Solution to HW 3

Guanhua FANG

October 9, 2017

2.2

There still could be linear relationship between X and Y. This hypothesis test whether there is a positive relationship between X and Y. It does not exclude the possibility that X and Y are negatively linearly correlated.

2.23

a.

Source of var	ss	df	MS
Regression	3.59	1	3.59
Error	45.82	118	0.39
Total	49.41	119	

Table 1: ANOVA table 2.23

b. MSE estimate σ^2, the variance of noise. MSR estimates the $\sigma^2 + \beta_1^2 \sum (X_i - \bar{X})^2$. When there is no correlation between X and Y which is to say $\beta_1 = 0$, then MSR and MSE estimate the same quantity.

c. The null hypothesis is $\beta_1 = 0$. The alternative is $\beta_1 \neq 0$. We want to use F test to do the testing. The F-statistic value here is $\frac{3.59}{0.39} = 9.21$. The critical cut off here is $F_{0.99}(1, 118)$ which is equal to 6.95. Value 9.21 is larger than 6.95. Hence, we reject the null.

d. The magnitude of absolute reduction is 3.59. The relative reduction is $3.59/49.41 = 7.2\%$. This measure is called R^2.

e. $r = 0.269481$, the sign is positive.

f. In the simple linear regression, r could be see as the measure of correlation between X and Y. It is more easy to compute r. On the other hand, in multiple linear regression, there is no statistical meaning of r any longer. Then, R^2 is better.

2.26

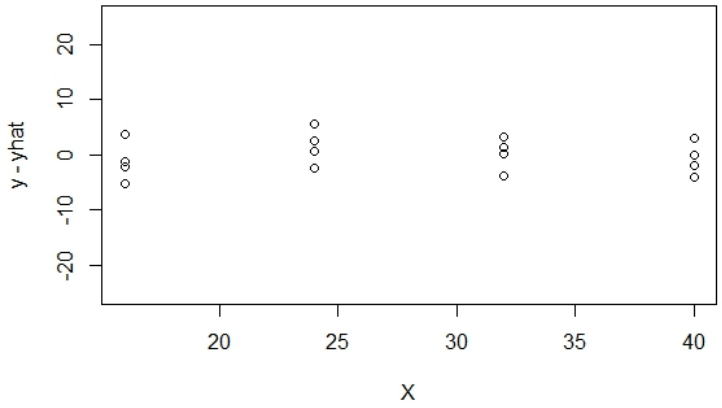
a. See Table 2.

Source of var	ss	df	MS
Regression	5297.51	1	5297.51
Error	146.43	14	10.46
Total	5443.94	15	

Table 2: ANOVA table for Problem 2.26

b. The null is $\beta_1 = 0$, the alternative is $\beta_1 \neq 0$. Hence, the F-statistics value is $5297.51/10.46 = 506.45$. The critical cut-off is $F_{0.99}(1, 14) = 8.86$. Since $506.45 > 8.86$. We reject the null.

c. From the Figure, we can see that the error term is much smaller than $\hat{Y} - \bar{Y}$. This indicates



that SSR is the larger component of $SSTO$ and R^2 should be very close to 1.

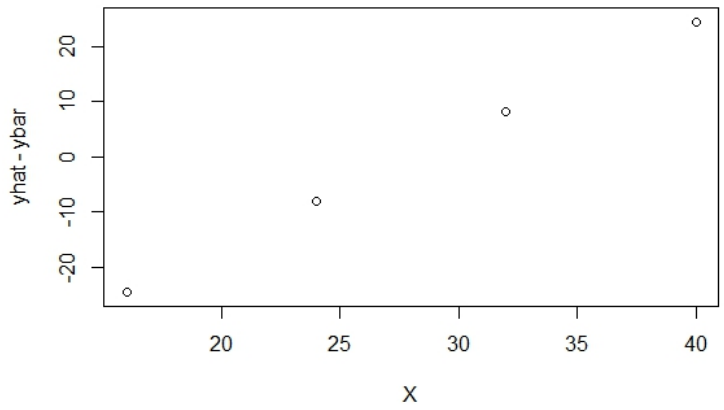
d. $R^2 = 5297.51/5443.94 = 0.97$. $r = 0.986$ and the sign is positive.

