Dragos Ailoae

Applied Statistics and Econometrics I

2021-10-12

Homework 2 - Suggested Solutions

Problem 1:

For the regression model $Y = \beta_0 + \beta_1 X + \varepsilon$,

a. Show that the least squares normal equations imply that

$$\sum_{i=1}^{N} e_i = 0 \text{ and } \sum_{i=1}^{N} x_i e_i = 0$$

Remember, the OLS optimization problem is to find parameters $\hat{\beta}_0, \hat{\beta}_1$ in order to minimize the sum of deviations in the vertical distance between our observed y_i and the predicted $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. We call this difference the residual $e_i = y_i - \hat{y}_i$ and we measure the deviation as the square of this difference in order to make it a positive value (absolute values are more cumbersome in calculus).

The problem is defined as:

$$\min_{\hat{\beta}_0, \hat{\beta}_1} Q = \sum_{i=1}^{N} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$$

Where Q is also known as the **sum of squared residuals** (SSR) or **residuals sum of squares**(RSS).

To solve the minimization problem, we need to take the partial derivatives with respect to our parameters and set them equal to 0. This gives us:

$$\frac{\partial Q}{\partial \hat{\beta}_0} = \sum_{i=1}^{N} -2 \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0 \tag{1}$$

$$\frac{\partial Q}{\partial \hat{\beta}_1} = \sum_{i=1}^{N} -2 x_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0 \tag{2}$$

Equations (1) and (2) are called the **normal equations**, they are basically the system of equations we solve to obtain our parameters $\hat{\beta}_0, \hat{\beta}_1$.

So problem 1a is basically asking: with the normal equations as given, prove these two corollaries $\sum_{i=1}^{N} e_i = 0$ and $\sum_{i=1}^{N} x_i e_i = 0$.

To prove $\sum_{i=1}^{N} e_i = 0$ we note that since $e_i = y_i - \hat{y}_i$ and $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ we can write

$$e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \tag{3}$$

so we essentially need to prove that

$$\sum_{i=1}^{N} e_i = \sum_{i=1}^{N} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0 \tag{4}$$

But this is just normal equation (1) i.e. dividing both sides of equation (1) by -2 gives us:

$$\sum_{i=1}^{N} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0 \tag{5}$$

Since (1) is true then (4) must also be true, QED.

Similarly, to prove $\sum_{i=1}^{N} x_i e_i = 0$ we use (3) to redefine the expression as:

$$\sum_{i=1}^{N} x_i e_i = \sum_{i=1}^{N} x_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0 \tag{6}$$

Then we note that dividing both sides of normal equation (2) by -2 gives us:

$$\sum_{i=1}^{N} x_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0 \tag{7}$$

Since (2) is true then (6) must be true, QED.

b. Show that the solution for the constant term is:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Let's start with the simplified version of normal equation 1 which we saw above as equation (5). It follows that

$$\begin{aligned}
\sum_{i=1}^{N} \hat{\beta}_0 &= \sum_{i=1}^{N} \left(y_i - \hat{\beta}_1 x_i \right) \\
N\hat{\beta}_0 &= \sum_{i=1}^{N} \left(y_i - \hat{\beta}_1 x_i \right) \\
\hat{\beta}_0 &= \frac{1}{N} \sum_{i=1}^{N} y_i - \frac{1}{N} \sum_{i=1}^{N} \hat{\beta}_1 x_i \\
\hat{\beta}_0 &= \frac{1}{N} \sum_{i=1}^{N} y_i - \hat{\beta}_1 \frac{1}{N} \sum_{i=1}^{N} x_i \\
\hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X}
\end{aligned}$$

Problem 2:

Show that the assumption $E(\varepsilon \mid x) = 0$ implies that $\operatorname{Cov}(x, \varepsilon) = 0$.

Per Greene Appendix:

THEOREM B.2 Covariance
In any bivariate distribution,

$$\operatorname{Cov}[x, y] = \operatorname{Cov}_x[x, E[y \mid x]] = \int_x (x - E[x])\, E[y \mid x] f_x(x)\, dx. \qquad \textbf{(B-67)}$$

(Note that this is the covariance of x and a function of x.)

In our case

$$\operatorname{Cov}[x, \varepsilon] = \operatorname{Cov}_x[x, E[\epsilon \mid x]] = \int_x (x - E[x]) E[\varepsilon \mid x] f_x(x) dx$$

Since $E[\varepsilon \mid x] = 0$ it follows that $Cov[x, \varepsilon] = 0$.

Alternately, we note that

$$Cov[x, \varepsilon] = E[x\varepsilon] - E[x]E[\varepsilon] = E[x\varepsilon] - E[x] \times 0 = E[x\varepsilon] \tag{8}$$

By the law of iterated expectations (see Greene Appendix B-66):

$$E[x\varepsilon] = E[E[x\varepsilon \mid x]] = E[xE[\varepsilon \mid x]] = E[x \times 0] = 0 \tag{9}$$

It follows then that $Cov[x, \varepsilon] = 0$

Problem 3:

Suppose we have the following data on GPA and Income.

