Homework 2

Juwon Lee, Economics and Statistics, UCLA

2023-01-13

tinytex::install_tinytex()

1.

(a) Show that the sum of residuals is always zero, i.e. $\sum_{i=1}^{n} \hat{e}_i = 0$.

Claim $\sum_{i=1}^{n} \hat{e}_i = 0$, having $E(\hat{\beta}_0) = \beta_0,\ E(\hat{\beta}_1) = \beta_1,\ E(e_i) = 0$.

$\sum_{i=1}^{n} \hat{e}_i = \sum_{i=1}^{n}(y_i - \hat{y}_i) = \sum_{i=1}^{n}[(\beta_0 + \beta_1 x_i + e_i) - (\hat{\beta}_0 + \hat{\beta}_1 x_i)] = \sum_{i=1}^{n}(\beta_0 - \hat{\beta}_0) + \sum_{i=1}^{n}(\beta_1 - \hat{\beta}_1)x_i + \sum_{i=1}^{n} e_i$.

Because $E(\hat{\beta}_0) = \frac{1}{n}\sum_{i=1}^{n} \hat{\beta}_0 = \beta_0$, $\sum_{i=1}^{n} \hat{\beta}_0 = n\beta_0 = \sum_{i=1}^{n} \beta_0$, so that $\sum_{i=1}^{n}(\beta_0 - \hat{\beta}_0) = 0$.

And $E(\hat{\beta}_1) = \frac{1}{n}\sum_{i=1}^{n} \hat{\beta}_1 = \beta_1$, $\sum_{i=1}^{n}(\hat{\beta}_1 - \beta_1) = 0$.

Thus, $\sum_{i=1}^{n} \hat{e}_i = \sum_{i=1}^{n} e_i$.

Because $E(e_i) = \frac{1}{n}\sum_{i=1}^{n} e_i = 0$, so that $\sum_{i=1}^{n} e_i = 0$. QED

(b) Show that $\hat{\beta}_0$ and $\hat{\beta}_1$ are the least square estimates, i.e. $\hat{\beta}_0$ and $\hat{\beta}_1$ minimizes $\sum \hat{e}^2$.

$\sum_{i=1}^{n} \hat{e}_i{}^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2 = \sum_{i=1}^{n}\{(\beta_0 + \beta_1 x_i) - (b_0 + b_1 x_i)\}^2$,

claim that $\hat{\beta}_0$ and $\hat{\beta}_1$ minimize $\sum_{i=1}^{n} \hat{e}_i{}^2$.

$\rightarrow \min_{b_0,\ b_1} \sum_{i=1}^{n}\{(\beta_0 + \beta_1 x_i) - (b_0 + b_1 x_i)\}^2 = \min_{b_0,\ b_1} \sum_{i=1}^{n}\{(\beta_0 - b_0) + (\beta_1 - b_1)x_i\}^2$

$\rightarrow b_0 = \hat{\beta}_0,\ b_1 = \hat{\beta}_1$. QED

(c) Show that S^2 is an unbiased estimator of σ^2.

Claim $E(S^2) = \sigma^2$ such that $S^2 = \frac{1}{n-2}\sum_{i=1}^{n} \hat{e}_i{}^2$.

$E(S^2) = E[\frac{1}{n-2}\sum_{i=1}^{n} \hat{e}_i{}^2] = E[\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2] = \frac{1}{n-2}\sum_{i=1}^{n} E[(y_i - \hat{y}_i)^2]$

$= \frac{1}{n-2}\sum_{i=1}^{n}[E(y_i^2) + E(\hat{y}_i{}^2) - 2E(y_i \hat{y}_i)]$

$= \frac{1}{n-2}\sum_{i=1}^{n}\{[E(y_i)^2 - (E(y_i))^2] + (E(y_i))^2 + [E(\hat{y}_i)^2 - (E(\hat{y}_i))^2 + (E(\hat{y}_i))^2] - 2E(y_i \hat{y}_i)\}$

$= \frac{1}{n-2}\sum_{i=1}^{n}[V(y_i) + (E(y_i))^2 + V(\hat{y}_i) + (E(\hat{y}_i))^2 - 2E(y_i \hat{y}_i)]$

