we are not willing to assume that ε has a Normal distribution, the RSD of T becomes approximately $N(0, 1)$.

Using the assumption that the errors have a Normal distribution: as $T_{obs} = \frac{18.50 - 0}{11.12} = 1.663 \not> 1.97149$, we do not reject the null-hypothesis that roe has no effect on salary.

```
qt(0.975, df=207)
```

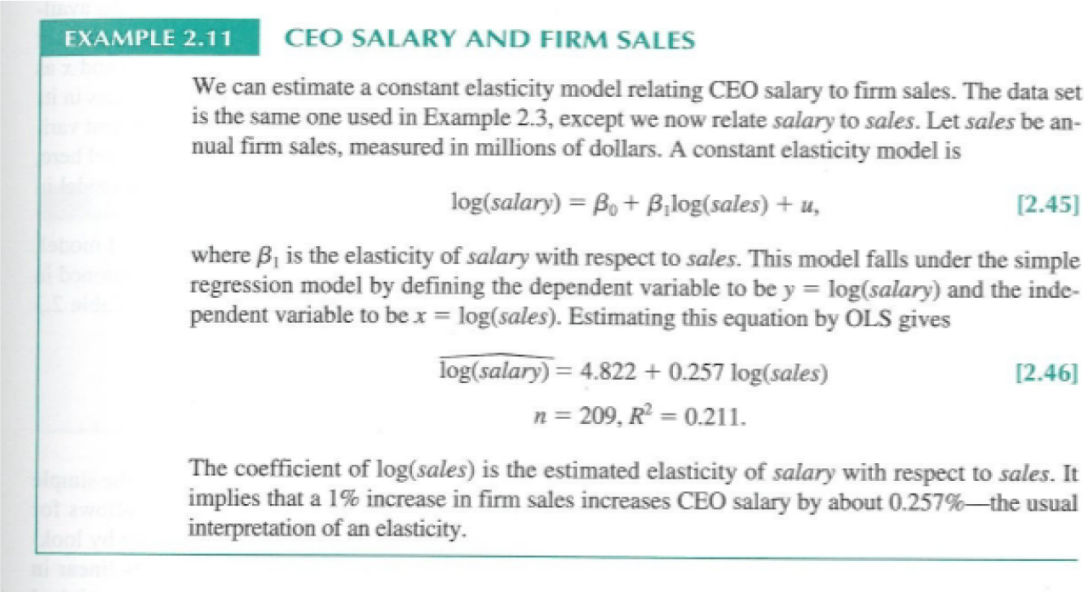
```
## [1] 1.97149
```

```
1 - pt(1.663, df=207) + pt(-1.663, df=207)
```

```
## [1] 0.09782576
```

Note that $T_{obs} = 1.663$ is given in the output of the lm command. As lm reports the p-value, we can read off the test-result directly from the output, no matter the value of α. We cannot infer the test-result from the confidence interval, as the latter used $\alpha = 1\%$.

Exercise 3



EXAMPLE 2.11 **CEO SALARY AND FIRM SALES**

We can estimate a constant elasticity model relating CEO salary to firm sales. The data set is the same one used in Example 2.3, except we now relate *salary* to *sales*. Let *sales* be annual firm sales, measured in millions of dollars. A constant elasticity model is

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + u, \qquad [2.45]$$

where β_1 is the elasticity of *salary* with respect to *sales*. This model falls under the simple regression model by defining the dependent variable to be $y = \log(salary)$ and the independent variable to be $x = \log(sales)$. Estimating this equation by OLS gives

$$\widehat{\log(salary)} = 4.822 + 0.257 \log(sales) \qquad [2.46]$$

$$n = 209, R^2 = 0.211.$$

The coefficient of $\log(sales)$ is the estimated elasticity of *salary* with respect to *sales*. It implies that a 1% increase in firm sales increases CEO salary by about 0.257%—the usual interpretation of an elasticity.

3a) Use the wooldridge package (and in particlar, the dataset ceosal1) to reproduce the estimation results given above.

Solution:

Using the log-log model we obtain:

```
library(wooldridge)
result2 <- lm(log(salary) ~ log(sales), data=ceosal1)
summary(result2)
```

```
##
## Call:
## lm(formula = log(salary) ~ log(sales), data = ceosal1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.01038 -0.28140 -0.02723  0.21222  2.81128
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.82200    0.28834  16.723  < 2e-16 ***
## log(sales)   0.25667    0.03452   7.436  2.7e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5044 on 207 degrees of freedom
## Multiple R-squared:  0.2108, Adjusted R-squared:  0.207
## F-statistic:  55.3 on 1 and 207 DF,  p-value: 2.703e-12
```

3b) Give a 95% confidence interval for $\beta_1 = \beta_{log(sales)}$. Is β_1 significantly different from zero?

**Solution:*

As $\hat{\beta}_1^{ols} \sim N\left(\beta_1, \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$, the pivotal quantity is equal to:

$$T = \frac{\hat{\beta}_1^{ols} - \beta_1}{\sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim t_{n-2}.$$

If we are not willing to assume that ε has a Normal distribution, the RSD of T becomes approximately $N(0, 1)$. Using the assumption that the errors have a Normal distribution, we obtain the quantiles as follows:

```
qt(0.975, df=207)
```

```
## [1] 1.97149
```

We can make the following probability statement:

$$P\left(-1.97149 < \frac{\hat{\beta}_1^{ols} - \beta_1}{\sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} < 1.97149\right) = 0.95.$$

Rewriting this statement gives

$$P\left(\hat{\beta}_1^{ols} - 1.97149 \times \sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} < \beta_1 < \hat{\beta}_1^{ols} + 1.97149 \times \sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}\right) = 0.95.$$

With $\hat{\beta}_{educ} = 0.25667$ and $\sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} = 0.03452$, we obtain

```
0.25667 - 1.97149 * 0.03452
```

```
## [1] 0.1886142
```

```
0.25667 + 1.97149 * 0.03452
```

```
## [1] 0.3247258
```

Hence,

$$P\left(0.1886142 < \beta_1 < 0.3247258\right) = 0.95.$$

As 0 is not included in this interval, we reject the null hypothesis that $\beta_1 = 0$ at the $\alpha = 5\%$ level. In other words: β_1 is significantly different from 0. We have statistical evidence that *sales* has a non-zero effect on *salary*.

3c) Test the null-hypothesis $H_0 : \beta_{sales} = 0.3$, using $\alpha = 5\%$.

Solution:

As $\hat{\beta}_1^{ols} \sim N\left(\beta_1, \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$, the pivotal statistic is equal to:

$$T = \frac{\hat{\beta}_1^{ols} - 0.3}{\sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \underset{H_0}{\sim} t_{n-2}.$$

If we are not willing to assume that ε has a Normal distribution, the RSD of T becomes approximately $N(0, 1)$.

Using the assumption that the errors have a Normal distribution: as $T_{obs} = \frac{0.25667 - 0.3}{0.03452} = -1.255214 \not< -1.97149$, we do not reject the null-hypothesis that $\beta_1 = 0.3$.

```
qt(0.975, df=207)
```

```
## [1] 1.97149
```

```
1 - pt(1.255214, df=207) + pt(-1.255214, df=207)
```

```
## [1] 0.2108163
```

Note that $T_{obs} = -1.255214$ is *not* given in the output of the lm command. Only the test-statistic for $\beta_1 = 0$ is reported. We could have seen the result of this test from the confidence interval for β_1: it includes the value 0.3.

Exercise 4

Consider the linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$, with the standard assumptions. Derive the OLS estimators for the constant and the slope.

Solution:

See the lecture notes.