GPA	Income
(Grade Point Average)	(Income of parents in $1000)
4.0	21.0
3.0	15.0
3.5	15.0
2.0	9.0
3.0	12.0
3.5	18.0
2.5	6.0
2.5	12.0

Calculate the regression of income on GPA and compare it with the regression of GPA on income. Why are the two different?

Method 1: Summations

$$\hat{\beta}_0 = \frac{1}{N}\sum_{i=1}^{N} y_i - \hat{\beta}_1 \frac{1}{N}\sum_{i=1}^{N} x_i = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} = \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)}$$

Model 1: Income ~ GPA

We'll define the population model as $Income = \beta_0 + \beta_1 GPA + \varepsilon$

In [1]:

```
GPA <- c(4.0,3.0,3.5,2.0,3.0,3.5,2.5,2.5)
Income <- c(21.0, 15.0, 15.0, 9.0, 12.0, 18.0, 6.0, 12.0)
m1.beta.hat.1 <- cov(GPA,Income) / var(GPA)
m1.beta.hat.1
```

6.5

In [2]:

```
m1.beta.hat.0 <-mean(Income) - m1.beta.hat.1 * mean(GPA)
m1.beta.hat.0
```

-6

So for model 1 we have $\hat{\beta}_0$ = -6 and $\hat{\beta}_1$ = 6.5. Let's look at model 2:

Model 2: GPA ~ Income

We'll define the population model as $GPA = \beta_0 + \beta_1 Income + \varepsilon$

In [3]:

```
m2.beta.hat.1 <- cov(GPA,Income) / var(Income)
round(m2.beta.hat.1,3)
```

0.12

In [4]:

```
m2.beta.hat.0 <-mean(GPA) - m2.beta.hat.1 * mean(Income)
m2.beta.hat.0
```

1.375

Conclusion: The reason for the difference is that, while both calculations for $\hat{\beta}_1$ are based on the covariance between the same two variables, in model 1 we scale the covariance by the variance of GPA while in model 2 we scale it by the variance of Income. Different y-intercepts, $\hat{\beta}_0$'s, are a consequence of these different slopes

Method 2: Matrices

The OLS estimator in matrix form is $\hat{\beta} = (X'X)^{-1}X'y$ so for model 1 we have:

Model 1: Income ~ GPA

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \left(\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 4.0 & 3.0 & 3.5 & 2.0 & 3.0 & 3.5 & 2.5 & 2.5 \end{bmatrix} \times \begin{bmatrix} 1 & 4.0 \\ 1 & 3.0 \\ 1 & 3.5 \\ 1 & 2.0 \\ 1 & 3.0 \\ 1 & 3.5 \\ 1 & 2.5 \\ 1 & 2.5 \end{bmatrix} \right)^{-1} \times \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 4.0 & 3.0 & 3.5 & 2.0 & 3.0 & 3.5 & 2.5 & 2.5 \end{bmatrix} \begin{bmatrix} 21 \\ 15 \\ 15 \\ 9 \\ 12 \\ 18 \\ 6 \\ 12 \end{bmatrix} \Bigg)$$

We'll first need to add the intercept column of 1s (also known as the 'bias') to our X (or data) matrix.

In [5]:
```
intercept <- rep(1, length(GPA))
intercept
```

1 · 1 · 1 · 1 · 1 · 1 · 1 · 1

Next we append it to our independent variable vector:

In [6]:
```
X <- cbind(intercept, GPA)
X
```

A matrix: 8 × 2 of type dbl

intercept	GPA
1	4.0
1	3.0
1	3.5
1	2.0
1	3.0
1	3.5
1	2.5
1	2.5

Setting our dependent variable y as Income we can then calculate our betas:

In [7]:
```
y <- Income
m1.betas <- solve(t(X) %*% X) %*% t(X) %*% y
m1.betas
```

A matrix: 2 × 1 of type dbl

intercept	-6.0
GPA	6.5

To print separately:

In [8]:
```
paste0("Model 1: Income ~ GPA")
paste0("Intercept is ",m1.betas[1])
paste0("GPA coefficient is ",round(m1.betas[2],3))
```

'Model 1: Income ~ GPA'

'Intercept is -6'

'GPA coefficient is 6.5'

Model 2: GPA ~ Income

In [9]:
```
X <- cbind(intercept, Income)
y <- GPA
m2.betas <- solve(t(X) %*% X) %*% t(X) %*% y
round(m2.betas,3)
```

A matrix: 2 × 1 of

type dbl	
intercept	1.375
Income	0.120

In [10]:

```
paste0("Model 2: GPA ~ Income")
paste0("Intercept is ",m2.betas[1])
paste0("Income coefficient is ",round(m2.betas[2],3))
```

'Model 2: GPA ~ Income'

'Intercept is 1.375'

'Income coefficient is 0.12'

Problem 4:

What happens to the least squares slope and intercept estimates when all observations on the independent variable are identical? Show how you arrived at your answer.

Note that

$$\hat{\beta}_1 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} = \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)}$$

Since all x's are the same that means $x_i = \bar{x}$ the denominator is 0 (no variance) meaning we cannot estimate our coefficients. Or, in matrix terms, $\hat{\beta} = (X'X)^{-1}X'y$ but the $(X'X)$ matrix is singular therefore non-invertible.

