LINEAR REGRESSION: Homework #3

Professor Jingchen Liu

Fan Yang
UNI: fy2232

Problem 1 (2.2)

No, this conclusion does not imply that X and Y have no linear association. This result only tells us that X and Y are negative correlated, which means when X grows, value of Y decreases.

Problem 2 (2.23)

(a)

```
> anova(lm.pr19)
Analysis of Variance Table
Response: da$V1
           Df Sum Sq Mean Sq F value   Pr(>F)
da$V2       1  3.588  3.5878  9.2402 0.002917 **
Residuals 118 45.818  0.3883
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

	DF	SS	MS	F
Regression	1	3.588	3.5878	9.239763
Error	118	45.818	0.3883	
Total	119	49.406		

(b)

$$E(MSR) = \sigma^2 + \beta_1^2 \sum (X_i - \bar{X})$$

So $s^2 + \hat{\beta}_1^2 \sum (X_i - \bar{X})$ is estimated by *MSR*

$$E(MSE) = \sigma^2$$

So s^2 is estimated by *MSE*

When $\hat{\beta}_1 = 0$, *MSR* and *MSE* estimate the same quantity.

(c)

$$H_0: \ \beta_1 = 0$$
$$H_1: \ \beta_1 \neq 1$$
$$F^* = \frac{MSR}{MSE} = 9.239763$$
if $F^* \leq F(0.99, 1, 118)$, then conclude H_0
else $F^* > F(0.99, 1, 118)$, reject H_0

While $F(0.99, 1, 118) = 6.855 \ < \ F^*$, so we can reject H_0 and conclude $\beta_1 \neq 0$

(d)

While $F(0.99, 1, 118) = 6.855 \ < \ F^*$, so we can reject H_0 and conclude $\beta_1 \neq 0$

pass

(e)

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{45.818}{49.406} = 0.0726$$
$$r = \sqrt{R^2} = \sqrt{0.0726} = +0.2695$$

(f)

I think R^2 has the more clear-cut operational interpretation. Because R^2 equals to Explained variation divided by Total variation, which represents the percentage of variation can be explained by our linear model.

Problem 3 (2.26)

(a)

```
> anova(model22)
Analysis of Variance Table
Response: d22$V1
          Df Sum Sq Mean Sq F value    Pr(>F)
d22$V2     1 5297.5  5297.5  506.51 2.159e-12 ***
Residuals 14  146.4    10.5
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

	DF	SS	MS	F
Regression	1	5297.5	5297.5	506.51
Error	14	146.4	10.5	
Total	15	5443.9		

(b)

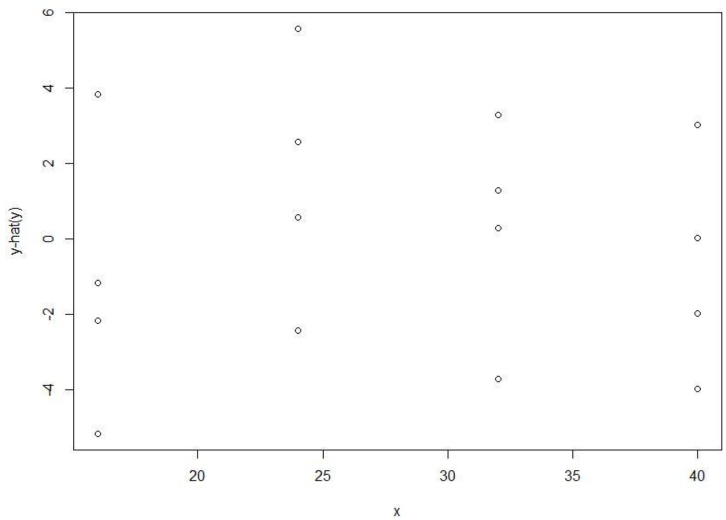
$$H_0: \ \beta_1 = 0$$
$$H_1: \ \beta_1 \neq 1$$
$$F^* = \frac{MSR}{MSE} = 506.51$$

