625.661 - Homework One

Eric Niblock

January 28, 2022

1. In a simple linear regression analysis, n independent paired data (y_1, x_1), ..., (y_n, x_n) are to be fitted to the model M1 given by,

$$y_i = \beta_0 + \beta_1(x_i - \mu) + \epsilon_i \text{ for } i = 1, ..., n$$

where the regressor x is a random variable with mean μ_x and variance σ_x^2. Conditional on x, the random error ϵ has mean zero and variance σ^2 which does not depend on x, β_0, β_1. The μ is a given real number (that is, the value of μ is known). The values of μ_x, σ_x^2, σ^2, β_0, and β_1 are all unknown. Before the values of the n independent paired data for (y, x) are available, we need to construct estimators for the parameters, μ_x, σ_x^2, σ^2, β_0, and β_1, respectively.

(a) Construct the ordinary least squares (OLS) estimator of β_1. Is the OLS estimator unbiased for β_1? Why or why not?

For simplification, we define $z_i = x_i - \mu$. Then, in order to find the OLS estimator of β_1 we must minimize the sum of the squared errors,

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 z_i)^2 \tag{1}$$

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 z_i) = 0 \tag{2}$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} z_i (y_i - \beta_0 - \beta_1 z_i) = 0 \tag{3}$$

We are left with the following set of equations,

$$\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1 z_i\right)=0 \quad \text{and} \quad \sum_{i=1}^{n} z_i\left(y_i-\hat{\beta}_0-\hat{\beta}_1 z_i\right)=0 \tag{4}$$

Beginning with the equation on the left of (4), we have that,

$$n\bar{y}-n\hat{\beta}_0-n\hat{\beta}_1\bar{z}=0 \tag{5}$$

$$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{z} \tag{6}$$

Now the right of equation (4) provides us with $\hat{\beta}_1$,

$$\sum_{i=1}^{n} z_i y_i - z_i(\bar{y}-\hat{\beta}_1\bar{z})-\hat{\beta}_1 z_i^2=0 \tag{7}$$

$$\sum_{i=1}^{n} z_i y_i - z_i\bar{y}+z_i\hat{\beta}_1\bar{z}-\hat{\beta}_1 z_i^2=0 \tag{8}$$

$$\sum_{i=1}^{n} z_i y_i - \bar{y}\sum_{i=1}^{n} z_i+\hat{\beta}_1\bar{z}\sum_{i=1}^{n} z_i-\hat{\beta}_1\sum_{i=1}^{n} z_i^2=0 \tag{9}$$

$$\sum_{i=1}^{n} z_i y_i - n\bar{y}\bar{z}-\hat{\beta}_1\left(-n\bar{z}^2+\sum_{i=1}^{n} z_i^2\right)=0 \tag{10}$$

$$\hat{\beta}_1=\frac{\left(\sum_{i=1}^{n} z_i y_i\right)-n\bar{y}\bar{z}}{\left(\sum_{i=1}^{n} z_i^2\right)-n\bar{z}^2} \tag{11}$$

With slight modification and further derivation, this becomes,

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}(z_i-\bar{z})(y_i-\bar{y})}{\sum_{i=1}^{n}\left(z_i-\bar{z}\right)^2} \tag{12}$$

Substituting in for z_i and $\bar{z}$ yields,

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}(x_i-\mu-\bar{x}+\mu)(y_i-\bar{y})}{\sum_{i=1}^{n}\left(x_i-\mu-\bar{x}+\mu\right)^2}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2} \tag{13}$$

In other words, shifting every point along the x-axis does not change the slope of the fitted line (which is to be expected).

The result is unbiased. We have from (11) that,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(z_i - \bar{z})y_i}{(\sum_{i=1}^{n} z_i^2) - n\bar{z}^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{(\sum_{i=1}^{n}(x_i - \mu)^2) - n(\bar{x} - \mu)^2} \tag{14}$$

$$E[\hat{\beta}_1|x_i] = \frac{\sum_{i=1}^{n}(x_i - \bar{x})E[y_i|x_i]}{(\sum_{i=1}^{n}(x_i - \mu)^2) - n(\bar{x} - \mu)^2} \tag{15}$$

$$E[\hat{\beta}_1|x_i] = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1(x_i - \mu))}{(\sum_{i=1}^{n}(x_i - \mu)^2) - n(\bar{x} - \mu)^2} \tag{16}$$

