MSDS 601 Homework 2

Tianqi Wang

9/3/2018

Question 1

(a)

I think the conclusion is warranted, since the 95% coffidence intervals of β_1 does not include 0, which shows the null hypotheis that $\beta_1 = 0$ could be rejected at 0.05 significance level. The implied significance level is 0.05.

(b)

I think argue the value of the lower interval confidence limit at X=0 is not appropriate, the interval confidence does not necessarrily provide meaningful information when 0 is not in the scope in the given model.

Question 2

```
Kelley <- read.csv("KelleyBlueBookData.csv")
summary(lm(Price ~ Mileage, data = Kelley))
```

```
##
## Call:
## lm(formula = Price ~ Mileage, data = Kelley)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13905  -7254  -3520   5188  46091
##
## Coefficients:
##                Estimate  Std. Error t value             Pr(>|t|)
## (Intercept) 24764.55901   904.36328  27.383 < 0.0000000000000002 ***
## Mileage        -0.17252     0.04215  -4.093            0.0000468 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9789 on 802 degrees of freedom
## Multiple R-squared:  0.02046,    Adjusted R-squared:  0.01924
## F-statistic: 16.75 on 1 and 802 DF,  p-value: 0.00004685
```

t^2 = 0 which is equal to F-statistic = 16.7545404

Question 3

- For one thing, t test could individually test each parameters, while F test could only be used when take all parameters as a whole.
- For another, t test could be used to test one-sided or two-sided tests, while F automatically test two-sided tests.
- And also, t test could be used to test null hypothesis that parameters equal to other values other than 0, while F test only test for parameters equal to zero or not.

Question 4

The F-statistics would be large if β_1^2 is large, so for testing β_1^2 is large should be same as testing $\beta_1 < 0$ and $\beta_1 > 0$.

Question 6

```
data <- read.csv("BodyFatPercentage.csv")
data <- filter(data, BODYFAT != 0)
new = data.frame(ABDOMEN = 60)
(summary(lm(BODYFAT ~ ABDOMEN, data = data)))
```

```
##
## Call:
## lm(formula = BODYFAT ~ ABDOMEN, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.4044  -3.5186  -0.0367   3.1052  11.9594
##
## Coefficients:
##              Estimate Std. Error t value            Pr(>|t|)
## (Intercept) -34.76949    2.48522  -13.99 <0.0000000000000002 ***
## ABDOMEN       0.58051    0.02665   21.79 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.51 on 249 degrees of freedom
## Multiple R-squared:  0.6559, Adjusted R-squared:  0.6545
## F-statistic: 474.6 on 1 and 249 DF,  p-value: < 0.00000000000000022
```

```
b0 = (summary(lm(BODYFAT ~ ABDOMEN, data = data)))$coefficients[1,
    1]
b1 = (summary(lm(BODYFAT ~ ABDOMEN, data = data)))$coefficients[2,
    1]
MSE = anova(lm(BODYFAT ~ ABDOMEN, data = data))["Mean Sq"][2,
    ]
n = nrow(data) - 2
```

Using R to generate both the predict interval and the confidence interval:

- The prediction interval covering 80% of all individuals with an abdominal circumference of 60 centimeters is

```
predict(lm(BODYFAT ~ ABDOMEN, data = data), new, interval = "predict",
    level = 0.8, df = n - 2)
```

```
##          fit       lwr      upr
## 1 0.06125596 -5.852351 5.974863
```

- The 90% confidence interval for the mean body fat percentage of all individuals with an abdominal circumference of 60 centimeters is

```
predict(lm(BODYFAT ~ ABDOMEN, data = data), new, interval = "confidence",
    level = 0.9, df = n - 2)
```

```
##          fit       lwr      upr
## 1 0.06125596 -1.450055 1.572567
```

Using the formulars shared in class to generate both the prediction interval and the confidence interval:

- The prediction interval covering 80% of all individuals with an abdominal circumference of 60 centimeters is
$$(\hat{\beta}_0 + \hat{\beta}_1 * 60 - t_{0.9}(249) * \sqrt{MSE(1 + \frac{1}{n} + \frac{(x_h - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2})},$$
$\hat{\beta}_0 + \hat{\beta}_1 * 60 + t_{0.9}(249) * \sqrt{MSE(1 + \frac{1}{n} + \frac{(x_h - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2})}$) which is (-5.852351,+5.972567), same as R got.

- The 90% confidence interval for the mean body fat percentage of all individuals with an abdominal circumference of 60 centimeters is
$$(\hat{\beta}_0 + \hat{\beta}_1 * 60 - t_{0.95}(249) * \sqrt{MSE(\frac{1}{n} + \frac{(x_h - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2})},$$
$\hat{\beta}_0 + \hat{\beta}_1 * 60 + t_{0.95}(249) * \sqrt{MSE(\frac{1}{n} + \frac{(x_h - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2})}$) which is (-1.45,+1.57), same as R got.

Question 7

Since We have

$$\sigma^2\{pred\} = \sigma^2(\frac{1}{n} + \frac{(x_h - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2})$$

and

$$\sigma^2\{\hat{Y}_h\} = \sigma^2(1 + \frac{1}{n} + \frac{(x_h - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2})$$

Since in $\sigma^2\{pred\}$, as n becomes large, both $\frac{1}{n}$ and $\frac{(x_h - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ would be brought increasingly close to zero, so $\sigma^2\{pred\}$ would be brought increasingly close to zero.

