Homework 2

Solution

Question 1. Sweetness of orange juice (Data set: OJUICE)

(a) (5 pts) Fit the model and find a 90% confidence interval for the true slope of the line. Interpret the result. (Hint: You need to decide which R code to use. confint()? or predict()?)

```
### Read the data
setwd("/cloud/project/STAT 3113 Data Sets")
df_ojuice = read.csv("OJUICE.csv", header=TRUE, sep=',', dec='.')
names(df_ojuice)
```

```
## [1] "Run"        "SweetIndex" "Pectin"
```

```
attach(df_ojuice)
### Fit the SLR model

fit_ojuice = lm(SweetIndex~Pectin)


### Find the confidence interval required
confint(fit_ojuice, level = 0.90)
```

```
##                      5 %          95 %
## (Intercept)  5.845753857  6.6583819614
## Pectin      -0.003864436 -0.0007568157
```

Solution: The 90% confidence interval for the slope is [-0.00386, -0.00076].

Interpretation: We are 90% confident that the change in the mean sweetness index for each one unit chang in the pectin is between -0.00386 and -0.00076.

(b) (10 pts) Fit the model and determine whether there is a positive or negative linear relationship between the amount of pectin x and the sweetness y. That is, determine if there is sufficient evidence (at α=0.05) to indicate that β_1, the slope of the straight-line model, is significantly different from zero. Hints:

- You'll need to use summary() in R to find the t-statistic or p-value.
- You could use either the critical value method or the p-value method.
- Feel free to write your answer to this question on a piece of paper, if needed. Then take a picture and upload to Blackboard.

```
summary(fit_ojuice)
```

```
##
## Call:
## lm(formula = SweetIndex ~ Pectin)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.54373 -0.11039  0.06089  0.13432  0.34638
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.2520679  0.2366220  26.422   <2e-16 ***
## Pectin      -0.0023106  0.0009049  -2.554   0.0181 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.215 on 22 degrees of freedom
## Multiple R-squared:  0.2286, Adjusted R-squared:  0.1936
## F-statistic:  6.52 on 1 and 22 DF,  p-value: 0.01811
```

Solution:

- Method I: P-value Method

The hypotheses are

$H_0 : \beta_1 = 0$

$H_a : \beta_1 \neq 0$

Because the p-value is $0.01811 < \alpha = 0.05$, we reject H_0. Therefore, we conclude β_1 is significant different from 0. There is a significant linear relationship between amount of pectin (x) and the sweetness (y).

- Method II: Critical Value Method

The hypotheses are

$H_0 : \beta_1 = 0$

$H_a : \beta_1 \neq 0$

The t-statistic is $t = -2.554$. The critical value is $t_{\alpha/2} = t_{0.025,df=22} = 2.07$ as follows.

Because $t = -2.554 < -2.07$, we reject H_0. We conclude there is a significant linear relationship between amount of pectin and sweetness.

Question 2. Joint Strike Fighter program (Data set: F35)

```
### Import data and fit the SLR model

setwd("/cloud/project/STAT 3113 Data Sets")
df_F35 = read.csv("F35.csv", header=TRUE, sep=',', dec='.')
names(df_F35)
```

```
## [1] "Year"  "Model" "Cost"
```

```
attach(df_F35)

fit_F35 = lm(Cost ~ Year)
summary(fit_F35)
```

```
##
## Call:
## lm(formula = Cost ~ Year)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.872  -9.973  -7.300   6.606  30.200
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3675.4833   723.7437  -5.078 0.000479 ***
## Year            1.8705     0.3644   5.133 0.000442 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.2 on 10 degrees of freedom
## Multiple R-squared:  0.7249, Adjusted R-squared:  0.6974
## F-statistic: 26.35 on 1 and 10 DF,  p-value: 0.0004422
```

```
anova(fit_F35)
```

```
## Analysis of Variance Table
##
## Response: Cost
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Year       1 6083.8  6083.8  26.348 0.0004422 ***
## Residuals 10 2309.1   230.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(a) (5 pts) Fit the simple linear regression model, $E(y) = \beta_0 + \beta_1 x$, to the data.

Solution: The estimated linear regression equation is $\hat{y} = -3675.48 + 1.87x$

(b) (5 pts) List assumptions required for the regression analysis.

Solution: Assumptions on random error, ϵ

1. $E(\epsilon) = 0$.
2. $var(\epsilon) = \sigma^2$ is constant.
3. ϵ has a normal distribution.
4. ϵ's are independent.

(c) (5 pts) Find the value of SSE.

Solution: SSE = 2309.1

(d) (5 pts) Find the estimated standard error of the regression model, s.

Solution: s = SE = 15.2

(e) (5 pts) Give a practical interpretation of s.

Solution: We would expect most of the observed values to fall within $2s$ or 2*15.2=30.4 units of the least square line.

(f) (5 pts) Find a 95% confidence interval for the true slope of the line. (Hint: You need to decide which R code to use. confint()? or predict()?)