$$E[\hat{\beta}_1|x_i] = \frac{\sum_{i=1}^{n} x_i\beta_0 - \bar{x}\beta_0 + x_i\beta_1(x_i - \mu) - \bar{x}\beta_1(x_i - \mu)}{(\sum_{i=1}^{n}(x_i - \mu)^2) - n(\bar{x} - \mu)^2} \tag{17}$$

$$E[\hat{\beta}_1|x_i] = \frac{\sum_{i=1}^{n} x_i\beta_1(x_i - \mu) - \bar{x}\beta_1(x_i - \mu)}{(\sum_{i=1}^{n}(x_i - \mu)^2) - n(\bar{x} - \mu)^2} \tag{18}$$

$$E[\hat{\beta}_1|x_i] = \frac{\beta_1 \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \mu)}{(\sum_{i=1}^{n}(x_i - \mu)^2) - n(\bar{x} - \mu)^2} \tag{19}$$

Simplifying the numerator and denominator yields,

$$E[\hat{\beta}_1|x_i] = \frac{\beta_1 \sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \beta_1 \tag{20}$$

And since,

$$E[\hat{\beta}_1] = E[E[\hat{\beta}_1|x_i]] = E[\beta_1] = \beta_1 \tag{21}$$

We have that $\hat{\beta}_1$ is an unbiased estimator for β_1.

(b) Construct an estimator for σ^2. Is the estimator unbiased for σ^2? Why or why not?

We begin with the following estimate of sample variance,

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(\epsilon_i - \bar{\epsilon})^2 \tag{22}$$

Furthermore, we have that,

$$\bar{\epsilon} = \frac{1}{n}\sum_{i=1}^{n}\epsilon_i = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{\beta}_0 - \hat{\beta}_1(x_i - \mu) \tag{23}$$

Though we can replace $\hat{\beta}_0$ from (6),

$$\bar{\epsilon} = \frac{1}{n}\sum_{i=1}^{n} y_i - \bar{y} + \hat{\beta}_1(\bar{x} - \mu) - \hat{\beta}_1(x_i - \mu) = 0 \tag{24}$$

So, we can simply write our estimator for σ^2 as,

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{25}$$

This estimator is biased, though we can produce an unbiased estimator. We begin by rearranging our calculation of $n\hat{\sigma}^2$,

$$
\begin{aligned}
\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 &= \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 z_i)^2 \\
&= \sum_{i=1}^{n}(y_i - \bar{y} + \hat{\beta}_1 \bar{z} - \hat{\beta}_1 z_i)^2 \\
&= \sum_{i=1}^{n}[(y_i - \bar{y}) - \hat{\beta}_1 (z_i - \bar{z})]^2 \\
&= \sum_{i=1}^{n}(y_i - \bar{y})^2 - 2\hat{\beta}_1 (z_i - \bar{z})(y_i - \bar{y}) + (z_i - \bar{z})^2 \\
&= \left(\sum_{i=1}^{n} y_i^2\right) - n\bar{y}^2 - S_{zz}\hat{\beta}_1^2
\end{aligned} \tag{26}
$$

Now, taking the expected value yields,

$$
\begin{aligned}
E\left[\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right] &= E\left[\sum_{i=1}^{n} y_i^2\right] - E[n\bar{y}^2] - E[S_{zz}\hat{\beta}_1^2] \\
&= \sum_{i=1}^{n} E[y_i^2] - nE[\bar{y}^2] - S_{zz}E[\hat{\beta}_1^2] \\
&= \left(\sum_{i=1}^{n} Var(y_i) + E[y_i]^2\right) + n\left(Var(\bar{y}) + E[\bar{y}]^2\right) \\
&\quad + S_{zz}\left(Var(\hat{\beta}_1) + E[\hat{\beta}_1]^2\right)
\end{aligned} \tag{27}
$$

An extensive expansion yields,

$$
E\left[\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right] = \sigma^2(n-2) \tag{28}
$$

Meaning that our original estimator would be unbiased had we used this $n-2$ factor. Therefore the following is an unbiased estimator of σ^2,

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{29}$$

(c) Derive the conditional variance of the OLS estimator of β_1 in (a) and construct an estimator of this conditional variance. Is the estimator unbiased for the conditional variance? [Note: conditional means "conditional on x".]

