1. a. $\sum_{i=1}^{n} \hat{e}_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$ $\quad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

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

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

$= 0 - \hat{\beta}_1 \cdot 0$

$= 0$ ■

So the sum of residuals is 0.

b. $S(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$

$= \sum_{i=1}^{n} \left((y_i + \hat{y} - \hat{y}) - \beta_0 - \beta_1 (x_i + \hat{x} - \hat{x}) \right)^2$

$= \sum_{i=1}^{n} (y_i + \hat{y} - \hat{y} - \beta_0 - \beta_1 x_i + \beta_1 \hat{x} - \beta_1 \hat{x})^2$

$= \sum_{i=1}^{n} (\hat{y} - \beta_0 - \beta_1 \hat{x})^2 + \sum_{i=1}^{n} (\beta_1 x_i + \beta_1 x + \hat{y} + y_i)^2$

$= n(\hat{y} - \beta_0 - \beta_1 \hat{x})^2 + \sum_{i=1}^{n} (\beta_1 (x_i - x) + (\hat{y} - y_i))^2$

$= n(\hat{y} - \beta_0 - \beta_1 \hat{x})^2 + \left(\sum_{i=1}^{n} \beta_1^2 (x_i - x)^2 - 2\beta_1 (x_i - \hat{x})(y_i - \hat{y}) + (y_i - y)^2 \right)$

$= n(\hat{y} - \beta_0 - \beta_1 \hat{x})^2 + \left(\beta_1^2 \sum_{i=1}^{n} (x_i - x)^2 - 2\beta_1 \sum_{i=1}^{n} (x_i - \hat{x})(y_i - \hat{y}) + \sum_{i=1}^{n} (y_i - y)^2 \right)$

$= n(\hat{y} - \beta_0 - \beta_1 \hat{x})^2 + \left(\sum_{i=1}^{n} (x_i - \hat{x})^2 \right) \left(\beta_1^2 - 2\beta_1 \frac{\sum_{i=1}^{n} (x_i - \hat{x})(y_i - \hat{y})}{\sum_{i=1}^{n} (x_i - \hat{x})^2} + \frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{\sum_{i=1}^{n} (x_i - \hat{x})^2} \right.$

$= n(\hat{y} - \beta_0 - \beta_1 \hat{x})^2 + \left(\sum_{i=1}^{n} (x_i - \hat{x})^2 \right) \left(1 - \left(\frac{\sum_{i=1}^{n} (x_i - \hat{x}) \cdot \sum_{i=1}^{n} (y_i - \hat{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \hat{x})^2 \cdot \sum_{i=1}^{n} (y_i - \hat{y})^2}} \right)^2 + \left(\beta_1 - \frac{\sum_{i=1}^{n} (x_i - \hat{x}) \cdot \sum_{i=1}^{n} (y_i - \hat{y})}{\sum_{i=1}^{n} (x_i - \hat{x})^2} \right)^2 \right)$

$= n(\hat{y} - \beta_0 - \beta_1 \hat{x})^2 + \left(\sum_{i=1}^{n} (y_i - \hat{y})^2 \right) \left(1 - \left(\frac{\sum_{i=1}^{n} (x_i - \hat{x}) \cdot \sum_{i=1}^{n} (y_i - \hat{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \hat{x})^2 \cdot \sum_{i=1}^{n} (y_i - \hat{y})^2}} \right)^2 + \left(\beta_1 - \frac{\sum_{i=1}^{n} (x_i - \hat{x}) \cdot \sum_{i=1}^{n} (y_i - \hat{y})}{\sum_{i=1}^{n} (x_i - \hat{x})^2} \right)^2 \right)$

$\geq \left(\sum_{i=1}^{n} (y_i - y) \right)^2 \left(1 - \left(\frac{\sum_{i=1}^{n} (x_i - \hat{x}) \cdot \sum_{i=1}^{n} (y_i - \hat{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \hat{x})^2 \cdot \sum_{i=1}^{n} (y_i - \hat{y})^2}} \right)^2 \right)$