```
confint(fit_F35, level=0.95)
```

```
##                   2.5 %       97.5 %
## (Intercept) -5288.08491 -2062.881779
## Year            1.05853     2.682401
```

Solution:

95% C.I. = (1.058, 2.682)

(g) (5 pts) Interpret the confidence interval in (f).

Solution: We are 95% confident that for each unit increase in the year of initial aircraft operation, the estimated annual cost increases between 1.058 to 2.682 millions of dollars.

(h) (5 pts) Find the p-value for testing $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$. Use this result to test the simple linear regression model is statistically useful for predicting the annual cost using the year of initial operation. (Test using $\alpha = 0.05$)

Solution:

The hypotheses are

$H_0 : \beta_1 = 0$

$H_a : \beta_1 \neq 0$

From the R output above, p-value = 0.000442. Because p-value = $0.000442 < \alpha = 0.05$, we reject H_0. Then we conclude there is a significant linear relationship between the estimated annual cost and the year of initial operation.

(i) (5 pts) Find and interpret the coefficient of determination, r^2.

Solution:

$r^2 = 0.7249$

Interpretation: About 72.49% of the total sample variability of the estimated annual cost could be explained by the linear relationship between the estimated annual cost and the year of initial operation.

(j) (5 pts) A researcher wants to estimate of the average annual cost of all the military aircraft with the year of initial operation in 1980. Which interval is desired by the researcher, a 95% prediction interval for y or a 95% confidence interval for $E(y)$? Use R to calculate the desired interval. (Hint: You need to decide which R code to use. confint()? or predict()?)

```
newdata_year = data.frame(Year = 1980)
CI = predict(fit_F35, newdata = newdata_year, se.fit=TRUE, interval = "confidence", level=0.95)

CI$fit
```

```
##       fit     lwr      upr
## 1 28.038 17.0869 38.98911
```

```
### In this question, as average annual cost is considered, CI should be calculated.
### If PI needs to be calculated, the code is as follows.
# PI = predict(fit_F35, newdata = newdata_year, interval = "prediction", level=0.95)
# PI$fit
```

As the average annual cost is estimated, the researcher should use confidence interval.

95% C.I. = (17.09, 38.99)

(k) (5 pts) Give a practical interpretation of the interval in part (j).

Solution: We are 95% confident the mean value of annual cost of the military aircraft with the year of initial operation in 1980 is between 17.09 millions and 38.99 millions of dollars.

Question 3. Fill in the blanks in the table and answer questions

For future planning and budgeting, the researchers want to analyze the relationship between the total area of structurally deficient bridges in a state and the number of deficient bridges. A simple linear regression model was fitted. In this analysis,

x = number of structurally deficient bridges,

y = the total area (thousands of square feet) of the deficient bridges.

The MINITAB output is as follows:

Regression Analysis: SDArea versus NumberSD

The regression equation is
SDArea = 120 + 0.346 NumberSD

Predictor	Coef	SE Coef	T	P
Constant	119.9	123.0	0.97	0.335
NumberSD	0.34560	0.06158	(1)	(2)

S = (3) R-Sq = 38.7% R-Sq(adj) = 37.4%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	12710141	(4)	31.50	0.000
Residual Error	50	20173111	(5)		
Total	51	32883252			

Figure 1: Fill in the blanks in the SLR table

(a) (5 pts) Please fill in the five blanks in the table.

(1) $= 0.3456/0.06158 = 5.6122$

(2) $= 0.000$

(3) $= \sqrt{MSE} = \sqrt{403462.22} = 635.19$

(4) $= 12710141/1 = 12710141$

(5) $= 20173111/50 = 403462.22$

(b) (5 pts) Find and interpret the coefficient of determination, r^2.

Solution: $r^2 = 38.7\%$

Interpretation: 38.7% of the sum of squares of deviations of the total area of the deficient bridges could be explained by using the number of structurally deficient bridges as a predictor.

(c) (5 pts) Calculate the coefficient of correlation, r.

Solution: As $\hat{\beta}_1 = 0.3456 >=$, the sign of β_1 is $+$.

$r = (\text{sign of } \beta_1)\sqrt{r^2} = \sqrt{0.387} = 0.622$

Question 4. Recalling student names (Data set: NAMEGAME2)

```
df_namegame = read.csv("STAT 3113 Data Sets/NAMEGAME2.csv", header=TRUE, sep=',', dec='.')

names(df_namegame)
```

```
## [1] "POSITION" "RECALL"
```

```r
attach(df_namegame)

fit_namegame = lm(RECALL~POSITION)