Here we calculate the unconditional variance,

$$\begin{aligned} Var(\hat{\beta}_1|x) &= Var\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) \\ &= Var\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) \\ &= Var\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1(x_i - \mu) + \epsilon_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) \end{aligned} \tag{30}$$

Where the only random variable is associated with the errors, so,

$$Var(\hat{\beta}_1|x) = \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}Var(\epsilon_i) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{31}$$

(d) Derive the unconditional variance of the OLS estimator of β_1 in (a). Is the unconditional variance equal to the conditional variance in (c)?

We know that,

$$Var(\hat{\beta}_1) = Var(E[\hat{\beta}_1|x]) + E[Var(\hat{\beta}_1|x)] \tag{32}$$

$$Var(\hat{\beta}_1) = Var(\beta_1) + E[Var(\hat{\beta}_1|x)] = E[Var(\hat{\beta}_1|x)] \tag{33}$$

$$Var(\hat{\beta}_1) = E\left[\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] = \sigma^2 E\left[\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] \tag{34}$$

(e) Construct an estimator for the expectation $E[y]$ using model M1.

We have that,

$$E[y] = E[\hat{\beta}_0 + \hat{\beta}_1(x - \mu) + \epsilon] = \hat{\beta}_0 + \hat{\beta}_1 E[x - \mu] + E[\epsilon] \tag{35}$$

$$E[y] = \hat{\beta}_0 + \hat{\beta}_1(\bar{x} - \mu) \tag{36}$$

So we call this an estimator of y,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1(\bar{x} - \mu) \tag{37}$$

(f) Construct an estimator for $E[y]$ not using the regression model.

The most straightforward estimator for y would be,

$$\hat{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{38}$$

(g) Are the two estimators in (e) and (f) equal? Why or why not?

The estimators are equal, as we can manipulate (38) into (37),

$$\hat{y} = \frac{1}{n}\sum_{i=1}^{n} \hat{\beta}_0 + \hat{\beta}_1(x_i - \mu) + \epsilon_i \tag{39}$$

$$\hat{y} = \frac{1}{n}\left(n\hat{\beta}_0 + \hat{\beta}_1(\sum_{i=1}^{n} x_i - n\mu) + \sum_{i=1}^{n} \epsilon_i\right) \tag{40}$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1(\bar{x} - \mu) \tag{41}$$

Where the error term is eliminated because we know that the sum of the errors must be zero.

2. **Consider a regression model $y = \beta_0 + \beta_1(x - \mu) + \epsilon$ where x is a non-random regressor and μ is a real number whose value may be known or unknown. Discuss whether the ordinary least-squares estimator of the slope β_1 is always unbiased and whether it always has the smallest variance compared to any estimator of β_1, irrespective of what the value of β_0 is. State assumptions in your discussion. Be careful about the word "any".**

First we outline the following three conditions:

(1) $Var(\epsilon_i) = \sigma^2 \quad \forall i$
(2) $Cov(\epsilon_i, \epsilon_j) = 0 \quad \forall i, j$
(3) $E[\epsilon_i] = 0 \quad \forall i$

Now, when all three conditions are met, it is true that β_1 is always unbiased. If, however, we have that the variance is not constant across the errors (violation of (1)), then β_1 would be biased. β_1 would also be biased if either of the other two conditions were violated.

With all three conditions satisfied, the Gauss–Markov theorem applies, meaning that β_1 has the smallest variance with respect to the class of unbiased estimators. However, there are other biased estimators which could possess a smaller variance. We could imagine a degenerate estimator (perhaps taking the form of something like a step function) which could reduce the variance. Additionally, the James-Stein estimator purports to be a biased estimator with lower variance.

3. **Select 25 rows from Table B.3. and complete the following:**

   (a) **Fit a simple linear regression model relating gasoline mileage y (miles per gallon) to engine displacement x_l (cubic inches).**

   Observe the attached code. The result is:

$$\hat{y} = 33.25 - 0.047x_1 \tag{42}$$

(b) Construct the analysis-of-variance table and test for significance of regression.

Observe the attached code, where the resulting table is displayed.

(c) What percent of the total variability in gasoline mileage is accounted for by the linear relationship with engine displacement?

Observe the attached code. This is given by the resulting R^2 value of 74.7%.