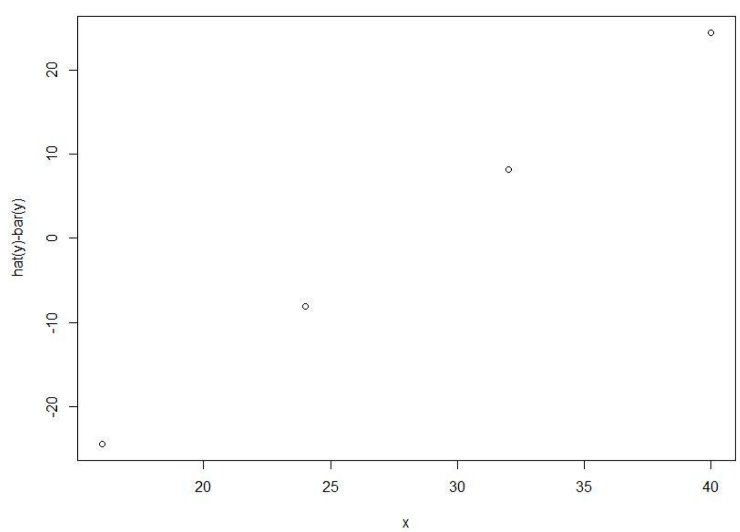
if $F^* \leq F(0.99, 1, 14)$, then conclude H_0

else $F^* > F(0.99, 1, 118)$, reject H_0

While $F(0.99, 1, 14) = 8.861593 \ < \ F^*$, so we can reject H_0 and conclude $\beta_1 \neq 0$

(c)





From the graph we can say that SSR appear to be the larger component of SST. While $R^2 = \dfrac{SSR}{SST}$ so R^2 is large.

Problem 4 (2.56)

(a)

$$\begin{aligned}\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X} = 5 + 3 \times 8 = 29 \\ E(MSE) = \sigma^2 \\ = 0.6^2 = 0.36 \\ E(MSR) &= \sigma^2 + \hat{\beta}_1^2 \sum (X_i - \bar{X})^2 \\ &= 0.36 + 9 \times \sum (X_i - 8)^2 \\ &= 1026.36\end{aligned}$$

(b)

Denote E as A wins, F as they played 2 games and game ends.

$$\begin{aligned}P(E) &= P(E \cap F) + P(E \cap F^c) \\ &= P(E \cap F) + P(E|F^c)P(F^c) \\ &= (p * p) + P(E) * 2p(1-p) \\ &= p^2 + P(E) * 2p(1-p)\end{aligned}$$

$$\implies (1 - 2p + 2p^2)P(E) = p^2$$
$$\implies P(E) = \frac{p^2}{(1 - 2p + 2p^2)}$$

Problem 5 (2.61)

$$\begin{aligned}\frac{SSR}{SST} &= \frac{b_1^2 \sum (X_i - \bar{X})^2}{\sum (Y_i - \bar{Y})^2} \\ &= \frac{\left(\frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}\right)^2 \sum (X_i - \bar{X})^2}{\sum (Y_i - \bar{Y})^2} \\ &= \frac{\left(\sum (X_i - \bar{X})(Y_i - \bar{Y})\right)^2}{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}\end{aligned}$$

As for this fraction, the denominator and numerator each has same expression for X and Y, which means X and Y have symmetric expressions. Therefore, the ratio is the same whether $X(orY_1)$ is regressed on $Y(orY_2)$ or $Y(orY_2)$ is regressed on $X(orY_1)$.

Problem 6 (2.66)

(a)

```
> e=rnorm(5, mean = 0, sd = 5)
[1] -4.9633070  6.4244096 -0.4761277 -2.8104307 -0.4428723
> X=c(4,8,12,16,20)
> Y=20+4*X+e
> Y
[1] 31.03669 58.42441 67.52387 81.18957 99.55713


> model66 <- lm(Y~X)
> summary(model66)
Call:
lm(formula = Y ~ X)
Residuals:
       1        2        3        4        5
-4.54844  6.85868 -0.02246 -2.33737  0.04959
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  19.6045     5.1806   3.784  0.03235 *
X             3.9952     0.3905  10.231  0.00199 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

So $b_0 = 19.6045$ and $b_1 = 3.9952$

When $X_h = 10,\ Y_h = 19.6045 + 3.9952 \times 10 = 59.55603$

The 95 percent confidence interval is

$$\hat{Y}_h \pm t(1-\alpha/2,\ n-2)s_{\{\hat{Y}_h\}}$$

which is

$$59.5565 \pm 3.182446 \times 2.343011$$
$$[\ 52.09952,\ 67.01254\ ]$$

(b)

```
for (i in 1:200){
  e=rnorm(5, mean = 0, sd = 5)
  X=c(4,8,12,16,20)
  Y=20+4*X+e
  model66 <- lm(Y~X)
  B1[i]=model66$coefficients[2]
}
```

(c)