2.56

a. $Y_i - \bar{Y} = \beta_1(X_i - \bar{X}) + (e_i - \frac{1}{5}(e_1 + e_2 + \ldots + e_5))$. Then, $(X_i - \bar{X})(Y_i - \bar{Y}) = \beta_1(X_i - \bar{X})^2 + (X_i - \bar{X})(e_i - \frac{1}{5}(e_1 + e_2 + \ldots + e_5))$. Sum them up, we get $\sum(X_i - \bar{X})(Y_i - \bar{Y}) = \beta_1 \sum(X_i - \bar{X})^2 + \sum(X_i - \bar{X})e_i$. Then, $SSR = \beta_1^2 \sum(X_i - \bar{X})^2 + 2\beta_1 \sum(X_i - \bar{X})e_i + \frac{(\sum(X_i - \bar{X})e_i)^2}{\sum(X_i - \bar{X})^2}$. We then compute $\mathrm{E}[SSR] = \beta_1^2 \sum(X_i - \bar{X})^2 + \mathrm{E}\frac{\sum(X_i - X)^2 e_i^2}{\sum(X_i - X)^2}$ by using the fact that $\mathrm{E}e_i = 0$ and e_i, e_j are independent for different i and j. Also, we know that $\mathrm{E}e_i^2 = \sigma^2$. We finally get $\mathrm{E}[SSR] = \beta_1^2 \sum(X_i - \bar{X})^2 + \sigma^2$.

Next, $SSTO = \sum(Y_i - \bar{Y})^2 = \sum \beta_1^2(X_i - \bar{X})^2 + 2\sum \beta_1(X_i - \bar{X})(e_i - \frac{1}{5}(e_1 + e_2 + \ldots + e_5)) + \sum(e_i - \frac{1}{5}(e_1 + e_2 + \ldots + e_5))^2$. Then, we take expectation and get $\mathrm{E}[SSTO] = \sum \beta_1^2(X_i - \bar{X})^2 + 4\sigma^2$. By using the all above displays, we plug the number into those expressions and get. $\mathrm{E}[MSE] = \sigma^2 = 0.36$. $\mathrm{E}[MSR] = \mathrm{E}[SSR] = 9 \cdot 114 + 0.6^2 = 1026.36$.

b. We can calculate the $\mathrm{E}[SSR] = 12.36$ when X = 6,7,8,9,10. It is smaller than the previous setting. Hence, it will have smaller R^2 value. It become worse. When our principle is to estimate the mean response at $X = 8$, the thing of interest is to estimate the confidence interval. Narrower the



confidence interval is, the better it can predict. We know that CI depends on $\sigma^2(Y_h) = \frac{1}{5} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}$. When $X_h = 8$, the $\sigma^2(Y_h)$ is free of data points since $X_h = \bar{X}$. But for other X_h other than 8, the $X's$ in part (a) will lead to narrower interval. Hence, the estimation becomes worse when we use $X = 6, 7, 8, 9, 10$. The reason is two folded. One is that the line is symmetric. Hence, we will lose half information if data given is symmetric around its mean. The second thing is that $X = 6, 7, 8, 9, 10$ has the smaller range, then it will give no information outside range $[6, 10]$. $X = 1, 4, 10, 11, 14$ has a relatively large range.

2.61

We use subscript 1 to denote the regression that Y_1 is regressed on Y_2 and 2 to denote the regression that Y_2 is regressed on Y_1. We know that $SSR_1 = \sum(\hat{Y}_{1i} - \bar{Y}_1)^2$, $SSTO_1 = \sum(Y_{1i} - \bar{Y}_1)^2$. We know that $b_{11} = \frac{\sum(Y_{2i} - \bar{Y}_2)(Y_{1i} - \bar{Y}_1)}{\sum(Y_{2i} - \bar{Y}_2)^2}$ and $b_{10} = \bar{Y}_1 - b_{11}\bar{Y}_2$. Then, we get $\hat{Y}_{1i} = \bar{Y}_1 - b_{11}\bar{Y}_2 + b_{11}Y_{2i}$. Therefore, $\hat{Y}_{1i} - \bar{Y}_1 = b_{11}(Y_{2i} - \bar{Y}_2)$. Then, $SSR_1 = \sum b_{11}^2(Y_{2i} - \bar{Y}_2)^2 = (\frac{\sum(Y_{2i} - \bar{Y}_2)(Y_{1i} - \bar{Y}_1)}{\sum(Y_{2i} - \bar{Y}_2)^2})^2\big(\sum(Y_{2i} - \bar{Y}_2)^2\big) = \frac{\big(\sum(Y_{2i} - \bar{Y}_2)(Y_{1i} - \bar{Y}_1)\big)^2}{\sum(Y_{2i} - \bar{Y}_2)^2}$. Hence, $SSR_1/SSTO_1 = \frac{\big(\sum(Y_{2i} - \bar{Y}_2)(Y_{1i} - \bar{Y}_1)\big)^2}{\sum(Y_{2i} - \bar{Y}_2)^2 \sum(Y_{1i} - \bar{Y}_1)^2}$. We can see that this quantity is symmetric about Y_1 and Y_2. We are safe to say that $SSR_2/SSTO_2$ will have the same value.