But for $\sigma^2\{\hat{Y}_h\}$, there is an additional 1 in the parenthesis, it could only be brought increasingly close to σ^2.

The implication is, when we add the observations, the we can narrow the prediction invervals and bring it closer to zero, but in this way we can not eliminate the prediction errors.

Question 8

```
air <- read.csv("AirFreightBreakage.csv")
(lm <- summary(lm(NumberBrokenAmpules ~ NumberTransfers, data = air)))
```

```
##
## Call:
## lm(formula = NumberBrokenAmpules ~ NumberTransfers, data = air)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -2.2   -1.2    0.3    0.8    1.8
##
## Coefficients:
##                 Estimate Std. Error t value    Pr(>|t|)
## (Intercept)      10.2000     0.6633  15.377 0.000000318 ***
## NumberTransfers   4.0000     0.4690   8.528 0.000027487 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.483 on 8 degrees of freedom
## Multiple R-squared:  0.9009, Adjusted R-squared:  0.8885
## F-statistic: 72.73 on 1 and 8 DF,  p-value: 0.00002749
```

(a)

The 95% confidence intervals for β_1 is (2.9183882,5.0816118) Interpretation: There is 1 95% chance that the true value of β_1 lies in this inverval.

(b)

Null Hypothesis: H_0: $\beta_1 = 0$ Alternative Hypothesis: H_a :β is not zero, which means there do exist linear assosiation between X a carton is transferred and the number of broken ampules Y. Decisiton rule: check the p-value of the t-test, if the p-value is less than the significance level which is $\alpha = 0.05$, then we can reject the null hypothesis. p-value is 0.0000275 which is less than significance level 0.05. So we can get reject the null hypothesis under the significance level 0.05.

(c)

The variance of $\hat{\beta}_0$ is $MSE(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2})$, so we can calculate the confidence, which is (8.6703699,11.7296301)

Interpretation: There is 1 95% chance that the true value of β_0 lies in this inverval.

Question 9

The formula for $\sigma^2\{\hat{Y}_h\}$ is

$$\sigma^2\{\hat{Y}_h\} = \sigma^2(1 + \frac{1}{n} + \frac{(x_h - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2})$$

Question 10

```
data10 <- read.csv("FiveYearsMAXSATvsFYGPA.csv")
```

(a)

```
lm10 <- lm(termgpa ~ maxscore, data = data10)
(anova10 <- anova(lm10))
```

```
## Analysis of Variance Table
##
## Response: termgpa
##             Df Sum Sq Mean Sq F value                Pr(>F)
## maxscore     1 111.22 111.220  524.17 < 0.000000000000000022 ***
## Residuals 2145 455.13   0.212
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- For the independent variable `maxscore`, Mean Sq is equal to Sum Sq divided by Df. That is:
  `anova10[1,2]/anova10[1,1]` = 111.2199685 = `anova10[1,3]`
- For the residuals, Mean Sq is also equal to Sum Sq divided by Df. That is
  `anova10[2,2]/anova10[2,1]` = 0.212183 = `anova10[2,3]`
- The F-value equals to Mean Sq of independent variable `maxscore` divided by Mean Sq of residuals, which is:
  `anova10[1,3]/anova10[2,3]` = 524.1700532 = `anova10[1,4]`

(b)

Null Hypothesis: $\beta_1 = 0$ Alternative Hypothesis: $\beta_1 \neq 0$ Decisiton rule: check the p-value of the F-test, if the p-value is less than the significance level which is $\alpha = 0.05$, then we can reject the null hypothesis. Since the p-value for F-test is 0 < 0.05, we can reject the null hypothesis at significance level 0.05 and accept the alternative hypothesis that $\beta_1 \neq 0$, so we can say there are linear relationship between dependent variable and independent variables at significance level 0.05

(c)

```r
summary(lm10 <- lm(termgpa ~ maxscore, data = data10))
```

```
##
## Call:
## lm(formula = termgpa ~ maxscore, data = data10)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.02465 -0.25605  0.06924  0.33196  1.10310
##
## Coefficients:
##              Estimate Std. Error t value             Pr(>|t|)
## (Intercept) 0.8015026  0.1057192   7.581   0.0000000000000506 ***
## maxscore    0.0018887  0.0000825  22.895 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4606 on 2145 degrees of freedom
## Multiple R-squared:  0.1964, Adjusted R-squared:  0.196
## F-statistic: 524.2 on 1 and 2145 DF,  p-value: < 0.00000000000000022
```

The p-value for the t-test of the β_1 is equal to the p-value of F-test in part(b), since it is the simple model, they are actually saying the samething.

(d)

```r
anova10
```

```
## Analysis of Variance Table
##
## Response: termgpa
##             Df Sum Sq Mean Sq F value                Pr(>F)
## maxscore     1 111.22 111.220  524.17 < 0.00000000000000022 ***
## Residuals 2145 455.13   0.212
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

R^2 = anova10[1,2]/anova10[2,2] = 0.2443683,which means that about 24.4% of the variation in Y could be explained by the variation in X.