LINEAR REGRESSION: Homework #6

Professor Jingchen Liu

Fan Yang
UNI: fy2232

Problem 1 (3.14)

(a)

Hypothesis:

$$H_0: \ E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 x_2$$
$$H_1: \ E(Y) \neq \beta_0 + \beta_1 X_1 + \beta_2 x_2$$
$$\text{if } F^* \leq F(0.99, 2, 12), \text{ then conclude } H_0$$
$$\text{if } F^* > F(0.99, 2, 12), \text{ then conclude } H_1$$

Let's compute the lack of fit test:

```
> anova(lm(Y~X,data=d22), lm(Y~factor(X),data=d22))
Analysis of Variance Table
Model 1: Y ~ X
Model 2: Y ~ factor(X)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     14 146.43
2     12 128.75  2    17.675 0.8237 0.4622
>
> qf(0.99,2,12)
[1] 6.926608
```

The lack of fit statistic is $F^* = 0.8237 < 6.926608$ with p-value 0.4622, supporting the linearity of the regression model.

(b)

Having an equal number of replications at each of1he X levels would lead to a smaller error term and get a better fit regression model. But all the effects covered in the error term could not vary at random from one repeated observation to the next because all the Y in the same group are set the same.

(c)

When it leads to nonlinear conclusion, the test in part (a) cannot indicate what regression function is appropriate. I will try to make box-cox transformation on the data and conduct regression on the transformed data.

Problem 2 (7.7)

(a)

```
> model18 <- lm(d18$V1~d18$V2 + d18$V3 + d18$V4 + d18$V5)
> model18_4 <- lm(d18$V1~d18$V5)
> model18_14 <- lm(d18$V1~d18$V2 + d18$V5)
> model18_124 <- lm(d18$V1~d18$V2 + d18$V3 + d18$V5)
> SSR4 <- sum((model18_4$fitted.values-mean(d18$V1))^2)
> SSR1_4 <- -SSR4 + sum((model18_14$fitted.values-mean(d18$V1))^2)
> SSR14 <- sum((model18_14$fitted.values-mean(d18$V1))^2)
> SSR2_14 <- -SSR14 + sum((model18_124$fitted.values-mean(d18$V1))^2)
> SSR124 <- sum((model18_124$fitted.values-mean(d18$V1))^2)
> SSR3_124 <- -SSR124 + sum((model18$fitted.values-mean(d18$V1))^2)
>
> SSR4; SSR1_4; SSR2_14; SSR3_124
[1] 67.7751
[1] 42.27457
[1] 27.85749
[1] 0.4197463
```

$$
\begin{aligned}
SSR(X_4) &= 67.7751, \\
SSR(X_1|X_4) &= 42.27457, \\
SSR(X_2|X_1, X_4) &= 27.85749, \\
SSR(X_3|X_1, X_2, X_4) &= 0.4197463
\end{aligned}
$$

(b)

Hypothesis:

$$
\begin{gathered}
H_0: \ \beta_3 = 0 \\
H_1: \ \beta_3 \neq 0 \\
\text{if } \ F^* \leq F(0.99, 1, n-5), \text{ then conclude } \ H_0 \\
\text{if } \ F^* > F(0.99, 1, n-5), \text{ then conclude } \ H_1 \\
F^* = \frac{SSE(X_1, X_2, X_4) - SSE(X_1, X_2, X_3, X_4)}{(n-4)-(n-5)} \Big/ \frac{SSE(X_1, X_2, X_3, X_4)}{n-5}
\end{gathered}
$$

```
> numerator <- sum(model18_124$residuals^2) - sum(model18$residuals^2)
> denominator <- sum(model18$residuals^2) / model18$df.residual
> Fstatistic = numerator / denominator
> Fstatistic
[1] 0.3247534
> qf(0.99, 1, model18$df.residual)
[1] 6.980578
> 1 - pf(Fstatistic,1,model18$df.residual)
[1] 0.5704457
```

Because $F^* = 0.3247534 \leq F(0.99, 1, n-5) = 6.980578$, with p-value 0.5704457, then we conclude H_0.

