Homework 6

UNI: qy2205 Eric Yuan

3.14

(a)

$H_0: Y_{ij} = \beta_0 + \beta_1 X_j + \epsilon_{ij}, reduced\ model$

$\mathrm{H}_0: Y_{ij} = \mu_{ij} + \epsilon_{ij}, full\ model$

$$SSR(R) = \sum_{j=1}^{4} \sum_{i} (\beta_0 + \beta_1 X_j - Y_{ij})^2$$

$$SSR(F) = \sum_{j=1}^{4} \sum_{i} (Y_{ij} - \overline{Y}_j)^2$$

$$F^* = \frac{SSE(R) - SSE(F)}{df_r - df_F} \div \frac{SSE(F)}{df_F}$$

$$= \frac{SSE - SSPE}{c - 2} \div \frac{SSPE}{n - c}$$

$$= \frac{146.425 - 128.75}{4 - 2} \div \frac{128.75}{16 - 4}$$

$$= 0.8237$$

Since $F^* \leq \mathrm{F}(0.99, 3, 11) = 6.9266$, so the conclusion is H_0.

```
miu <- tapply(data$V1, data$V2, mean)
se <- 0
for (i in as.numeric(names(miu))){
     se <- se + sum((data[data$V2 == i, 1] - miu[which(names(miu) == i)])^2)
}
df <- nrow(data1) - length(miu)
yfit <- lm(V1 ~ V2, data = data)$fitted.values
sse <- sum((data1[, 1] - yfit) ^ 2)
sslf <- sse - se
Fstat <- (sslf/(length(miu)-2))/(se/df)
```

```
F-stat 0.82369
```

```
6.926608
```

(b)

There is no substantial advantage or disadvantage.

(c)

The test indicate that the linear regression line is invalid and transformation to data may generate better results.

7.7

(a)

```
y.mean <- mean(data$Y)
SST <- sum((Y-y.mean)^2)

L <- lm(Y~X1+X2+X3+X4)
SSE <- sum((L$residuals)^2)

SSR <- SST-SSE

SSR4 <- SST-sum(((lm(Y~X4)$residuals))^2)

SSE4 <- sum(((lm(Y~X4))$residuals)^2)
SSE14 <- sum(((lm(Y~X1+X4))$residuals)^2)
SSR1_4 <- SSE4-SSE14

SSE14 <- sum(((lm(Y~X1+X4))$residuals)^2)
SSE142 <- sum(((lm(Y~X1+X4+X2))$residuals)^2)
SSR2_14 <- SSE14-SSE142

SSE142 <- sum(((lm(Y~X1+X4+X2))$residuals)^2)
SSE1423 <- sum(((lm(Y~X1+X4+X2+X3))$residuals)^2)
SSR3_142 <- SSE142-SSE1423

SS <- c(SSR,SSR4,SSR1_4,SSR2_14,SSR3_142,SSE,SST)
df <- c(4,1,1,1,1,76,80)
MS <- SS/df
MS[7] <- NA
df <- data.frame(SS=SS,df=df,MS=MS)
rownames(df) <- c("Regression","X4","X1|X4","X2|X1,X4","X3|X1,X2,X4","Error","Total")
df
```

```
##                     SS df         MS
## Regression  138.3269061  4 34.5817265
## X4           67.7750980  1 67.7750980
## X1|X4        42.2745683  1 42.2745683
## X2|X1,X4     27.8574935  1 27.8574935
## X3|X1,X2,X4   0.4197463  1  0.4197463
## Error        98.2305939 76  1.2925078
## Total       236.5575000 80         NA
```

(b)

H_0: $\beta_3 = 0$

$\mathrm{H_A}$: $\beta_3 \neq 0$

$$F^* = \frac{SSE(R) - SSE(F)}{df_r - df_F} \div \frac{SSE(F)}{df_F}$$

$$= \frac{SSR(X_3|X_1, X_2, X_4)}{(n-4)-(n-5)} \div \frac{SSE(X_1, X_2, X_3, X_4)}{n-5}$$

$$= \frac{0.4197}{1} \div \frac{98.2306}{76}$$

$$= 0.3247$$

Since $F^* \leq \mathrm{F}(0.99, 1{,}76) = 6.9806$, so the conclusion is H_0.

```
> p_value <- 1 - pf(0.3247,1,76)
> p_value
[1] 0.5704773
```

7.10

```
sser <- sum((data$V1 + 0.1*data$V2 - 0.4*data$V3 - lm(V1+0.1*V2-0.4*V3 ~
               V4 + V5, data = data)$fitted.values)^2)
Fstat <- ((sser - sse)/2)/(sse/76)
cat(Fstat)
cat(qf(0.99, 2, 76))
```

