Stat302 Assignment 2 Solution

Daisy Yu

6/18/2023

Q2.8 c (1 point)

Q2.12 (5 points)

a. (3 points)

$t = \frac{5.3}{2.8} = 1.89$ (1 point) p-value $= 0.06 > 0.05$ (1 point) We fail to reject H_0 (1 point) ### b. (2 points) t critical value = 1.99 (1 point) Confidence interval: (-0.272, 10.872) (1 point)

```
df <- 82-2
t_star <- qt(0.975,df)
lower <- 5.3-t_star*2.8;print(lower)
```

```
## [1] -0.2721776
```

```
upper <- 5.3+t_star*2.8;print(upper)
```

```
## [1] 10.87218
```

Q2.16 (7 points)

a. (4 points)

$H_0 = \beta = 0$ vs $H_1 = \beta \neq 0$ (1 point) $t = 7.653$ (1 point) p-value < 0.05 (1 point) We reject H_0, and the number of pages is a useful precitor for predicting the price of the book. (1 point)

```
library(Stat2Data)
data(TextPrices)
fit <- lm(Price~Pages,data=TextPrices)
summary(fit)
```

```
##
## Call:
## lm(formula = Price ~ Pages, data = TextPrices)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -65.475 -12.324  -0.584  15.304  72.991
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.42231   10.46374  -0.327    0.746
## Pages        0.14733    0.01925   7.653 2.45e-08 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.76 on 28 degrees of freedom
## Multiple R-squared:  0.6766, Adjusted R-squared:  0.665
## F-statistic: 58.57 on 1 and 28 DF,  p-value: 2.452e-08
```

b. (3 points)

Confidence interval: (0.108, 0.187) (1 point) Interpretation of confidence interval: we are 95% confident that the true value of slope is between 0.108 and 0.187. (1 point) Interpretation of slope: when number of pages increase by 1, the price of the book increase by 0.14733. (1 point)

```
confint(fit)
```

```
##                   2.5 %    97.5 %
## (Intercept) -24.8563229 18.011694
## Pages         0.1078959  0.186761
```

Q2.18 (8 points)

a. (4 points)

$H_0 = \beta = 0$ vs $H_1 = \beta \neq 0$ (1 point) t-statistics = 13.463 (1 point) p-value < 0.05 (1 point) We reject H_0. (1 point)

```
data("Sparrows")
model <- lm(Weight~WingLength,data=Sparrows)
summary(model)
```

```
##
## Call:
## lm(formula = Weight ~ WingLength, data = Sparrows)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5440 -0.9935  0.0809  1.0559  3.4168
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.36549    0.95731   1.426    0.156
## WingLength   0.46740    0.03472  13.463   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.4 on 114 degrees of freedom
## Multiple R-squared:  0.6139, Adjusted R-squared:  0.6105
## F-statistic: 181.3 on 1 and 114 DF,  p-value: < 2.2e-16
```

b. (2 points)

Confidence interval: (0.399, 0.536) (1 point) Interpretation of confidence interval: we are 95% confident that the true value of slope is between 0.399 and 0.536. (1 point)

```
confint(model)
```

```
##                  2.5 %    97.5 %
## (Intercept) -0.5309316 3.2619109
## WingLength   0.3986288 0.5361792
```

c. (2 points)

Confidence interval does not contain 0, so it supports our conclusion in part (a) that WingLength is significant predicting Weight.

Q2.22 (2 points)

$r^2 = \frac{SSModel}{SST} = \frac{38}{102} = 0.3725$ (1 point) The linear regression model explain 37.25% of variation within the response. (1 point)

Q2.26 (10 points)

a. (4 points)