$= \frac{1}{n-2}\sum_{i=1}^{n}[\sigma^2 + (E(y_i))^2 + 0 + (E(\hat{y}_i))^2 - 2E(y_i \hat{y}_i)]$

$= \frac{1}{n-2}\sum_{i=1}^{n}[\sigma^2 + 2(E(y_i))^2 - 2E(y_i \hat{y}_i)] = \frac{1}{n-2}\sum_{i=1}^{n}[\sigma^2 - 2(E(y_i \hat{y}_i) - (E(y_i))^2)]$

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

$= \frac{n\sigma^2}{n-2} - \frac{2\sigma^2}{n-2} = \frac{(n-2)\sigma^2}{n-2} = \sigma^2$. QED

2.

```r
indicators <- read.table('indicators.txt', header=T)
indicators_lm <- indicators[c(2,3)]

PriceChange <- as.vector(indicators_lm[1])
LoanPaymentsOverdue <- as.vector(indicators_lm[2])

data_lm_hw2_2 <- lm(PriceChange~LoanPaymentsOverdue, data=indicators)
data_lm_hw2_2
```

```
##
## Call:
## lm(formula = PriceChange ~ LoanPaymentsOverdue, data = indicators)
##
## Coefficients:
##         (Intercept)  LoanPaymentsOverdue
##               4.514               -2.249
```

```r
summary(data_lm_hw2_2)
```

```
##
## Call:
## lm(formula = PriceChange ~ LoanPaymentsOverdue, data = indicators)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6541 -3.3419 -0.6944  2.5288  6.9163
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.5145     3.3240   1.358   0.1933
## LoanPaymentsOverdue  -2.2485     0.9033  -2.489   0.0242 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.954 on 16 degrees of freedom
## Multiple R-squared:  0.2792, Adjusted R-squared:  0.2341
## F-statistic: 6.196 on 1 and 16 DF,  p-value: 0.02419
```

```r
Sxx_hw2_2 <- colSums((indicators_lm[2] - colMeans(indicators_lm[2]))^2)

len_hw2_2 <- length(indicators_lm$PriceChange)

Syy_hw2_2 <- colSums((indicators_lm[1] - colMeans(indicators_lm[1]))^2)
s_hw2_2 <- (sum(data_lm_hw2_2$residuals^(2)) / (len_hw2_2-2))^(1/2)
se_1_hw2_2 <- s_hw2_2 / (Sxx_hw2_2)^(1/2)

data_lm_hw2_2$coefficients[2] - qt(0.975, len_hw2_2-2) * se_1_hw2_2
```

```
## LoanPaymentsOverdue
##           -4.163454
```

```r
data_lm_hw2_2$coefficients[2] + qt(0.975, len_hw2_2-2) * se_1_hw2_2
```

```
## LoanPaymentsOverdue
```

```
##          -0.3335853
```

(a) Thus, the 95% confidence interval is (-4.163454, -0.333585). If $H_0: \ \beta_1 > 0, \ H_1: \ \beta_1 < 0$, then the p-value $= 0.02419 < 0.05$, so that we can reject the null. It means that we can't say that $\beta_1 > 0$.

(b) $E(Y|X=4) = 4.514 - 2.249 * 4$

```
data_lm_hw2_2$coefficients[1] + data_lm_hw2_2$coefficients[2] * 4
```

```
## (Intercept)
##   -4.479585
```

Thus, $E(Y|X=4) = -4.479585.

If we take the interval estimation,

```
E_hw2_2 <- (data_lm_hw2_2$coefficients[1] + data_lm_hw2_2$coefficients[2] * 4)
barx_hw2_2 <- colMeans(indicators_lm[2])

E_hw2_2 - qt(0.975, len_hw2_2-2) * s_hw2_2 * (1/len_hw2_2 + ((4-barx_hw2_2)^2 / Sxx_hw2_2))^(1/2)
```

```
## (Intercept)
##   -6.648849
```

```
E_hw2_2 + qt(0.975, len_hw2_2-2) * s_hw2_2 * (1/len_hw2_2 + ((4-barx_hw2_2)^2 / Sxx_hw2_2))^(1/2)
```

```
## (Intercept)
##   -2.310322
```

```
data_hw2_2 = data.frame(LoanPaymentsOverdue=4)
predict(data_lm_hw2_2, newdata=data_hw2_2, interval='confidence', level=0.95)
```

```
##         fit       lwr       upr
## 1 -4.479585 -6.648849 -2.310322
```

Thus, the 95% confidence interval for $E(Y|X=4)$ is (-6.648849, -2.310322). It means that 0% is not a feasible value for $E(Y|X=4)$ for $\alpha = 0.05$.