2.66

a. $b_0 = 20.27$ and $b_1 = 4.26$. $\hat{Y}_h = 63.34$ and 95 % confidence interval is $(58.69, 67.98)$.

b. See the appendix.

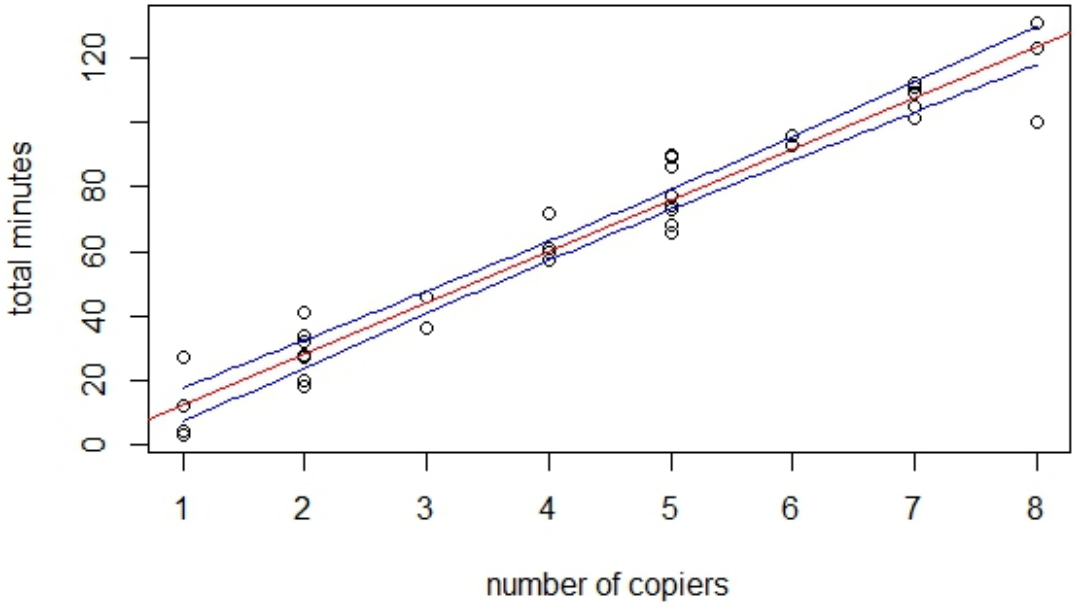
c. The mean is 3.95 and standard deviation is 0.37. The corresponding theoretical value is 4 and 0.39. It is quite close.

d. The proportion is 0.94 which is quite close to 0.95. The result is consistent with theoretical value. (The code is attached in appendix.)

2.68

a. See the corresponding figure.

b. The band is somewhat narrower. There are still a few points outside the 90 percent confidence band which indicates that the true linear regression line could be steeper or more flat. In other words, the line could differ a lot when there are more observed data at X = 1 and 8.



A Code for 2.66

```
set.seed(5205)

e <- 5*rnorm(5);
X <- c(4,8,12,16,20);
y <- 4*X + 20 + e;

fit <- lm(y~X);

coef <- fit$coefficients;
b0 <- coef[1];
b1 <- coef[2];

yh <- b0 + b1*10;

yh + qnorm(0.025)*sqrt(25/5 + 25*4/160)
yh + qnorm(0.975)*sqrt(25/5 + 25*4/160)



#-----------------------------------------------------------------

set.seed(5205)

nrep <- 200;

E <- matrix(rnorm(5*nrep),nrep,5)*5;

ub <- rep(0,nrep);
lb <- rep(0,nrep);
```

```
b1 <- rep(0,nrep);

#-----------------------------------------------------------------

for(i in 1:nrep){
X <- c(4,8,12,16,20);
e <- E[i,];
y <- 4*X + 20 + e;
fit <- lm(y~X);
b1[i] <- fit$coefficients[2];
}

mean(b1)
sd(b1)

5/sqrt(160) # theoretical variance for b1



#-----------------------------------------------------------------

count = 0;
for(i in 1:nrep){
X <- c(4,8,12,16,20);
e <- E[i,];
y <- 4*X + 20 + e;
fit <- lm(y~X);
b0 <- fit$coefficients[1];
b1<- fit$coefficients[2];
```

```
yh <- b0 + b1*10;

lb <- yh + qnorm(0.025)*sqrt(25/5 + 4/160*25)
ub <- yh + qnorm(0.975)*sqrt(25/5 + 4/160*25)

if(ub > 60 && lb < 60){count = count + 1;}

}

count/nrep    # proportion
```