MAT 3375, Fall 2018

Assignment 1

Due in class on October 2, 2018

1. Do the following problems from the textbook.

Textbook Questions:

Solution to 2.16.

```
>df = read.table("data2-16.csv",header=T)
> df
    vol   Pre
1  2084  4599
2  2084  4600
3  2273  5044
4  2273  5043
5  2273  5044
6  2463  5488
7  2463  5487
8  2651  5931
9  2652  5932
10 2652  5932
11 2842  6380
12 2842  6380
13 3030  6818
14 3031  6817
15 3031  6818
16 3221  7266
17 3221  7268
18 3409  7709
19 3410  7710
20 3600  8156
21 3600  8158
22 3788  8597
23 3789  8599
24 3789  8600
25 3979  9048
26 3979  9048
27 4167  9484
28 4168  9487
```

```
29 4168  9487
30 4358  9936
31 4358  9938
32 4546 10377
33 4547 10379
>model=lm(df$Pre~df$vol)
> summary(model)

Call:
lm(formula = df$Pre ~ df$vol)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3276 -0.9227  0.0773  1.2676  2.9577

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.907e+02  1.355e+00  -214.6   <2e-16 ***
df$vol       2.346e+00  4.007e-04  5855.4   <2e-16 ***

> anova(model)
Analysis of Variance Table

Response: df$Pre
          Df    Sum Sq   Mean Sq  F value    Pr(>F)
df$vol     1 103947022 103947022 34286009 < 2.2e-16 ***
Residuals 31        94         3
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> cor(df$vol,df$Pre)
[1] 0.9999995
```

The model

$$Y = -290.7 + 2.346x + \epsilon$$

seems to be a good model.

2.23. Here is the program:

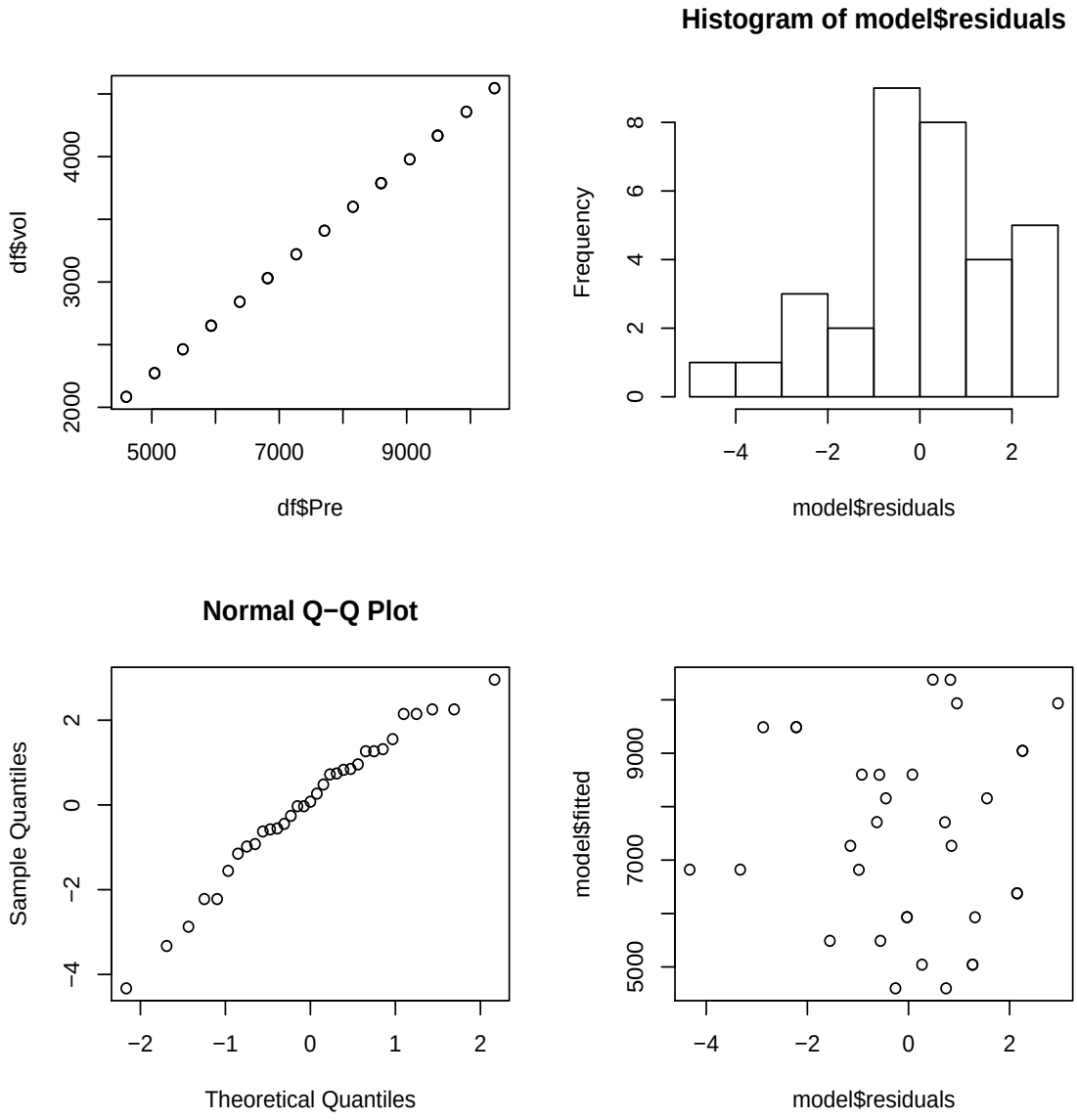
df$vol
df$Pre
Histogram of model$residuals
Frequency
model$residuals
Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
model$fitted
model$residuals