3.

```
invoices <- read.table('invoices.txt', header=T)

invoices_lm <- invoices[c(2,3)]

data_lm_hw2_3 <- lm(Time~Invoices, data=invoices)
data_lm_hw2_3
```

```
##
## Call:
## lm(formula = Time ~ Invoices, data = invoices)
##
## Coefficients:
## (Intercept)     Invoices
##     0.64171      0.01129
```

```
summary(data_lm_hw2_3)
```

```
##
## Call:
## lm(formula = Time ~ Invoices, data = invoices)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59516 -0.27851  0.03485  0.19346  0.53083
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.6417099  0.1222707   5.248 1.41e-05 ***
## Invoices    0.0112916  0.0008184  13.797 5.17e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3298 on 28 degrees of freedom
## Multiple R-squared:  0.8718, Adjusted R-squared:  0.8672
## F-statistic: 190.4 on 1 and 28 DF,  p-value: 5.175e-14
```

```
barx_hw2_3 <- colMeans(invoices_lm[1])

Sxx_hw2_3 <- colSums((invoices_lm[1] - colMeans(invoices_lm[1]))^2)

len_hw2_3 <- length(invoices$Invoices)

Syy_hw2_3 <- colSums((invoices_lm[2] - colMeans(invoices_lm[2]))^2)
s_hw2_3 <- (sum((data_lm_hw2_3$residuals)**2) / (len_hw2_3-2))^(1/2)

se_0_hw2_3 <- sd(data_lm_hw2_3$residuals) * ((1/len_hw2_3) + ((barx_hw2_3)^2 / Sxx_hw2_3))^(1/2)

data_lm_hw2_3$coefficients[1] - qt(0.975, len_hw2_3-2) * se_0_hw2_3
```

```
## (Intercept)
##   0.3956058
```

```
data_lm_hw2_3$coefficients[1] + qt(0.975, len_hw2_3-2) * se_0_hw2_3
```

```
## (Intercept)
##    0.887814
```

(a) Thus, the 95% confidence level is $(0.3956058, 0.887814)$.

(b) $H_0 : \beta_1 = 0.01$ vs. $H_1 : \beta_1 \neq 0.01$.

Then

```
se_1_hw2_3 <- s_hw2_3 / (Sxx_hw2_3)^(1/2)
se_1_hw2_3
```

```
##    Invoices
## 0.000818402
```

```
Statistic_hw2_3 <- (data_lm_hw2_3$coefficients[2] - 0.01) / se_1_hw2_3
Statistic_hw2_3
```

```
## Invoices
## 1.578251
```

```
qt(0.975, len_hw2_3-2)
```

```
## [1] 2.048407
```

Thus, the statistic $1.578251 < 2.048407$, so we cannot reject the null. Then, we can't say that β_1 is not 0.01.

(c) Suppose that $x = 130$. Then $Y|(X = 130) = 0.64171 + 0.01129 * 130 = 2.109624$.

```
data_lm_hw2_3
```

```
##
## Call:
## lm(formula = Time ~ Invoices, data = invoices)
##
## Coefficients:
## (Intercept)     Invoices
##     0.64171      0.01129
```

```
invoices130_hw2_3 <- data_lm_hw2_3$coefficients[1] + data_lm_hw2_3$coefficients[2] * 130
se130_hw2_3 <- s_hw2_3 * (1 + 1/len_hw2_3 + (130-barx_hw2_3)^2/Sxx_hw2_3)^(1/2)

invoices130_hw2_3 - qt(0.975, len_hw2_3-2) * se130_hw2_3
```

```
## (Intercept)
##    1.422947
```

```
invoices130_hw2_3 + qt(0.975, len_hw2_3-2) * se130_hw2_3
```

```
## (Intercept)
##      2.7963
```

```
data_hw2_3 <- data.frame(Invoices=130)
predict(data_lm_hw2_3, newdata=data_hw2_3, interval='prediction', level=0.95)
```

```
##        fit      lwr    upr
## 1 2.109624 1.422947 2.7963
```

so the 95% prediction interval is (1.422947, 2.7963).