To convince ourselves let's try an example:

In [11]:

```
X <- matrix(c(1, 3,
              1, 3,
              1, 3),byrow=T, nrow=3)
X
```

A matrix: 3 × 2 of type dbl

1	3
1	3
1	3

In [12]:

```
XprimeX <- t(X) %*% X
XprimeX
```

A matrix: 2 × 2 of type dbl

3	9
9	27

In [13]:

```
solve(XprimeX)
```

```
Error in solve.default(XprimeX): Lapack routine dgesv: system is exactly singular: U[2,2] = 0
Traceback:

1. solve(XprimeX)
2. solve.default(XprimeX)
```

Inverting fails because "system is exactly singular"

Also note that the columns of X are exactly collinear (more of this later in the course) so the X matrix is not full rank.

In [15]:

```
options(warn=-1) # Suppress warnings
qui <- suppressPackageStartupMessages # quiet! - suppress library load messages
qui(if(!require(Matrix)){install.packages('Matrix')}) # for 'rankMatrix' function

paste('Number of Columns: ',dim(X)[2])
paste('Matrix Rank: ',rankMatrix(X))
```

'Number of Columns: 2'

'Matrix Rank: 1'

Problem 5:

Suppose that the coefficient in the regression of income on GPA in problem 3 was -1.7. How would you interpret it?

A one point increase in GPA is associated with a $1700 decrease in parents' income.

Problem 6:

Assume that least squares estimates are obtained for the relationship $Y=\beta_0+\beta_1 X+\varepsilon$. After the work is completed, it is decided to multiply the units of the X variable by a factor of 10 . What will happen to the least squares slope and intercept?

Remember,

$$\hat{\beta}_0=\frac{1}{N} \sum_{i=1}^N y_i-\hat{\beta}_1 \frac{1}{N} \sum_{i=1}^N x_i$$

$$\hat{\beta}_1=\frac{\frac{1}{N} \sum_{i=1}^N\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\frac{1}{N} \sum_{i=1}^N\left(x_i-\bar{x}\right)^2}$$

Multiplying X by 10 results in the following change to the slope coefficient:

$$\hat{\beta}_1^*=\frac{\frac{1}{N} \sum_{i=1}^N\left(10 x_i-10 \bar{x}\right)\left(y_i-\bar{y}\right)}{\frac{1}{N} \sum_{i=1}^N\left(10 x_i-10 \bar{x}\right)^2}=\frac{10 \frac{1}{N} \sum_{i=1}^N\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{100 \frac{1}{N} \sum_{i=1}^N\left(x_i-\bar{x}\right)^2}=\frac{1}{10} \hat{\beta}_1$$

While for the intercept:

$$\begin{aligned}
\hat{\beta}_0^* & =\frac{1}{N} \sum_{i=1}^N y_i-\hat{\beta}_1^* \frac{1}{N} \sum_{i=1}^N 10 x_i \\
& =\frac{1}{N} \sum_{i=1}^N y_i-10 \hat{\beta}_1^* \frac{1}{N} \sum_{i=1}^N x_i \\
& =\frac{1}{N} \sum_{i=1}^N y_i-10 \frac{1}{10} \hat{\beta}_1 \frac{1}{N} \sum_{i=1}^N x_i \\
& =\frac{1}{N} \sum_{i=1}^N y_i-\hat{\beta}_1 \frac{1}{N} \sum_{i=1}^N x_i \\
& =\hat{\beta}_0
\end{aligned}$$

Conclusion: If we only scale X the slope coefficient gets rescaled (inverse proportionally) while the intercept stays the same.

What happens to the slope and intercept if both the X and Y variables are multiplied by a factor of 10?

In this case:

$$\hat{\beta}_1^*=\frac{\frac{1}{N} \sum_{i=1}^N\left(10 x_i-10 \bar{x}\right)\left(10 y_i-10 \bar{y}\right)}{\frac{1}{N} \sum_{i=1}^N\left(10 x_i-10 \bar{x}\right)^2}=\frac{100 \frac{1}{N} \sum_{i=1}^N\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{100 \frac{1}{N} \sum_{i=1}^N\left(x_i-\bar{x}\right)^2}=\hat{\beta}_1$$

While for the intercept:

$$\begin{aligned}
\hat{\beta}_0^* & =\frac{1}{N} \sum_{i=1}^N 10 y_i-\hat{\beta}_1^* \frac{1}{N} \sum_{i=1}^N 10 x_i \\
& =10 \frac{1}{N} \sum_{i=1}^N y_i-10 \hat{\beta}_1 \frac{1}{N} \sum_{i=1}^N x_i \\
& =10\left(\frac{1}{N} \sum_{i=1}^N y_i-\hat{\beta}_1 \frac{1}{N} \sum_{i=1}^N x_i\right) \\
& =10 \hat{\beta}_0
\end{aligned}$$

Conclusion: If we scale both X an Y by the same factor the slope coefficient does not change but the intercept gets proportionally rescaled.