(d) Find a 95% CI on the mean gasoline mileage if the engine displacement is 275 $in.^3$.

Observe the attached code. The resulting 95% confidence interval is $(19.680, 21.146)$.

(e) Suppose that we wish to predict the gasoline mileage obtained from a car with a $in.^3$ engine. Give a point estimate of mileage. Find a 95% prediction interval on the mileage.

Observe the attached code. The resulting 95% prediction interval is $(16.677, 24.148)$.

(f) Compare the two intervals obtained in parts (d) and (e). Explain the difference between them. Which is wider, and why?

It is evident that the prediction interval obtained in part (e) is wider than the confidence interval obtained in part (d). This is because the prediction interval estimates an interval on a future observation, which must take into account the error from the model, as well as the observation itself. The confidence interval is only concerned with the error from the model.

Data Loading and Selection

In [7]:
```
# Load Data

import pandas as pd
df = pd.read_excel(r'--MASKED--\linear_regression_5e_data_sets\linear_regression_5e_dat
```

```
WARNING *** file size (10909) not 512 + multiple of sector size (512)
WARNING *** OLE2 inconsistency: SSCS size is 0 but SSAT size is non-zero
```

In [15]:
```
# Selected and View Data

use = df[['y','x1']]
sample = use.sample(25)
x = list(sample['x1'])
y = list(sample['y'])
print('x-values: ', x)
print('y-values: ', y)
```

Part (a)

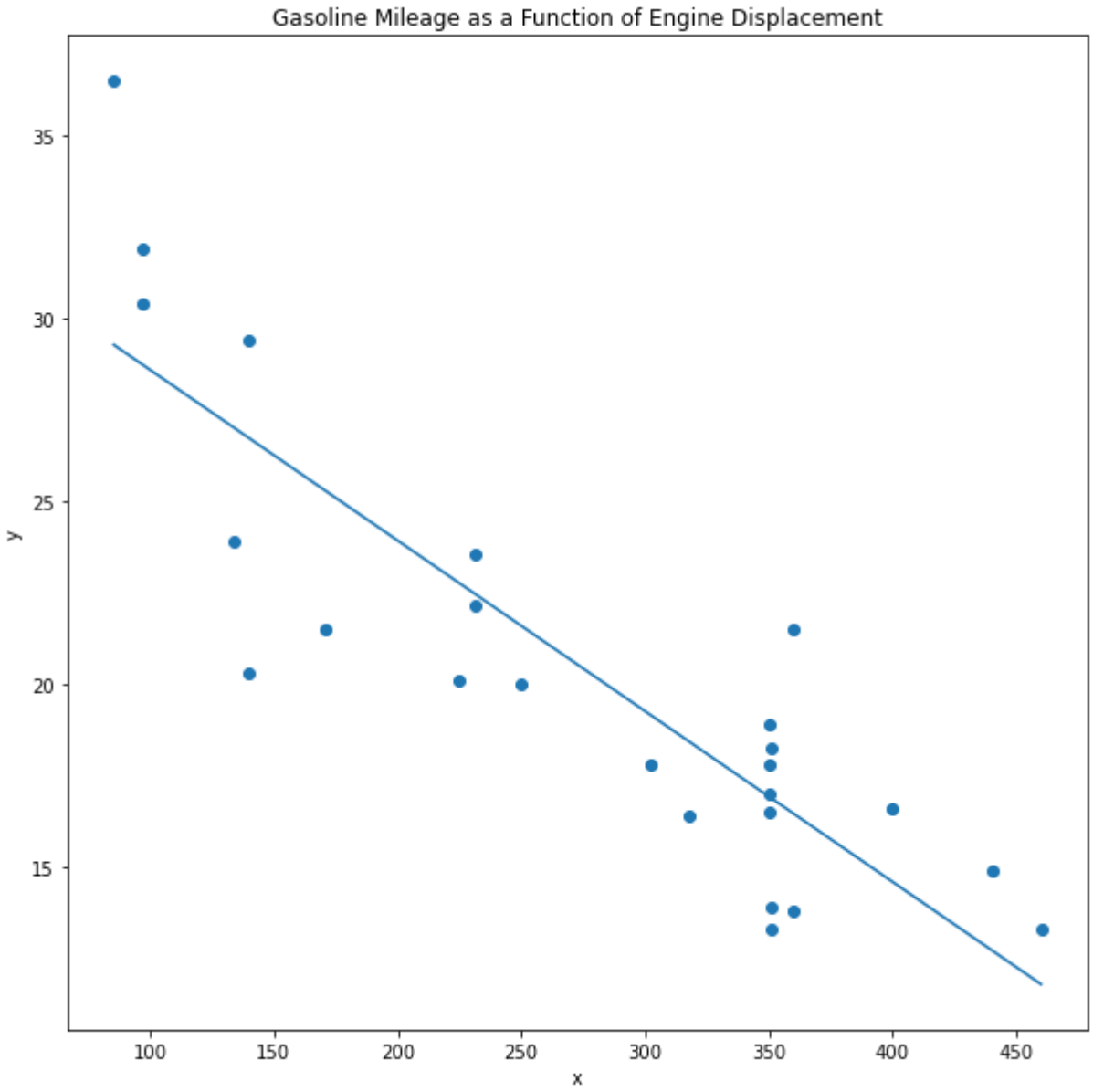
In [20]:
```
## Means of x and y Data

barx = sum(x)/len(x)
bary = sum(y)/len(y)
```