$n(\hat{y} - \beta_0 - \beta_1 \hat{x})^2 = 0$

$0 = \left(\beta_1 - \frac{\sum_{i=1}^{n} (x_i - \hat{x}) \cdot \sum_{i=1}^{n} (y_i - \hat{y})}{\sum_{i=1}^{n} (x_i - \hat{x})^2} \right)^2$

$\Rightarrow \beta_0 = \hat{y} - \beta_1 x$

$\beta_1 = \frac{\sum_{i=1}^{n} (x_i - \hat{x}) \cdot \sum_{i=1}^{n} (y_i - \hat{y})}{\sum_{i=1}^{n} (x_i - \hat{x})^2}$

So $\hat{\beta}_0$ and $\hat{\beta}_1$ are the least square estimates

c. Unbiased estimator : zero bias

Expected value of parameter = true value of the parameter

WTS: $E(S^2) = E\left(\frac{RSS}{n-2}\right) = \sigma^2$

$RSS = \sum (Y_i - \hat{Y}_i)^2$

$E(\sum X_i) = \sum E(X_i)$ $\quad E(cX) = cE(X)$

$Var(X) = E(X^2) - (E(X))^2$

$E(X^2) = \sigma^2 + \mu^2$

$E(\bar{X}^2) = \frac{\sigma^2}{n} + \mu^2$

WTS: $E(S^2) = E\left(\frac{\sum_{i=1}^{n}(X_i - X)^2}{n-1}\right) = \sigma^2$

$E(\sum(X_i - X)^2) = E(\sum(X_i^2 - 2X_i\bar{X} + \bar{X}^2))$

$= E(\sum X_i^2 - \sum 2X_i\bar{X} + \sum \bar{X}^2)$

$\bar{X} = \frac{\sum X_i}{n} \quad n\bar{X} = \sum X_i$

$= E(\sum X_i^2 - 2\bar{X}\sum X_i + n\bar{X}^2)$

$= E(\sum X_i^2 - \underbrace{2\bar{X}\cdot n\bar{X}}_{2n\bar{X}^2} + n\bar{X}^2)$

$= E(\sum X_i^2 - n\bar{X}^2)$

$= \sum E(X_i^2) - E(n\bar{X}^2)$

$= \sum E(X_i^2) - nE(\bar{X}^2)$

$= \sum(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)$

$= n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2$

$= (n-1)\sigma^2$

$E(S^2) = E\left(\frac{\sum_{i=1}^{n}(X_i - X)^2}{n-1}\right)$

$= \frac{1}{n-1}E(\sum(X_i - X)^2)$

$= \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2$

$\Rightarrow E(S^2) = \sigma^2$

So S^2 is an unbiased estimator of σ^2

Stats 1021 - Homework 2

Damien Ha

2023-01-27

Problem 2

A story by James R. Hagerty entitled *With Buyers Sidelined, Home Prices Slide* published in the Thursday October 25, 2007 edition of the *Wall Street Journal* contained data on so-called fundamental housing indicators in major real estate markets across the US. The author argues that… *prices are generally falling and overdue loan payments are piling up*. Thus, we shall consider data presented in the article on

Y **= Percentage change in average price from July 2006 to July 2007 (based on the S&P/Case-Shiller national housing index); and**

x **= Percentage of mortgage loans 30 days or more overdue in latest quarter (based on data from Equifax and Moody's).**

The data are available on the book web site in the file indicators.txt. Fit the following model to the data: $Y = \beta_0 + \beta_1 x + e$. Complete the following tasks:

(a) Find a 95% confidence interval for the slope of the regression model, b1. On the basis of this confidence interval decide whether there is evidence of a significant negative linear association.

(b) Use the fitted regression model to estimate $E(Y \mid X = 4)$. Find a 95% confidence interval for $E(Y \mid X = 4)$. Is 0% a feasible value for $E(Y \mid X = 4)$? Give a reason to support your answer.

Part A

```
indicators <- read.table("indicators.txt", header = T)
```

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

```
##
## 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
```

```
confint(model, "LoanPaymentsOverdue", level = 0.95)
```

```
##                         2.5 %     97.5 %
## LoanPaymentsOverdue -4.163454 -0.3335853
```

Since we are 95% confident that the true value of β_1 falls between -4.16 and -0.33, there is evidence of a significant negative linear trend.