### Calculate CI and PI when POSITION=5
# Method I
CI = predict(fit_namegame, se.fit=TRUE, interval = "confidence", level = 0.99)
cbind(df_namegame, CI$fit)[df_namegame$POSITION==5,]
```

```
##    POSITION RECALL       fit       lwr       upr
## 55        5   0.97 0.7025529 0.6459537 0.7591521
## 56        5   0.27 0.7025529 0.6459537 0.7591521
## 57        5   0.39 0.7025529 0.6459537 0.7591521
## 58        5   0.99 0.7025529 0.6459537 0.7591521
## 59        5   0.99 0.7025529 0.6459537 0.7591521
## 60        5   0.99 0.7025529 0.6459537 0.7591521
## 61        5   0.99 0.7025529 0.6459537 0.7591521
## 62        5   0.73 0.7025529 0.6459537 0.7591521
## 63        5   0.99 0.7025529 0.6459537 0.7591521
## 64        5   0.85 0.7025529 0.6459537 0.7591521
## 65        5   0.01 0.7025529 0.6459537 0.7591521
## 66        5   0.76 0.7025529 0.6459537 0.7591521
## 67        5   0.99 0.7025529 0.6459537 0.7591521
## 68        5   0.15 0.7025529 0.6459537 0.7591521
## 69        5   0.33 0.7025529 0.6459537 0.7591521
## 70        5   0.85 0.7025529 0.6459537 0.7591521
## 71        5   0.26 0.7025529 0.6459537 0.7591521
## 72        5   0.93 0.7025529 0.6459537 0.7591521
```

```r
PI = predict(fit_namegame, se.fit=TRUE, interval = "prediction", level = 0.99)
cbind(df_namegame, PI$fit)[df_namegame$POSITION==5,]
```

```
##    POSITION RECALL       fit        lwr      upr
## 55        5   0.97 0.7025529 0.03656847 1.368537
## 56        5   0.27 0.7025529 0.03656847 1.368537
## 57        5   0.39 0.7025529 0.03656847 1.368537
## 58        5   0.99 0.7025529 0.03656847 1.368537
## 59        5   0.99 0.7025529 0.03656847 1.368537
## 60        5   0.99 0.7025529 0.03656847 1.368537
## 61        5   0.99 0.7025529 0.03656847 1.368537
## 62        5   0.73 0.7025529 0.03656847 1.368537
## 63        5   0.99 0.7025529 0.03656847 1.368537
## 64        5   0.85 0.7025529 0.03656847 1.368537
## 65        5   0.01 0.7025529 0.03656847 1.368537
## 66        5   0.76 0.7025529 0.03656847 1.368537
## 67        5   0.99 0.7025529 0.03656847 1.368537
## 68        5   0.15 0.7025529 0.03656847 1.368537
## 69        5   0.33 0.7025529 0.03656847 1.368537
## 70        5   0.85 0.7025529 0.03656847 1.368537
## 71        5   0.26 0.7025529 0.03656847 1.368537
## 72        5   0.93 0.7025529 0.03656847 1.368537
```

```r
# Method II
new.position = data.frame(POSITION=5)
CI = predict(fit_namegame, newdata=new.position, se.fit=TRUE, interval="confidence", level=0.99)
```

```
CI$fit
```

```
##         fit       lwr       upr
## 1 0.7025529 0.6459537 0.7591521
```

```
PI = predict(fit_namegame, newdata=new.position, se.fit=TRUE, interval="prediction", level=0.99)
PI$fit
```

```
##         fit        lwr      upr
## 1 0.7025529 0.03656847 1.368537
```

(a) (5 pts) Find a 99% confidence interval for the mean recall proportion for students in the fifth position during the "name game." Interpret the result.

Solution: From the R output above, we find

99% CI = [0.646, 0.759]

Interpretation: We are 99% confident that the mean recall of all those in the 5th position is between 0.646 and 0.759.

(b) (5 pts) Find a 99% prediction interval for the recall proportion of a particular student in the fifth position during the "name game." Interpret the result.

Solution: From the R output above, we find

99% PI = [0.037, 1.369]

Interpretation: We are 99% confident that the actual recall of a person in the 5th position is between 0.037 and 1.369.

(c) (5 pts) Compare the two intervals, part (a) and part (b). Which interval is wider? Will this always be the case? Explain.

Solution: The prediction interval in part b is wider than the confidence interval in part a. The prediction interval will always be wider than the confidence interval. The confidence interval for the mean is an interval for predicting the mean of all observations for a particular value of x. The prediction interval is a confidence interval for the actual value of the dependent variable for a particular value of x. The variation of predicting the actual value is larger than estimating the mean value.