In [45]:
```
## Calculation of Model Parameters

Sxy = 0
Sxx = 0
for i in range(len(x)):
    Sxy += (x[i]-barx)*(y[i]-bary)
    Sxx += (x[i]-barx)**2
b1 = Sxy/Sxx
b0 = bary - b1*barx
print('beta_0 value: ', b0)
print('beta_1 value: ', b1)
```

```
beta_0 value:  33.24580559847236
beta_1 value:  -0.04666509130971887
```

In [46]:
```
## Viewing the Model

%matplotlib inline
import matplotlib.pyplot as plt
plt.figure(figsize=(10,10))
plt.title('Gasoline Mileage as a Function of Engine Displacement')
plt.xlabel('x')
plt.ylabel('y')
plt.scatter(x,y)
plt.plot([min(x),max(x)],[b1*min(x)+b0, b1*max(x)+b0])
```

Out[46]: [<matplotlib.lines.Line2D at 0x1b420308880>]



Part (b)

In [48]:
```
## Analysis of variance

types = ['Regression', 'Residual', 'Total']
SS = [b1*Sxy, tot, b1*Sxy + tot]
degf = [1, len(y)-2, len(y)-1]
MS = [b1*Sxy, MS_res, '-']
F0 = [b1*Sxy/MS_res, '-', '-']

pd.DataFrame({'Type':types, 'Sum Squares': SS, 'DoF': degf, 'Mean Square': MS, 'F_0':F0
```

Out[48]:

	Type	Sum Squares	DoF	Mean Square	F_0
0	Regression	669.837655	1	669.838	213.616
1	Residual	226.151545	23	3.13571	-

	Type	Sum Squares	DoF	Mean Square	F_0
2	Total	895.989200	24	-	-

Part(c)

In [51]:
```
## R^2

r2 = 1 - (SS[1]/SS[2])
print('r^2 value:  ', r2)
```

```
r^2 value:   0.7475956798276959
```

Part(d)

In [52]:
```
## Confidence Interval

tot = 0
for i in range(len(y)):
    tot += (y[i] - (b0+b1*x[i]))**2
MS_res = (tot/(len(y)-2))**0.5
radical = (MS_res*( (1/len(y)) + (((275-barx)**2)/Sxx )  ))**0.5

pred_val = b0 + 275*b1
err = 2.068658*radical
print('95% CI for Predicted Value:    (',pred_val-err,', ',pred_val+err,')')
```

```
95% CI for Predicted Value:    ( 19.680255583284502 ,  21.14555539331484 )
```

Part(e)

In [53]:
```
## Prediction Interval

tot = 0
for i in range(len(y)):
    tot += (y[i] - (b0+b1*x[i]))**2
MS_res = (tot/(len(y)-2))**0.5
radical = (MS_res*(1+ (1/len(y)) + (((275-barx)**2)/Sxx )  ))**0.5

pred_val = b0 + 275*b1
err = 2.068658*radical
print('95% PI for Predicted Value:    (',pred_val-err,', ',pred_val+err,')')
```

```
95% PI for Predicted Value:    ( 16.677190996026695 ,  24.148619980572647 )
```