Part B

```
predict(model, data.frame("LoanPaymentsOverdue" = 4), interval='confidence')
```

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

0% is not a reasonable estimate of $E[X \mid Y = 4]$ because our 95% confidence interval lies entirely below 0

Problem 3

The manager of the purchasing department of a large company would like to develop a regression model to predict the average amount of time it takes to process a given number of invoices. Over a 30-day period, data are collected on the number of invoices processed and the total time taken (in hours). The data are available on the book web site in the file invoices.txt. The following model was fit to the data: $Y = \beta_0 + \beta_1 x + e$ where Y is the processing time and x is the number of invoices. A plot of the data and the fitted model can be found in Figure 2.7. Utilizing the output from the fit of this model provided below, com- plete the following tasks.

(a) Find a 95% confidence interval for the start-up time, i.e., β_0.

(b) Suppose that a best practice benchmark for the average processing time for an additional invoice is 0.01 hours (or 0.6 minutes). Test the null hypothesis

$H_0 : \beta_1 = 0.01$ **against a two-sided alternative. Interpret your result.**

(c) Find a point estimate and a 95% prediction interval for the time taken to process 130 invoices.

Part A

```
invoices <- read.table("invoices.txt", header = T)
model2 <- lm(Time ~ Invoices, data = invoices)
summary(model2)
```

```
##
## 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
```

```
confint(model2, "(Intercept)", level = 0.95)
```

```
##                 2.5 %    97.5 %
## (Intercept) 0.3912496 0.8921701
```

We are 95% confident that the true value of β_0 falls between 0.39 and 0.89

Part B

```
2 * pt((0.0112916 - 0.01) / 0.0008184, 29, lower.tail = F)
```

```
## [1] 0.1253666
```

The p-value is 0.1253666 which is greater than the significance level of 0.05, so we fail to reject the null hypothesis that $\beta_1 = 0.01$ hours

Part C

```
predict(model2, data.frame("Invoices" = 130), interval="prediction")
```

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

The 95% prediction interval is (1.422947, 2.7963). It takes about 2.109624 hours to process 130 invoices

4.

6. In this problem we will show that SST=SSreg+RSS . To do this we will show that $\sum_{i=1}^{n}(y_i-\hat{y}_i)\,(\hat{y}_i-\overline{y})=0$.

(a) Show that $(y_i-\hat{y}_i)=(y_i-\overline{y})-\hat{\beta}_1(x_i-\overline{x})$.

(b) Show that $(\hat{y}_i-\overline{y})=\hat{\beta}_1(x_i-\overline{x})$.

(c) Utilizing the fact that $\hat{\beta}_1=\frac{SXY}{SXX}$, show that $\sum_{i=1}^{n}(y_i-\hat{y}_i)\,(\hat{y}_i-\overline{y})=0$.

a. $y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ $\qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$= y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i$

$= \boxed{(y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})}$ ∎

b. $\cancel{y_i} - \hat{y}_i = (\cancel{y_i} - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})$ (from part A)

$-y_i + \bar{y} \qquad -y_i + \bar{y}$

$(-1)\ \bar{y} - \hat{y}_i = -\hat{\beta}_1 (x_i - \bar{x})\ (-1)$

$\boxed{\hat{y}_i - \bar{y} = \hat{\beta}_1 (x_i - \bar{x})}$ ∎

c. $\hat{\beta}_1 = \frac{SXY}{SXX}$

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

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

$= \hat{\beta}_1 (SXY - 0 - \hat{\beta}_1 SXX)$

$= \hat{\beta}_1 (SXY - 0 - \frac{SXY}{\cancel{SXX}} \cancel{SXX})$

$= \hat{\beta}_1 (SXY - SXY)$

$= 0$

$\boxed{\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0}$ ∎

5. Suppose that the data from 17 flights contains the two variables, airfares and distance. Below are the R results to find the linear model of predicting airfares from the distance of the flight. It is given that the average of 17 airfares is 228.35 and their SD is 129.74. Also, the average of distances is 816.53 and their SD is 588.79.

