BIS 623 Homework 1

Due on 09/12/2019 before the lecture

Joanna Chen

9/6/2019

Problem 1

Show that for SLM, the least square estimates are unbiased; and also $Var(\hat{\beta}_0) = \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i-\bar{x})^2})$ and $Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum(x_i-\bar{x})^2}$.

We can state the Simple Linear Model as follows:

$$Y_i = \beta_0 + \beta_1 X_i$$

Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the estimators. To show the least square estimates are unbiased, we need to show that the expectation value of the estimators is equal to the true value, i.e. $E\{\hat{\beta}_0\} = \beta_0$ and $E\{\hat{\beta}_1\} = \beta_1$.

$$\begin{aligned}
E\left\{\hat{\beta}_1\right\} &= E\left\{\frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}\right\} \\
&= E\left\{\frac{\sum_{i=1}^n \left(X_i - \overline{X}\right) Y_i}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2} - \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)\overline{Y}}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}\right\} \\
&= E\left\{\frac{\sum_{i=1}^n \left(X_i - \overline{X}\right) Y_i}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}\right\} \quad \text{since} \quad \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)\overline{Y}}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2} = \frac{0 \cdot \overline{Y}}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2} = 0 \\
&= \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2} \cdot E\left\{Y_i\right\} \\
&= \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2} \cdot (\beta_0 + \beta_1 X_i) \\
&= \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2} \cdot \beta_0 + \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2} \cdot \beta_1 X_i \\
&= 0 \cdot \beta_0 + \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)\left(X_i - \overline{X}\right)}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2} \cdot \beta_1 \quad \text{since} \quad \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2} = \frac{0}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2} = 0 = \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)\overline{X}}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2} \\
&= 0 + 1 \cdot \beta_1 \\
&= \beta_1
\end{aligned}$$

$$
\begin{aligned}
E\left\{\hat{\beta}_0\right\} &= E\left\{\overline{Y} - \hat{\beta}_1\ \overline{X}\right\} \\
&= E\left\{\overline{Y}\right\} - \overline{X}E\left\{\hat{\beta}_1\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n} E\left\{Y_i\right\} - \overline{X}\beta_1 \\
&= \frac{1}{n}\sum_{i=1}^{n} E\left\{\beta_0 + \beta_1 X_i + \epsilon_i\right\} - \overline{X}\beta_1 \\
&= \frac{1}{n}\sum_{i=1}^{n} E\left\{\beta_0\right\} + \frac{1}{n}\sum_{i=1}^{n} E\left\{\beta_1 X_i\right\} + \frac{1}{n}\sum_{i=1}^{n} E\left\{\epsilon_i\right\} - \overline{X}\beta_1 \\
&= \frac{1}{n} n\beta_0 + \frac{1}{n}\beta_1 \sum_{i=1}^{n} X_i - \overline{X}\beta_1 \\
&= \beta_0 + \overline{X}\beta_1 - \overline{X}\beta_1 \\
&= \beta_0
\end{aligned}
$$

$$
\begin{aligned}
Var\left(\hat{\beta}_0\right) &= Var\left\{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right\} \\
&= Var\left\{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)Y_i}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} - \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\overline{Y}}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right\} \\
&= Var\left\{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)Y_i}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right\} \quad \text{since} \quad \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\overline{Y}}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} = \frac{0 \cdot \overline{Y}}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} = 0 \\
&= \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^4} Var\left\{Y_i\right\} \\
&= \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^4}\sigma^2 \\
&= \frac{\sigma^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}
\end{aligned}
$$

$$
\begin{aligned}
Var\left(\hat{\beta}_0\right) &= Var\left(\overline{Y} - \hat{\beta}_1\overline{X}\right) \\
&= Var(\frac{\sum_{i=1}^{n} Y_i}{n}) + \overline{X}^2 Var\left(\hat{\beta}_1\right) - 2\overline{X}Cov\left(\frac{\sum_{i=1}^{n} Y_i}{n}, \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right) \\
&= \frac{1}{n^2} Var(\sum_{i=1}^{n} Y_i) + \overline{X}^2 Var\left(\hat{\beta}_1\right) - \frac{2\overline{X}}{n\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} \operatorname{Cov}\left(\sum_{i=1}^{n} Y_i, \sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)\right) \\
&= \frac{1}{n^2} Var(\sum_{i=1}^{n}(\beta_0 + \beta_1 X_i + \epsilon_i)) + \overline{X}^2 Var\left(\hat{\beta}_1\right) - \frac{2\overline{X}}{n\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} \sum_{i=1}^{n}\left(X_i - \overline{X}\right) \operatorname{Cov}\left(Y_i, Y_i - \overline{Y}\right) \\
&= \frac{1}{n^2} \sum_{i=1}^{n} Var(\epsilon_i) + \overline{X}^2 Var\left(\hat{\beta}_1\right) - \frac{2\overline{X}}{n\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} \sum_{i=1}^{n}\left(X_i - \overline{X}\right) \operatorname{Cov}\left(Y_i, Y_i\right) \\
&= \frac{n\sigma^2}{n^2} + \overline{X}^2 \operatorname{Var}\left(\hat{\beta}_1\right) - \frac{2\overline{X}}{n\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} \sum_{i=1}^{n}\left(X_i - \overline{X}\right)\sigma^2 \\
&= \frac{\sigma^2}{n} + \overline{X}^2 Var\left(\hat{\beta}_1\right) \quad \text{since} \quad \sum_{i=1}^{n}\left(X_i - \overline{X}\right) = 0 \\
&= \frac{\sigma^2}{n} + \overline{X}^2 \frac{\sigma^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} \\
&= \sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right)
\end{aligned}
$$

