Stat500(Section002): Homework #2

Due on Spet. 29, 2021 at 11:59pm

Instructor:Naisyin Wang

Tiejin Chen **tiejin@umich.edu**

Problem 1

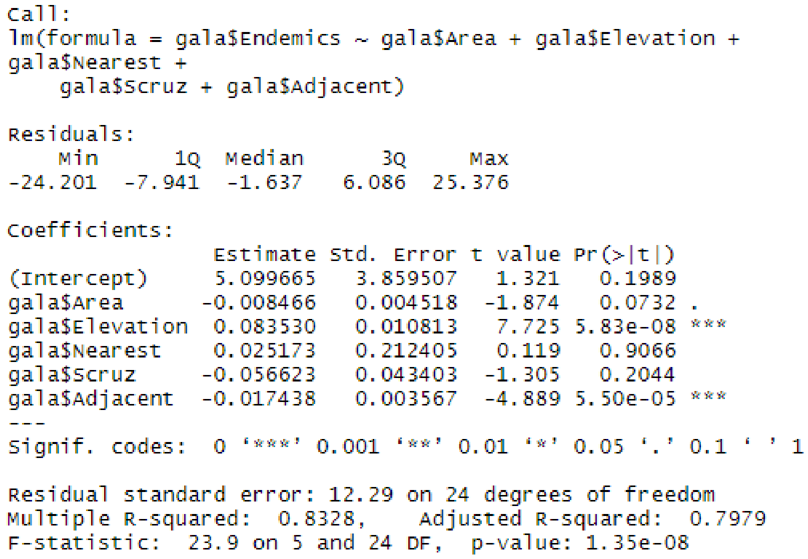
Part a

we use the following code to create a linear regression model, and print the summary of it.

R_code.Rmd

```
linear = lm(gala$Endemics~gala$Area+gala$Elevation+
gala$Nearest+gala$Scruz+gala$Adjacent)
summary(linear)
```

and we get a summary



```
Call:
lm(formula = gala$Endemics ~ gala$Area + gala$Elevation +
gala$Nearest +
    gala$Scruz + gala$Adjacent)

Residuals:
    Min      1Q  Median      3Q     Max 
-24.201  -7.941  -1.637   6.086  25.376 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     5.099665   3.859507   1.321   0.1989    
gala$Area      -0.008466   0.004518  -1.874   0.0732 .  
gala$Elevation  0.083530   0.010813   7.725 5.83e-08 ***
gala$Nearest    0.025173   0.212405   0.119   0.9066    
gala$Scruz     -0.056623   0.043403  -1.305   0.2044    
gala$Adjacent  -0.017438   0.003567  -4.889 5.50e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.29 on 24 degrees of freedom
Multiple R-squared:  0.8328,	Adjusted R-squared:  0.7979 
F-statistic:  23.9 on 5 and 24 DF,  p-value: 1.35e-08
```

we can find that t-test p-value of Elevation and Adjacent is less than 0.001 which proves these features do work in this model. And we can find that the p-value of f-test for all model is also less than 0.01. And R^2 is 0.8328. All this shows that it is suitable to use linear regression model to fit the data.
we use the following code to draw boxplot of residuals.

R_code.Rmd

```
resd = residuals(linear)
boxplot(resd)
```

Problem 1 continued on next page...

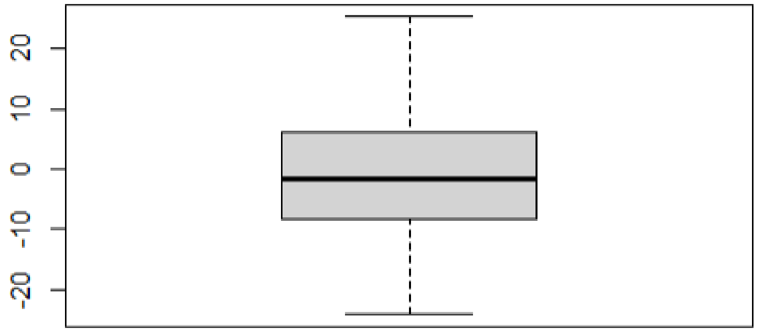


Figure 1: the boxplot of residuals

Part b

we use the following code to get the case number of largest residual.

R_code.Rmd

```
abs_res = abs(resd)
which(abs_res == max(abs_res))
```

case number 25 has the largest residual.

Part c

R_code.Rmd

```
mea = mean(resd)
med = median(resd)
print(mea)
print(med)
```

we can get the mean of resduals is 3.698521e-16, and the median is -1.637416. That is because their are more points have residual below 0. I think we should worry about this. there is more points have residuals below 0 and the mean however is above 0 shows that there might me some 'big wrong' points have residual far from 0. It might be outlier. Hence, we should worry about it.

Part d

R_code.Rmd

```
fit_data = fitted(linear)
cor_data = cor(fit_data ,abs_res)
plot(abs_res ,fit_data)
print(cor_data)
```

We compute the correlation is 0.3013027 and get the plot:

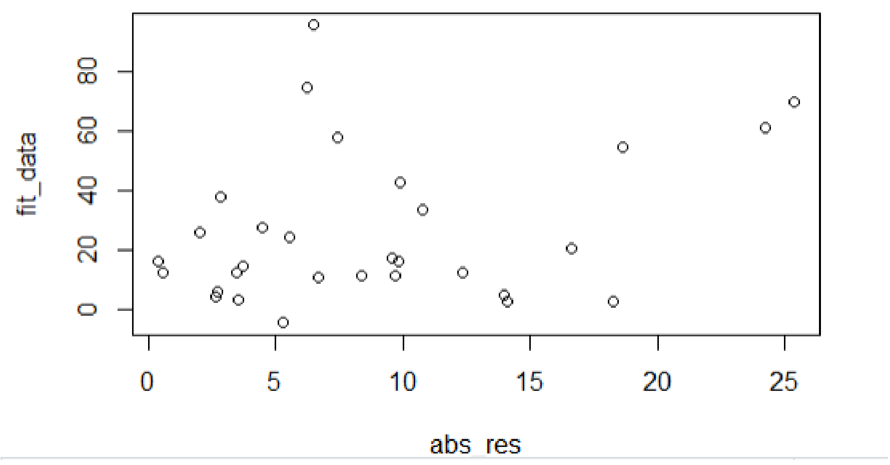


Figure 2: the plot of absolute residuals and fit_data

We can find that there are some linear relationship between fitted data abd absolute residuals. It means when we fit the data higher, we get the the higher residual. That shows that might be some large outliers. which is the same as part c.

Problem 2

For simple linear regression, we have:

$$\hat{y} = a + \beta_x x$$

And we want to make the following formula min.:

$$\sum_{i=1}^{n}(y - \hat{y})^2$$

we can get:

$$\frac{\partial Rss}{\partial a} = \sum_{i=1}^{n} 2(y_i - a - \beta_x x_i) = 0, \frac{\partial Rss}{\partial \beta_x} = \sum_{i=1}^{n} 2(y_i - a - \beta_x x_i)x_i = 0$$

From two formulas, we can derive that:

$$\hat{\beta_x} = \frac{\sum_{i=1}^{n}(x - \overline{x})(y - \overline{y})}{\sum_{i=1}^{n}(x - \overline{x})^2}, \hat{a} = \overline{y} - \hat{\beta_x}\overline{x}$$

We know that:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x - \overline{x})(y - \overline{y})}{\sqrt{\sum_{i=1}^{n}((x - \overline{x})\sum_{i=1}^{n}((y - \overline{y})}}$$

We simplify them to get:

$$\hat{\beta_x} = \frac{Cov(x, y)}{sd(x)^2}, r_{xy} = \frac{Cov(x, y)}{sd(x)sd(y)}$$

Hence:

$$\hat{\beta_x} = r_{xy}\frac{sd(y)}{sd(x)}$$

Problem 3

No, it won't be in most time. From the problem above we know that:

$$\hat{\beta_x} = r_{xy}\frac{sd(y)}{sd(x)}$$

If we change the X and Y, we can get:

$$\hat{\beta_y} = r_{xy}\frac{sd(x)}{sd(y)} \neq \frac{sd(x)}{r_{xy}sd(y)} \ when \ r_{xy} \neq 1$$

Hence unless r_{xy} is 1,$\hat{\beta_y} \neq 1/\hat{\beta_x}$.

Problem 4

we knew that:

$$\frac{\beta_{ddpi} - 0.5}{sd(\beta_{ddpi})} = -0.4602 \sim t_{45}$$

we use the following code to get the p-value:

R_code.Rmd

```
2*(1-pt(0.4602,df=45))
```

The result is 0.6475 which is the same as F-test.Next, we prove they are same mathematically.

Proof. Assumed $H0 : \beta_j = 0.5$ We know that:

$$F = \frac{RSS_{H_0} - RSS_{Full}}{Rss_{Full}/n - (p+1)} \sim F(1, n-(p+1))$$

And:

$$\begin{aligned}
RSS_{H_0} - RSS_{Full} &= \sum_{i=1}^{n}(y_i - \theta^T x_i - 0.5x_{ij})^2 - (y_i - \theta^T x_i - \hat{\beta}_j x_{ij})^2 \\
&= \sum_{i=1}^{n}(\hat{\beta}_j - 0.5)x_{ij}(2y_i - 2\theta^T x_i - (0.5 + \hat{\beta}_j)x_{ij}) \\
&= \sum_{i=1}^{n}(\hat{\beta}_j - 0.5)x_{ij}(2y_i - 2\theta^T x_i - 2\hat{\beta}_j x_{ij} - (0.5 - \hat{\beta}_j)x_{ij}) \\
&= \sum_{i=1}^{n}(\hat{\beta}_j - 0.5)^2 x_{ij}^2 + 2e_i x_{ij} \\
&= \sum_{i=1}^{n}(\hat{\beta}_j - 0.5)^2 x_{ij}^2 + 2\sum_{i=1}^{n} e_i x_{ij} \\
&= \sum_{i=1}^{n}(\hat{\beta}_j - 0.5)^2 x_{ij}^2 = (X^T X)_{jj}(\hat{\beta}_j - 0.5)^2
\end{aligned} \tag{1}$$

Here, the x_i in the formula represent $[x_i 1, ..., x_i(j-1), x_i(j+1), .., x_p]$. X represents $[\hat{x_1}, \hat{x_2}, ..., \hat{x_n}]$ where $\hat{x_i}$ represents $[x_i 1, ..., x_i(j-1), x_i j, x_i(j+1), .., x_p]$ is the full data.
Hence, we get:

$$F = \frac{(\hat{\beta}_j - 0.5)^2}{(X^T X)_{jj} Rss_{Full}/n - (p+1)} = \frac{(\hat{\beta}_j - 0.5)^2}{se(\hat{\beta}_j)^2}$$

Problem 4 continued on next page. . .

We have T for t-test:

$$T = \frac{\hat{\beta}_j - 0.5}{se(\hat{\beta}_j)} \sim t_{n-(p+1)}$$

We also have:

$$t_n^2 \sim F(1, n)$$

Hence, we get:

$$F = T^2$$

And their disturbtions also have square relationship. Thus, they compute the same value, and do the same test indeed. So they must have the same p-value. □