$H_0 = \beta = 0$ vs $H_1 = \beta \neq 0$ (1 point) t-statistics = -4.029 (1 point) p-value < 0.05 (1 point) We reject H_0. (1 point)

```
data(LeafWidth)
reg <- lm(Width~Year,data=LeafWidth)
summary(reg)
```

```
##
## Call:
## lm(formula = Width ~ Year, data = LeafWidth)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1214 -1.1253 -0.3136  0.9320  5.4144
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.723091   8.574977   4.399 1.61e-05 ***
## Year        -0.017560   0.004358  -4.029 7.43e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.424 on 250 degrees of freedom
## Multiple R-squared:  0.06098,    Adjusted R-squared:  0.05723
## F-statistic: 16.24 on 1 and 250 DF,  p-value: 7.425e-05
```

b. (1 point)

$R^2 = 6.1\%$ ### c. (3 points) F = 16.236 (1 point) p-value < 0.05 (1 point) We reject H_0. (1 point)

```
anova(reg)
```

```
## Analysis of Variance Table
##
## Response: Width
##            Df Sum Sq Mean Sq F value    Pr(>F)
## Year        1  32.91  32.911  16.236 7.425e-05 ***
## Residuals 250 506.76   2.027
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

d. (2 point)

$\sqrt{F} = \sqrt{16.236} = 4.029 = t$

Q2.44 (10 points)

a. (1 point)

Confidence Interval: (51.73, 74.02) ### b. (1 point) Prediction Interval: (0.90, 124.85) ### c. (1 point) The midpoints of both intervals are the same. Both intervals have the general form of $\hat{y}\pm$ some margin of error. ### d. (2 points) The confidence interval for the mean price is much narrower than the prediction interval for the price of an individual textbook. We need a much wider interval to capture most of the textbook prices than we do to just capture the mean of those prices. ### e. (2 points) The narrowest possible interval is when the number of pages is $x^* = \bar{x} = 464.5$ so that the term involving term involvingterm involving term involvingterm involvingterm involving term involving $(x^* - \bar{x})^2$ contributes nothing to the standard error, $SE_{\hat{y}}$. ### f. (3 points) Prediction Interval: (143.36, 291.78) The 95% prediction interval for textbook price when the number of pages is 1500 goes from 143.36 to 291.78. However, 1500 pages is much larger than any of the textbooks in the sample (as seen in the note in the output) and much of the prediction interval covers prices that are much larger than any of the prices in the sample. We should avoid this sort of extrapolation and thus would have less than 95% confidence that the interval would capture the price for a particular 1500-page textbook.

Q3.2 (2 points)

a. (1 point)

$\hat{calories} = 109.3 + 1x11 - 3.7x1 = 116.6$ ### b. (1 point) Residual = 110 - 116.6 = -6.6

Q3.6 (2 points)

As the number of grams of fiber per serving goes up by 1, after accounting for the amount of sugar, the average number of calories goes down by 3.7.

Q3.8 (4 points)

a. (2 points)

This is true. Adding a new predictor to a multiple regression model can never decrease the percentage of variability explained by that model. ### b. (2 point) This is false. When a weak predictor is added to a multiple regression model, the R^2_{adj} can decrease if the decrease in the SSE is not enough to offset the

decrease in the error degrees of freedom. If the second variable (the one with the lower original R^2) is a very weak predictor, it may actually decrease the R^2_{adj}.

Q3.10 (4 points)

a. (2 points)

Positive. Every year people drive their cars, so each year the car is older, the moretotal miles the car will have been driven. ### b. (2 points) Negative. The more miles the car has been driven,the lower the price of the used car will be.

Q3.22 (8 points)

a. (2 points)

$\frac{SSR}{SST} = \frac{9350}{17190} = 54.4\%$ We can explain 54.4% of the variability in calories in these cereals by using grams of sugar and grams of fiber in a multiple regression model. ### b. (2 points) $\hat{\sigma}_\epsilon = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{\frac{7840}{36-2-1}} = 15.4$ ### c. (2 points) $F = \frac{MSModel}{MSE} = \frac{9350/2}{7840/33} = 19.7$ ### d. (2 points) $H_0 : \beta_1 = \beta_2 = 0$ vs $H_1 : \beta_1 \neq 0$ or $\beta_2 \neq 0$ F = 19.7 p-value = $0.000002 < 0.05$ We reject H_0. In other words, either sugar or fiber, or both of them contribute to the calarie content.