

# Advanced Data Analysis HW3

Ao Liu, al3472

1.

Consider a regression model with  $p$  predictors, that is,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \dots + \beta_p X_{pi} + \epsilon_i, i = 1, 2, \dots, n$$

(a)

Show that

$$F = \frac{n-p-1}{p} \frac{R^2}{1-R^2}$$

Answer:

$$\begin{aligned} F &= \frac{SSR/dfR}{SSE/dfE} \\ &= \frac{SSR}{SST - SSR} \times \frac{n-p-1}{p} \\ &= \frac{1}{\frac{SST}{SSR} - 1} \times \frac{n-p-1}{p} \\ &= \frac{1}{1/R^2 - 1} \times \frac{n-p-1}{p} \end{aligned}$$

(b)

If  $n=20$ ,  $p=3$ ,  $R^2 = 0.572$ . Test  $H_0: \beta_1 = \beta_2 = \beta_3 = 0$  against  $H_a$ : at least one of them is not zero.

Answer:

```
1 > n=20
2 > p=3
3 > R2=0.572
4 > F=(n-p-1)/p*R2/(1-R2)
5 > F
6 > qf(.95, df1=p, df2=n-p-1)
```

```
1 7.12772585779782
2 3.23887151745358
```

Since F statistic is greater than  $F(0.95, 3, 16)$ , then we cannot reject the Null hypothesis that the parameters are all 0.

2.

Comp-U-Systems, a computer manufacturer, sells and services the Comp-Y-Systems Microcomputers. Let  $x_i$  = the number of microcomputer serviced on the  $i$ th service call  $y_i$  = the number of minutes required to perform service on the  $i$ th service call the data is in file CompUSys.csv. Suppose the model  $y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, 2, \dots, n$ , is used to model the relationship between the number of minutes required to perform a service and the number of microcomputers serviced.

(a)

**Estimate  $\beta_0$  and  $\beta_1$  using the least square method. Interpret the estimate of  $\beta_1$ .**

**Answer:**

```
1 > data = read.csv("CompUSys.csv")
2 > y = data[,2]
3 > x = data[,1]
4 > lm(y~x)
```

```
1 Call:
2 lm(formula = y ~ x)
3
4 Coefficients:
5 (Intercept)          x
6      11.46       24.60
```

By using the least square method, the estimation of  $\beta_0$  is 11.46 and the estimation of  $\beta_1$  is 24.60. If the number of microcomputer serviced on the  $i$ th call increases by 1, then the number of minutes required to perform service on  $i$ th service call will increase by 24.60 on average.

(b)

**Use a 95% confidence interval to estimate  $\beta_1$ . Interpret your result**

**Answer:**

```
1 > confint(lm(y~x))

1 2.5 % 97.5 %
2 (Intercept) 3.684472 19.24371
3 x          22.782272 26.42215
```

According to the result we got above, the 95% confidence interval for  $\beta_1$  is:

$$(22.7822, 26.4221)$$

Which means we are 95% confident that one unit increase in the number of microcomputer serviced on the  $i$ th service call will increase the number of minutes required to perform service on the  $i$ th call by a range from 22.7822 to 26.4221 on average.

(c)

**Estimate the average time it will take to serve 6 microcomputer using a 95% confidence interval. Interpret your result.**

**Answer:**

```
1 > predict(lm(y~x), newdata=data.frame(x=6), interval="confidence")

1      fit      lwr      upr
2 1 159.0773 154.1388 164.0159
```

The output shows that a 95% confidence interval for the average number of minutes required to perform service for 6 microcomputers is

$$[154.1388, 164.0159]$$

Interpretation: We are 95% confident that the average number of minutes it takes to serve 6 microcomputers ranges between 154.1388 and 164.0159.

(d)

Compute a 95% prediction interval for the amount of time it will take to service 6 microcomputers. Interpret your result.

**Answer:**

```
1 > predict(lm(y~x),newdata=data.frame(x=6),interval="prediction")
2
```

```
1      fit      lwr      upr
2 1 159.0773 147.5279 170.6268
```

The output shows that a 95% prediction interval for the average number of minutes required to perform service for 6 microcomputers is

[147.5279, 170.6268]

Interpretation: We are 95% confident that the number of minutes it takes to serve 6 microcomputers ranges between 154.1388 and 164.0159.

(e)

Use the Bonferroni method and to find a joint confidence intervals for the mean amounts of time it will take to serve 6 and 7 microcomputers.