Problem 3 (7.10)

Hypothesis:

$$H_0: \ \beta_1 = -.1, \quad \beta_2 = .4$$
$$H_1: \ \beta_3 \neq -.1, \quad \beta_2 \neq .4$$
$$\text{if } F^* \leq F(0.99, 2, n-5), \text{ then conclude } H_0$$
$$\text{if } F^* > F(0.99, 2, n-5), \text{ then conclude } H_1$$

```
> redy <- d18$V1 + 0.1*d18$V2 - 0.4*d18$V3
> model18_red <- lm(redy ~ d18$V4 + d18$V5)
> numerator <- (sum(model18_red$residuals^2) - sum(model18$residuals^2)) / 2
> denominator <- sum(model18$residuals^2) / model18$df.residual
> Fstatistic = numerator / denominator
> Fstatistic
[1] 4.60764
> qf(0.99, 2, model18$df.residual)
[1] 4.89584
> 1 - pf(Fstatistic,2,model18$df.residual)
[1] 0.0129197
```

Because $F^* = 4.60764 \leq F(0.99, 2, n-5) = 4.89584$, with p-value 0.0129197, then we conclude H_0.

Problem 4 (7.16)

(a)

```
> n = nrow(d5)
> sy = sd(d5$V1)
> sx1 = sd(d5$V2)
> sx2 = sd(d5$V3)
> d5$V1 = 1/sqrt(n-1)*(d5$V1-mean(d5$V1))/sy
> d5$V2 = 1/sqrt(n-1)*(d5$V2-mean(d5$V2))/sx1
> d5$V3 = 1/sqrt(n-1)*(d5$V3-mean(d5$V3))/sx2
>
> model5_trans <- lm(d5$V1 ~ d5$V2 + d5$V3)
> summary(model5_trans)
Call:
lm(formula = d5$V1 ~ d5$V2 + d5$V3)
Residuals:
      Min        1Q    Median        3Q       Max
-0.099209 -0.039740  0.000564  0.035794  0.094699
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.626e-17  1.518e-02   0.000        1
d5$V2        8.924e-01  6.073e-02  14.695 1.78e-09 ***
d5$V3        3.946e-01  6.073e-02   6.498 2.01e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

(b)

As computed in part (a), $b_0^* = -2.626 \times 10^{-17},\ b_1^* = 0.8924,\ b_2^* = 0.3946$

(c)

```
> b <- model5_trans$coefficients
> b1 = (sy/sx1) * b[2]
> b2 = (sy/sx2) * b[3]
> b0 = mean(d5$V1) - b1*mean(d5$V2) - b2*mean(d5$V3)
> b0 ; b1 ; b2
[1] 37.65
[1] 4.425
[1] 4.375
```

which is the same result as Problem 6.5b.

Problem 5 (7.24)

(a)

```
> model24a <- lm(d5$V1 ~ d5$V2)
> model24a
Call:
lm(formula = d5$V1 ~ d5$V2)
Coefficients:
(Intercept)        d5$V2
     50.775        4.425
```

The fitted regression function is $Y = 50.775 + 4.425 \times X_1$

(b)

The two regression coefficients for moisture content are both 4.425.

(c)

```
> model24 <- lm(d5$V1 ~ d5$V2 + d5$V3)
> model24a <- lm(d5$V1 ~ d5$V2)
> model24b <- lm(d5$V1 ~ d5$V3)
> SSR1 <- sum((model24a$fitted.values-mean(d5$V1))^2)
> SSR2 <- sum((model24b$fitted.values-mean(d5$V1))^2)
> SSR12 <- -SSR2 + sum((model24$fitted.values-mean(d5$V1))^2)
> SSR1; SSR12
[1] 1566.45
[1] 1566.45
```

Therefore, $SSR(X_1) = SSR(X_1|X_2)$

(d)

In the correlation matrix obtained in Problem 6.5a, the correlation between X_1 and X_2 is 0, which corresponds to the results in parts (b) and (c). The two variables contribute independently to Y.

Problem 6 (7.37)

(a)