Problem 2

Show that for ANOVA, $\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum(y_i - \hat{y}_i)^2$.

$$
\begin{aligned}
\sum(y_i - \bar{y})^2 &= \sum(\hat{y}_i - \bar{y} + y_i - \hat{y}_i)^2 \\
&= \sum(\hat{y}_i - \bar{y})^2 + \sum(y_i - \hat{y}_i)^2 + 2\sum(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) \\
&= \sum(\hat{y}_i - \bar{y})^2 + \sum(y_i - \hat{y}_i)^2
\end{aligned}
$$

$$
\begin{aligned}
\text{This is because} \sum(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) &= \sum \hat{y}_i(y_i - \hat{y}_i) - \sum \bar{y}(y_i - \hat{y}_i) \\
&= \sum \hat{y}_i e_i - \sum \bar{y} e_i \\
&= \sum \hat{y}_i e_i - \bar{y}\sum e_i \\
&= 0 \quad \text{by Book 1.20 and 1.17}
\end{aligned}
$$

Problem 3

Complete the ANOVA table for SLR

Source	df	MS	F	Pvalue
Model				
Error		200		
Total	11	450		

- Is there a significant linear association between Y and X at $\alpha = 0.05$ level? Explain.
- What is the R^2 value for the regression? Interpret this value.
  Since the degree of freedom(dof) for the total is 11 and for the SLR model is always 1, then $11 = n - 1$ gives $n = 12$. Then dof(error) $= n - 2 = 12 - 2 = 10$.
  Note that mean squares are not additive. By definition, we know that

$$MSTO = SSTO/11, MSE = SSE/10, MSR = SSR/1.$$

  It follows that

$$SSTO = MSTO \cdot 11 = 450 \cdot 11 = 4950, SSE = MSE \cdot 10 = 200 \cdot 10 = 2000, SSR = MSR \cdot 1 = MSR.$$

  Therefore,

$$SSR = SSTO - SSE = 4950 - 2000 = 2950 = MSR.$$

  Hence,

$$F = MSR/MSE = 2950/200 = 14.75.$$

  We can use R to compute p-value as following:

```
pf(14.75, df1=1, df2=10,lower.tail=F)
```

```
## [1] 0.003262033
```

Source	df	MS	F	Pvalue
Model	1	2950	14.75	0.03
Error	10	200		
Total	11	450		

(1) H_0: There is no linear association between Y and X.
Since $0.003262033 < \alpha = 0.05$, we reject the null hypothesis and conclude that there is a significant linear association between Y and X at $\alpha = 0.05$ level.

(2)

$$R^2 = SSR/SSTO = 2950/4950 \approx 0.6$$

This means 60% of variance in Y can be explained by the model regressed on X.

Problem 4

In a regression analysis of on-the-job head injuries of warehouse laboreres caused by failling objects, we have a response as a measure of severity of injury (continous), and we will consider weight of the object, distance it fell and the type of protection worn at the time of accident (hard hat; bump cap or none).

- Develop a MLR for the problem and write down the response function ($E(y)$) on the predictors, and interpret the coefficients.
- For each of the following questions, specify H_0 and H_a:
  - With weight and distance fixed, does wearing a bump cap reduce the expected severity of injury as compared with wearing no protection?
  - With weight and distance fixed, is the expected severity of injury the same when wearing a hard hat as when wearing a bump hat?

(1)

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \quad \text{where}$$

β_0 is the intercept represents the mean response $E(Y)$ when the weight of the object $= 0$, the distance it fell $= 0$, and worn none protection at the time of accident.

β_1 indicates the change in the mean response $E(Y)$ per unit increase in the weight when other predictor held constant.

β_2 indicates the change in the mean response $E(Y)$ per unit increase in the distance when other predictor held constant.

β_3 indicates how much higher (lower) the response function for wearing hard is than the one for wearing none or bump, for any given level of weight and distance.

β_4 indicates how much higher (lower) the response function for wearing bump is than the one for wearing none or hard, for any given level of weight and distance.

x_1 is the weight of the object.

x_2 is the distance it fell.

$$x_3 = \begin{cases} 1 & \text{if wore hard} \\ 0 & \text{otherwise (if worn none/bump)} \end{cases}$$
$$x_4 = \begin{cases} 1 & \text{if wore bump} \\ 0 & \text{otherwise (if worn none/hard)} \end{cases}$$

For the case the wore neither protection, $E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$;

For the case the wore hard, $E(Y) = \beta_0 + \beta_3 + \beta_1 x_1 + \beta_2 x_2$;

For the case the wore bump, $E(Y) = \beta_0 + \beta_4 + \beta_1 x_1 + \beta_2 x_2$.

(2) $H_0 : \beta_4 \geq 0, H_a : \beta_4 < 0$
(3) $H_0 : \beta_3 = \beta_4, H_a : \beta_3 \neq \beta_4$