**Answer:**

Since there are two intervals in this method, we divide  $\alpha$  by 4:

```
1 > predict(lm(y~x),newdata=data.frame(x=6),interval="confidence",level=1-0.05/4)
2 > predict(lm(y~x),newdata=data.frame(x=7),interval="confidence",level=1-0.05/4)
```

```
1      fit      lwr      upr
2 1 159.0773 152.2858 165.8689
3      fit      lwr      upr
4 1 183.6796 174.8148 192.5443
```

The joint confidence intervals for the mean amounts of time it will take to serve 6 microcomputers is [152.285, 165.8689] and [174.8148, 192.5443] for serving 7 microcomputers.

(f)

Test

$H_0 : E(Y|X=x) = \beta_0 + \beta_1 x$   $H_a : \text{Not } H_0$  Using  $\alpha = 0.05$ .

**Answer:**

```
1 > reduced = lm(y~x)
2 > full = lm(y~factor(x))
3 > anova(reduced, full)
```

```
1 Res.Df RSS      Df Sum of Sq F      Pr(>F)
2 9      191.7017 NA    NA      NA    NA
3 4      100.0000 5     91.70166 0.7336133 0.6353456
```

3.

International Oil Inc. Is attempting to a develop a reasonably priced minimum unleaded gasoline that will deliver higher gasoline mileage than can be achieved by its current premium unleaded gaso- lines. As part of its development process, International Oil Inc. wishes to study the effect of one qualitative variable, x1, premium gasoline un- leaded type (A, B, C) and one quantitative variable x2 amount of gaso- line additive VST (0, 1, 2, 3 units) on the gasoline mileage y obtained by an automobile called Encore. For testing purposes a sample of 22 Encores is randomly selected and driven under normal driving condi- tions. The combination of x1 and x2 used in the experiment along with the corresponding values of y are in file mileage.csv. Define  $[A,x]$ ,  $[B,x]$  and  $[C,x]$  to be the mean unleaded gasoline mileage by Encore when using AST amount x and premium unleaded gasoline types A, B and C, respectively. Consider the model

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 x_2 + \epsilon_i$$

where  $D_{1i} = 1$  gas type is B and 0 other wise and  $D_{2i} = 1$  is gas type is C and 0 otherwise.

(a)

Estimate the  $\beta_{a_{is}}$  and interpret your result (see note for how to fit this model )

**Answer:**

```
1 > data = read.csv("mileage.csv")
2 > y = data[,1]
3 > x1 = data[,2]
4 > x2 = data[,3]
5 > lm(y ~ factor(x1)+x2)
6 > summary(lm(y ~ factor(x1)+x2))
```

```
1 Call:
2 lm(formula = y ~ factor(x1) + x2)
3
4 Residuals:
5      Min       1Q   Median       3Q      Max
6 -4.6171 -1.6321  0.5508  1.3756  4.0021
7
8 Coefficients:
9             Estimate Std. Error t value Pr(>|t|)
10 (Intercept)  32.0171     1.0005   32.002  <2e-16 ***
11 factor(x1)B   1.5218     1.2650    1.203   0.245
12 factor(x1)C   0.5252     1.6194    0.324   0.749
13 x2           -0.4192     0.6042   -0.694   0.497
14 ---
15 Signif. codes:  0   ***    0.001   **    0.01   *    0.05   .    0.1    1
16
17 Residual standard error: 2.532 on 18 degrees of freedom
18 Multiple R-squared:  0.09453, Adjusted R-squared:  -0.05638
19 F-statistic: 0.6264 on 3 and 18 DF,  p-value: 0.6072
```

The estimation for  $\beta_0$  is 32.0171, meaning that when the two factors have no effect, the mileage on average is 32.0171;

The estimation for  $\beta_1$  is 1.5218, meaning that when VST amount is the same, choosing type B will have 1.5218 more milage than choosing type A on average;

The estimation for  $\beta_2$  is 0.5252, meaning that when VST amount is the same, choosing type C will have 0.5252 more milage than chooSing type A on average;  
The estimation for  $\beta_3$  is -0.4192, meaning that when the type is the same, increasing 1 unit of VST will cause 0.4192 less milage on average.

(b)

**Test  $H_0 : \beta_1 = \beta_2 = 0$  against  $H_a$ : Not  $H_0$  using  $\alpha = 0.05$ .**

**Answer:**

According to the result in (a), the F-statistic we got is 0.6264 on 3 and 18 DF, and p-value is 0.6072. Since p-value is greater than 0.05, we cannot reject the null hypothesis.