4.

(a) Claim $(y_i - \hat{y_i}) = (y_i - \bar{y}) - \hat{\beta}_1(x_i - \bar{x})$.

$\rightarrow -(\hat{\beta}_0 + \hat{\beta}_1 x_i) = -\bar{y} - \hat{\beta}_1 x_i + \hat{\beta}_1 \bar{x}$.

$\rightarrow \ -\hat{\beta}_0 = -\bar{y} + \hat{\beta}_1 \bar{x}$.

$\rightarrow \ \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$.

$\leftrightarrow \ E(Y) = \hat{\beta}_0 + \hat{\beta}_1 E(X) = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$. QED

(b) Claim $(\hat{y_i} - \bar{y}) = \hat{\beta}_1(x_i - \bar{x})$.

$\rightarrow \ (\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y}) = \hat{\beta}_1 x_i - \hat{\beta}_1 \bar{x}$.

$\rightarrow \ \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$, which is same with (a). QED

(c) Claim $\sum_{i=1}^n (y_i - \hat{y_i})(\hat{y_i} - \bar{y}) = 0$, using $\hat{\beta}_1 = \frac{Sxy}{Sxx} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$.

$\rightarrow \ \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))((\hat{\beta}_0 + \hat{\beta}_1 x_i) - \bar{y}) = 0$

$\rightarrow \ \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))((\hat{\beta}_0 + \hat{\beta}_1 x_i) - (\hat{\beta}_0 + \hat{\beta}_1 \bar{x})) = 0$

$\rightarrow \ \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)) * \hat{\beta}_1(x_i - \bar{x}) = 0$

thus, if $\sum_{i=1}^n (x_i - \bar{x}) = 0$, then it is clear.

$\rightarrow \ \sum_{i=1}^n x_i - n\bar{x} = 0, \ \bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$. QED

5.

X is the distance, Y is airfares.
And len(Y) = 17, E(Y) = $\frac{1}{n}\sum_{i=1}^n y_i = 228.35$, sd($Y$)= $\frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2 = 129.74$,

E(X) =$\frac{1}{n}\sum_{i=1}^n x_i$= 816.53, sd($X$) =$\frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2 = 588.79$.

(a) First of all, because $\frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2 = 129.74$,

$\rightarrow \ Syy = 129.74 * (n-1) = 129.74 * 16$.

```
Syy_hw2_5 <- 129.74 * 16
Syy_hw2_5
```

```
## [1] 2075.84
```

$\therefore \ s = \sqrt{\frac{Syy}{n-2}} = \sqrt{\frac{129.74*16}{15}}$

```
s_hw2_5 <- (Syy_hw2_5 / 15)^(1/2)
s_hw2_5
```

```
## [1] 11.7639
```

And, because $\frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2 = 588.79$,

```
Sxx_hw2_5 <- 588.79 * 16
Sxx_hw2_5
```

```
## [1] 9420.64
```

Therefore, because we want to find $se(\hat{\beta}_0) = s\sqrt{(\frac{1}{n} + \frac{\bar{x}^2}{Sxx})}$ and $se(\hat{\beta}_1) = \frac{s}{\sqrt{Sxx}}$.

```
se_0_hw2_5 <- s_hw2_5 * (1/17 + (816.53)^2/Sxx_hw2_5)^(1/2)
se_1_hw2_5 <- s_hw2_5 / (Sxx_hw2_5)^(1/2)

se_0_hw2_5
```

```
## [1] 99.00649
```

```
se_1_hw2_5
```

```
## [1] 0.1212024
```

Thus, (1) = 99.00649, (4) = 0.1212024.

```
48.97177 / se_0_hw2_5
```

```
## [1] 0.4946319
```

```
0.219687 / se_1_hw2_5
```

```
## [1] 1.812564
```

Thus, (2) = 0.4946319, (5) = 1.812564.

```
2 * (1 - pt(0.4946319, 15))
```

```
## [1] 0.6280259
```

```
2 * (1 - pt(1.812564, 15))
```

```
## [1] 0.0899601
```

Thus, (3) = 0.6280259, (6) = 0.0899601

And because Adjusted R-squared $= 1 - \frac{n-1}{n-k-1}(1-R^2) = 1 - \frac{16}{15}(1-R^2) = 0.9936$, $(\because\ k=1)$

so that $R^2 = 1 - (1-0.9936) * \frac{15}{16}$.

```
1 - (1-0.9936) * 15 / 16
```

```
## [1] 0.994
```

Thus, (7) = 0.994.