(a) Complete the R results.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 48.971770    (1)      (2)     (3)
Distance     0.219687    (4)      (5)     (6)
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.41 on 15 degrees of freedom
Multiple R-squared:  (7)  ,     Adjusted R-squared:  0.9936
F-statistic:  2469 on 1 and 15 DF,  p-value: < 2.2e-16
```

```
Analysis of Variance Table

Response: Fare
          Df Sum Sq Mean Sq F value    Pr(>F)
Distance  (1)  (2)     (3)     (7)        (8)
Residuals (4)  (5)     (6)
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(b) Write a linear regression equation.

(c) Is the slope significant at 0.05 level? How about the intercept? Why or why not?

(d) Interpret R^2.

(e) State the null and alternative hypotheses for the ANOVA.

(f) Make a conclusion for the ANOVA. Is it consistent to the hypothesis testing for the slope?

a.) R results:

(1) $Se(\hat{\beta}_0) = S \cdot \sqrt{\frac{1}{n} + \frac{x^2}{Sxx}}$

$= 10.41 \cdot \sqrt{\frac{1}{17} + \frac{(816.53)^2 \leftarrow \text{avg}}{16(588.79)^2 \leftarrow \text{SD}}}$

$= 4.4058$

(2) $T = \frac{\hat{\beta}_0}{Se(\hat{\beta}_0)} = \frac{48.971770}{4.0458}$

$= 11.1184$

(3) $P(>|t|)$

$= 1 - p(T, df, df^2)$

$= 1 - p(11.1184, 1, 15)$

$= 0.001 \, (1.215 \times 10^{-18})$

≈ 0

(4) $Se(\hat{\beta}_1) = \frac{s}{\sqrt{Sxx}}$

$= \frac{10.41}{\sqrt{16(588.79)^2}}$

$= 0.0044$

(5) $T = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{0.219687}{0.0044}$

$= 49.7029$

(6) F-statistic w/ p-value: 2.2×10^{-16}

$\Rightarrow P(>|t|) \approx 0$

(7) Multiple $R^2 = \frac{SST - RSS}{SST}$

$\frac{((n-1) \cdot Var(Y)) - (df \cdot (res.st\ err)^2)}{(n-1)(SD\ airfares)^2}$

$= \frac{16 \cdot (129.74)^2 - 15 \cdot (10.41)^2}{16(129.74)^2}$

$= 0.9936$

ANOVA Table

(1) F-statistic:

2469 on 1 and 15 df

1

(2) Sum sq = SST − RSS

$= ((n-1) - var(Y)) - (df \cdot res\text{-}std\text{-}error^2)$

$= (16 \cdot 129.74^2) - (15 \cdot 10.41^2)$

$= 267693.9601$

(3) mean $= \frac{\text{sum sq}}{df}$

$= 267693.9601$

(4) $df = n-2 = 17-2$

$= 15$

(5) sum sq. = RSS

$= df \cdot res\text{-}std\text{-}error^2$

$= 15 \cdot 10.41^2 = 162552.15$

(6) mean sq. $= \frac{RSS}{residuals\ df}$

$= \frac{df \cdot res\text{-}std\text{-}err}{n-2}$

$= \frac{15 \cdot 10.41^2}{15}$

$= 108.3681$

b.) $\hat{y}$: intercept ~ estimate + (distance ~ estimate) · x

$\hat{y}$: 48.9718 + 0.4197x

c.) p-value for β_0 : 0

p-value for β_1 : 4.622×10^{-18}

$0 < 0.05$, $4.622 \times 10^{-18} < 0.05$

=> The slope and intercept are both significant

d.) Adjusted R^2: 0.9936

The model explains 99.36 % of the variability in airfares

e.) $H_0: y = \beta_0 + e$

$H_A: y = \beta_0 + \beta_1 x + e$

f.) F-statistic = 2469

= slope F test

Each test had about the same F statistics or p-values

=> consistent to

$H_A: y = \beta_0 + \beta_1 x + e$

$H_0: \beta_1 = 0$, $H_A: \beta_1 \neq 0$