```
> model37_12 <- lm(‘Number of active physicians‘ ~
+                    ‘Total population‘+‘Total personal income‘, data = d37)
> model37_123 <- lm(‘Number of active physicians‘ ~
+                    ‘Total population‘+‘Total personal income‘ +
+                    ‘Land area‘, data = d37)
> model37_124 <- lm(‘Number of active physicians‘ ~
+                    ‘Total population‘+‘Total personal income‘ +
+                    ‘Percent of population 65 or older‘, data = d37)
> model37_125 <- lm(‘Number of active physicians‘ ~
+                    ‘Total population‘+‘Total personal income‘ +
+                    ‘Number of hospital beds‘, data = d37)
> model37_126 <- lm(‘Number of active physicians‘ ~
+                    ‘Total population‘+‘Total personal income‘ +
+                    ‘Total serious crimes‘, data = d37)
> SSE12 <- sum((d37$‘Number of active physicians‘-model37_12$fitted.values)^2)
> SSR12 <- sum((model37_12$fitted.values -
+ mean(d37$‘Number of active physicians‘))^2)
> SSR3_12 <- -SSR12 + sum((model37_123$fitted.values -
+                            mean(d37$‘Number of active physicians‘))^2)
> SSR4_12 <- -SSR12 + sum((model37_124$fitted.values -
+                            mean(d37$‘Number of active physicians‘))^2)
> SSR5_12 <- -SSR12 + sum((model37_125$fitted.values -
+                            mean(d37$‘Number of active physicians‘))^2)
> SSR6_12 <- -SSR12 + sum((model37_126$fitted.values -
+                            mean(d37$‘Number of active physicians‘))^2)
>
> R3_12 = SSR3_12 / SSE12
> R4_12 = SSR4_12 / SSE12
> R5_12 = SSR5_12 / SSE12
> R6_12 = SSR6_12 / SSE12
>
> R3_12; R4_12; R5_12; R6_12
[1] 0.02882495
[1] 0.003842367
[1] 0.5538182
[1] 0.007323408
```

(b)

The variable X_5 number of hospital beds has the largest coefficient of partial determination. When X_5 is added to the model, the error sum of squares is reduced 6.17%, which is the most among the four new variables.

```
> SSR3_12; SSR4_12; SSR5_12; SSR6_12
[1] 4063370
[1] 541647.3
[1] 78070132
[1] 1032359
```

The extra sum of squares of X_5 is the largest among the 4 variables.

(c)

Hypothesis:

$$H_0: \ \beta_3 = 0$$
$$H_1: \ \beta_3 \neq 0$$
$$\text{if } F^* \leq F(0.99, 1, n-4), \text{ then conclude } H_0$$
$$\text{if } F^* > F(0.99, 1, n-4), \text{ then conclude } H_1$$
$$F^* = \frac{SSE(X_1, X_2) - SSE(X_1, X_2, X_5)}{(n-3)-(n-4)} \Big/ \frac{SSE(X_1, X_2, X_5)}{n-4}$$

```
> numerator <- sum(model37_12$residuals^2) - sum(model37_125$residuals^2)
> denominator <- sum(model37_125$residuals^2) / model37_125$df.residual
> Fstatistic = numerator / denominator
> Fstatistic
[1] 541.1801
> qf(0.99, 1, model37_12$df.residual)
[1] 6.693223
> 1 - pf(Fstatistic,1,model37_12$df.residual)
[1] 0
```

Because $F^* = 541.1801 > F(0.99, 1, n-4) = 6.693223$, with a very small p-value, then we reject H_0, which means the variable X_5 number of hospital beds is helpful in the regression model.

The coefficient of partial determination of the other 3 variables are much smaller than that of X_5, so the F* test statistics for the other three potential predictor variables will not be as large as the one here.

Actually, the F* test statistics for X_3, X_4, X_6 are 12.94069, 1.681734, 3.216562, respectively, which is much smaller than that of X_5 541.1801.