Figure 1:

```
>sum((x-mean(x))**2)
[1]166.25
> x=seq(0.5,10,0.5)
> b0=rep(0,500)
> b1=rep(0,500)
> for(i in 1:500){
+ y=50+10*x+rnorm(20,0,4)
+ reg=lm(y~x)
+ s=sum(reg$residuals^2)/18
+ }
> l=rep(0,500)
> u=rep(0,500)
> x=seq(0.5,10,0.5)
> b0=rep(0,500)
> b1=rep(0,500)
> for(i in 1:500){
+ y=50+10*x+rnorm(20,0,4)
+ reg=lm(y~x)
+ s=sqrt(sum(reg$residuals^2)/18)
+ l[i]=reg$coef[2]-2.100922*s/sqrt(166.25)
+ u[i]=reg$coef[2]+2.100922*s/sqrt(166.25)}
> > sum(l<10 &10<u)/500
[1] 0.954
```

Your result can be a number near 0.95. as well.

2.28. We proved in class that

$$E(\sum_{i=1}^{n} e_i^2/(n-2)) = \sigma^2.$$

Therefore

$$E(\sum_{i=1}^{n} e_i^2/n) = \frac{(n-2)\sigma^2}{n}.$$

Therefore

$$Bias = E(\sum_{i=1}^{n} e_i^2/n) - \sigma^2 = \frac{(n-2)\sigma^2}{n} - \sigma^2 \to 0$$

as $n \to \infty$.

2.32 (a). Since β_0 is known we need to minimize

$$Q(\beta_1) = \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i)^2$$

with respect to β_1. Differentiating with respect to β_1

$$Q^{'}(\beta_1) = \sum_{i=1}^{n} x_i(Y_i - \beta_0 - \beta_1 x_i) = 0$$

implies that

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i(Y_i - \beta_0)}{\sum_{i=1}^{n} x_i^2}.$$

This is reasonable as β_0 is known.

(b) Let

$$w_i = \frac{x_i}{\sum_{i=1}^{n} x_i^2}.$$

Therefore

$$\hat{\beta}_1 = \sum_{i=1}^{n} w_i(Y_i - \beta_0).$$

Note that $\hat{\beta}_1$ is a linear combinations of n independent normal random variables and

$$E(\hat{\beta}_1) = \sum_{i=1}^{n} w_i \beta_1 x_i = \beta_1 \sum_{i=1}^{n} w_i x_i = \beta_1.$$

Also

$$Var(\hat{\beta}_1) = \sum_{i=1}^{n} w_i^2 \sigma^2 = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}.$$

(c) The confidence interval for β_1 is

$$\hat{\beta}_1 \pm t_{\alpha/2}(n-1)\frac{s}{\sum_{i=1}^{n} x_i^2}.$$

In here

$$s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-1}$$

and

$$e_i = Y_i - \beta_0 - \hat{\beta}_1 x_i, i = 1, \ldots, n.$$

Notice that since β_0 is known we do not need to estimate it. Notice that the degrees of freedom for the t distribution is $n-1$

2. Consider the regression model with no intercept

$$Y_i = \beta X_i + \epsilon_i, i = 1, 2, \ldots, n$$

where β is unknown parameter, $X_1, \ldots, X_n$ are given constants and $\epsilon_i, i = 1, 2, \ldots, n$ are i.i.d. random variables with $N(0, \sigma^2)$. Assume σ^2 is also unknown.

(i) Find the MLE for β and σ^2 (denote by $\hat{\beta}$ and $\hat{\sigma}^2$).

Solution. Similar to question 2.32 we need to minimize

$$Q(\beta) = \sum_{i=1}^{n} (Y_i - \beta x_i)^2$$

with respect to β. Differentiating with respect to β

$$Q^{'}(\beta_1) = \sum_{i=1}^{n} x_i(Y_i - \beta_1 x_i) = 0$$

implies that

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2}.$$

Also similar to the proof in class we get

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n}.$$

where

$$e_i = Y_i - \hat{\beta} x_i, i = 1, 2, \ldots, n.$$

(ii) Find the distribution for $\hat{\beta}$ and explain how we can use this result to find a confidence interval for β.