Finally, we can say that

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(intercept)	48.971770	99.00649	0.4946319	0.6280259
Distance	0.219687	0.1212024	1.812564	0.0899601
Multiple R-squared	0.994			

And, because we have known that the F-statistic is 2469 and p-value is $2.2e-16$, so

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Distance	1	SSA	SSA	2469	$2.2e-16$
Residuals	15	SSE	SSE/15		

Also, $\frac{1}{n}\sum_{i=1}^{n} y_i = 228.35$ and $\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2 = 129.74$,

$\rightarrow \ \sum_{i=1}^{n} y_i^2 - 228.35 * 2\sum_{i=1}^{n} y_i + 17 * (228.35)^2 = 129.34 * 16$,

$\rightarrow \ \sum_{i=1}^{n} y_i^2 = 228.35 * 2\sum_{i=1}^{n} y_i + 129.34 * 16 - 17 * (228.35)^2 = 888519.1$

then, $CT = \frac{T^2}{N} = \frac{(17*228.35)^2}{17} = 886443.3$

Thus, $SST = \sum_{i=1}^{n} y_i^2 - CT = 2075.84$.

```
228.35 * 2 * 228.35 * 17 + 129.74 * 16 - 17 * (228.35)^2
```

```
## [1] 888519.1
```

```
(17*228.35)^2 / 17
```

```
## [1] 886443.3
```

```
228.35 * 2 * 228.35 * 17 + 129.74 * 16 - 17 * (228.35)^2 - (17*228.35)^2 / 17
```

```
## [1] 2075.84
```

Now, we have $\begin{cases} \frac{SSA}{SSE/15} = 2469 \\ 2075.84 = SST = SSA + SSE \end{cases}$

$\rightarrow \ SSA = 2063.305, \ SSE = 12.53527$.

Finally, we can conclude that

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Distance	1	2063.305	2063.305	2469	$2.2e-16$
Residuals	15	12.53527	0.8356844		

```
15 * 2075.84 / (15 + 2469)
```

```
## [1] 12.53527
```

```
2075.84 - (15 * 2075.84 / (15 + 2469))
```

```
## [1] 2063.305
```

```
(15 * 2075.84 / (15 + 2469)) / 15
```

```
## [1] 0.8356844
```

```
(2075.84 - (15 * 2075.84 / (15 + 2469))) / ((15 * 2075.84 / (15 + 2469)) / 15)
```

```
## [1] 2469
```

(b) $y = 48.97177 + 0.219687 * x$.

(c) $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$.

$T = \frac{0.219687}{0.1212024} = 1.812563$,

$P(|t| < 1.812563) = 0.08996026 > 0.025$, so we can't reject the null.

Thus, we can't say that β_1 is not zero.

```
0.219687 / 0.1212024
```

```
## [1] 1.812563
```

```
2 * (1 - pt(1.812563, 15))
```

```
## [1] 0.08996026
```

For β_0, $H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$.

$T = \frac{48.971770}{99.00649} = 0.4946319$,

$P(|t| < 0.4946319) = 0.6280259 > 0.025$, so we can't reject the null.

Thus, we can't say that β_0 is not zero.

```
48.971770 / 99.00649
```

```
## [1] 0.4946319
```

```
2 * (1 - pt(0.4946319, 15))
```

```
## [1] 0.6280259
```

(d) $R^2 = 0.994$ explains that, 99.4% percentage of sum-of-square of y(airfares) is explained by x(distance).

(e) $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$.

```
qf(0.95, 1, 15)
```

```
## [1] 4.543077
```

Thus, $F = 2469 > F_{0.05}(1,\ 15) = 4.543077$, so we can reject the null.

```
qt(0.975, 15)^2
```

```
## [1] 4.543077
```

Then, it is not consistent to the hypothesis testing for the slope,
but $F_{0.05}(1,\ 15) = (t_{0.025}(15))^2$.