```
> mean(B1)
[1] 3.967998
> sd(B1)
[1] 0.3660616
```

$$\sigma^2(b_1) = \frac{\sigma^2}{\sum(X_i - \bar{X})^2}$$

```
> sqrt(25/sum((X-mean(X))^2))
[1] 0.3952847
```

So, the theoretical expectation of $\sigma(b_1)$ should be 0.3952847. The result differs from the theoretical expectation a little.

PASS

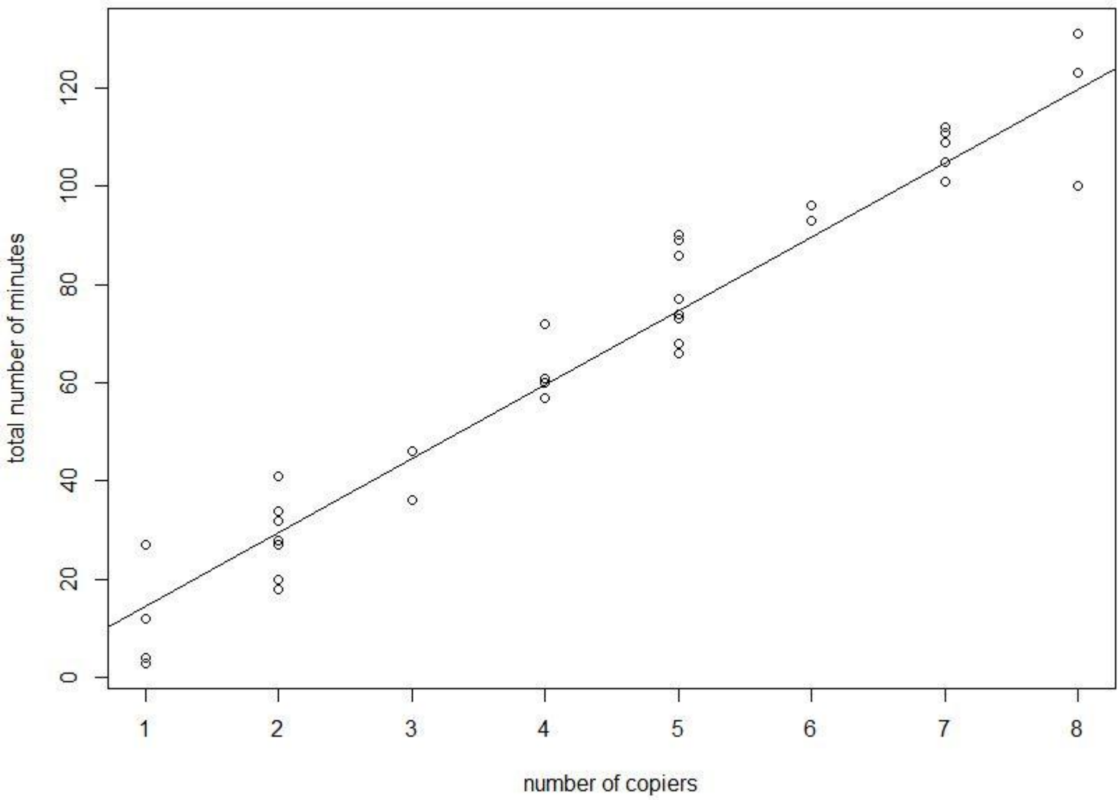
(d)

```
> for (i in 1:200){
+   e=rnorm(5, mean = 0, sd = 5)
+   X=c(4,8,12,16,20)
+   Y=20+4*X+e
+   model66 <- lm(Y~X)
+   newdata=data.frame(X=10)
+   newy=predict(model66, newdata)
+   Bool66[i]<-(newy%in%predict(model66,newdata,interval="confidence"))*1
+ }
> sum(Bool66) / length(Bool66)
[1] 1
```

From the results above, we find 100% of the 200 confidence intervals include $E\{Y_h\}$. This result is consistent with theoretical expectations.

Problem 7 (2.68)

(a)



(b)

$$
\begin{aligned}
P(\frac{1}{2} < X < \frac{3}{2}) &= P(\frac{1}{2} < X < 1) + P(1 \le X < \frac{3}{2}) \\
&= \lim_{y \to 1^-} F(y) - F(\frac{1}{2}) + F(\frac{3}{2}) - F(1) \\
&= (\frac{1}{4} - \frac{1/2}{4}) + (\frac{1}{2} + \frac{3/2 - 1}{4} - \frac{1}{2}) \\
&= \frac{1}{4}
\end{aligned}
$$