Solution. We have

$$\hat{\beta} = \sum_{i=1}^{n} w_i Y_i, \quad w_i = \frac{x_i}{\sum_{i=1}^{n} x_i^2}.$$

This shows that $\hat{\beta}$ is also in linear combination of independent normal random variables. Therefore

$$E(\hat{\beta}) = \beta, Var(\hat{\beta}) = \sigma^2 \sum_{i=1}^{n} w_i^2 = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}.$$

(iii), (iv). We have $e_i = Y_i - \hat{Y}_i$ for $i = 1, \ldots, n$. This gives $E(e_i) = 0$. On the otehr hand we have

$$Var(e_i) = E(e_i^2) = Var(Y_i - \hat{Y}_i) = Var(Y_i) + Var(\hat{Y}_i) - 2Cov(Y_i, \hat{Y}_i) = \sigma^2 + \sigma^2 \frac{x_i^2}{\sum_{i=1}^n x_i^2} - 2\sigma^2 \frac{x_i^2}{\sum_{i=1}^n x_i^2}.$$

Therefore

$$E(\sum_{i=1}^n e_i^2) = n\sigma^2 + \sigma^2 - 2\sigma^2 = (n-1)\sigma^2.$$

For the prediction interval notice that for a given constant x,

$$Var(Y - \hat{Y}) = Var(Y) + Var(\hat{Y}) - 2Cov(Y, \hat{Y}).$$

We have $Var(Y) = \sigma^2$,

$$Var(\hat{Y}) = Var(\hat{\beta}x) = x^2 \frac{\sigma^2}{\sum_{i=1}^n x_i^2}$$

and $Cov(Y, \hat{Y}) = 0$. Therefore

$$Var(Y - \hat{Y}) = N\left(0, \sigma^2\left(1 + \frac{x^2}{\sum_{i=1}^n x_i^2}\right)\right).$$

Therefore the prediction interval is

$$\hat{\beta}x + t_{n-1,\alpha/2} s \sqrt{1 + \frac{x^2}{\sum_{i=1}^n x_i^2}}.$$

3. (i) Let (X, Y) be a a continuous random point with correlation coefficient ρ such that

$$E(Y|X) = \int_{-\infty}^{\infty} y f(y|x) dy = \beta_0 + \beta_1 X.$$

Prove

$$\beta_0 + \beta_1 X = E(Y) + \rho \frac{\sigma_Y}{\sigma_X}(X - E(X))$$

where σ_X and σ_Y are standard deviations for random variables X and Y, respectively. What if the distribution is of discrete type?

Solution. We have

$$E(Y|X = x) = \int_{-\infty}^{\infty} y f(y|x) dy = \int_{-\infty}^{\infty} y \frac{f(x, y)}{f_1(x)} dy = \beta_0 + \beta_1 x.$$

Therefore

$$\int_{-\infty}^{\infty} yf(x,y)dy = (\beta_0 + \beta_1 x)f_1(x) \qquad (1).$$

Integrating over x we get

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yf(x,y)dydx = \int_{-\infty}^{\infty}(\beta_0 + \beta_1 x)f_1(x)dx.$$

This gives

$$E(Y) = \beta_0 + \beta_1 E(X) \qquad (2).$$

Multiplying (1) by x and integrating over x gives

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)dydx = \int_{-\infty}^{\infty}(\beta_0 + \beta_1 x)xf_1(x)dx.$$

This gives

$$E(XY) = \beta_0 E(X) + \beta_1 E(X^2) \qquad (3).$$

Use the equations (2) and (3) together to solve for β_0 and β_1 and results follow easily. Notice that

$$E(XY) - E(X)E(Y) = Cov(X,Y) = \rho\sigma_x\sigma_Y$$

and

$$\sigma_X^2 + E(X^2) - (E(X))^2.$$

Bonus mark (ii) If $Var(Y|X = x)$ is free from x, then

$$Var(Y|X = x) = \sigma_2^2(1-\rho)^2.$$

Hint. Note that

$$Var(Y) = E(Var(Y|X)) + Var(E(Y|X)).$$

Solution. Since $Var(Y|X)$ is a constant, we have

$$Var(Y) = \sigma_Y^2 = E(Var(Y|X)) + Var(E(Y|X)) = Var(Y|X = x) + Var\left(E(Y) + \rho\frac{\sigma_Y}{\sigma_X}(X - E(X))\right)$$

$$= \rho^2\left(\frac{\sigma_Y}{\sigma_X}\right)^2 Var(X) = \rho^2\left(\frac{\sigma_Y}{\sigma_X}\right)^2\sigma_X^2.$$

Therefore

$$Var(Y|X) = \sigma_Y^2(1-\rho^2).$$