$F^* = 4.6076 \leq F(0.99; 2, 76) = 4.8958$

$H_0: \beta_1 = -0.1$ and $\beta_2 = 0.4$

$\mathrm{H_A}: \beta_1 \neq -0.1$ or $\beta_2 \neq 0.4$

Based on the calculation result, we can conclude H_0

7.16

(a)

```
data16 <- read.table("http://users.stat.ufl.edu/~rrandles/sta4210/Rclassnote
n <- nrow(data16)
y <- (data16$V1 - mean(data16$V1))/sd(data16$V1)/(n-1)^.5
x1 <- (data16$V2 - mean(data16$V2))/sd(data16$V2)/(n-1)^.5
x2 <- (data16$V3 - mean(data16$V3))/sd(data16$V3)/(n-1)^.5
coef(lm(y ~ x1  + x2))
beta1s <- coef(lm(y ~ x1 + x2))[2]
beta1s
beta1 <- beta1s*sd(data16$V1)/sd(data16$V2)
beta1
coef(lm(V1~V2+V3, data = data16))[2]
```

```
> coef(lm(y     ~      x1      +      x2))
  (Intercept)            x1            x2
-1.238444e-17  8.923929e-01  3.945807e-01
> beta1s <-      coef(lm(y ~ x1 + x2))[2]
> beta1s
       x1
0.8923929
> beta1 <- beta1s*sd(data16$V1)/sd(data16$V2)
> beta1
   x1
4.425
> coef(lm(V1~V2+V3, data = data16))[2]
   V2
4.425
```

$$\mathrm{Y}^* = -1.238 \times 10^{-17} + 0.892\mathrm{X}_1^* + 0.395\mathrm{X}_2^*$$

(b)

When X_1^* increase 1-unit, Y^* will increase 0.892 if X_2^* is constant.

(c)

Yes, they are the same.

7.24

```
> coef(lm(V1 ~ V2, data = data16))
(Intercept)          V2
     50.775       4.425
```

```
> ssr1<-sum((mean(data16$V1)-lm(V1~V2,  data=data16)$fitted.values)^2)
> ssr2<-sum((mean(data16$V1)-lm(V1~V3,  data=data16)$fitted.values)^2)
> ssr12<-sum((mean(data16$V1)-lm(V1~V2+V3,data=data16)$fitted.values)^2)
> ssr12-ssr2
[1] 1566.45

> ssr1
[1] 1566.45
> ##    [1]     1566.45
> cor(data16)
          V1        V2        V3
V1 1.0000000 0.8923929 0.3945807
V2 0.8923929 1.0000000 0.0000000
V3 0.3945807 0.0000000 1.0000000
```

The fitted line is:

$Y = 50.775 + 4.425X_1$

The coefficient of first order regression function with Y is the same as the coefficient obtained before. So, they are the same.

Yes, $\mathrm{SSR}(\mathrm{X}_1) = \mathrm{SSR}(\mathrm{X}_1|\mathrm{X}_2)$

It shows that X_1 and X_2 are independent.

7.37

(a)

```
data737 <- read.table("http://users.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/

lm1 <- lm(V8~V5+V16, data=data737)
lm2 <- lm(V8~V5+V16+V4, data=data737)
lm3 <- lm(V8~V5+V16+V7, data=data737)
lm4 <- lm(V8~V5+V16+V9, data=data737)
lm5 <- lm(V8~V5+V16+V10, data=data737)

sse1 <- sum((data737$V8 - lm1$fitted.values)^2)
sse2 <- sum((data737$V8 - lm2$fitted.values)^2)
sse3 <- sum((data737$V8 - lm3$fitted.values)^2)
sse4 <- sum((data737$V8 - lm4$fitted.values)^2)
sse5 <- sum((data737$V8 - lm5$fitted.values)^2)

Rsquare12 <- (sse1  - sse2)/sse1; Rsquare12
Rsquare13 <- (sse1  - sse3)/sse1; Rsquare13
Rsquare14 <- (sse1  - sse4)/sse1; Rsquare14
Rsquare15 <- (sse1  - sse5)/sse1; Rsquare15

> Rsquare12 <- (sse1 - sse2)/sse1;     Rsquare12
[1] 0.02882495
> Rsquare13 <- (sse1 - sse3)/sse1;     Rsquare13
[1] 0.003842367
> Rsquare14    <- (sse1 - sse4)/sse1;  Rsquare14
[1] 0.5538182
> Rsquare15    <- (sse1 - sse5)/sse1;  Rsquare15
[1] 0.007323408
```

(b)

Number of hospital beds is the best one, It has the largest coefficients of partial determination.

(c)

```
Fstat <- (sse1 - sse4)/(sse4/(nrow(data737)-4));  Fstat
qf(0.99,1,436)
(sse1-sse2)/(sse2/(nrow(data737)-4))
(sse1-sse3)/(sse3/(nrow(data737)-4))
(sse1-sse4)/(sse4/(nrow(data737)-4))
(sse1-sse5)/(sse5/(nrow(data737)-4))
```

```
> Fstat <- (sse1 - sse4)/(sse4/(nrow(data737)-4));      Fstat
[1] 541.1801
> qf(0.99,1,436)
[1] 6.693358
> (sse1-sse2)/(sse2/(nrow(data737)-4))
[1] 12.94069
> (sse1-sse3)/(sse3/(nrow(data737)-4))
[1] 1.681734
> (sse1-sse4)/(sse4/(nrow(data737)-4))
[1] 541.1801
> (sse1-sse5)/(sse5/(nrow(data737)-4))
[1] 3.216562
```

$F^* = 541.1801 \leq F(0.99; 1{,}437) = 6.693358$

So, we conclude $\mathrm{H_A}$, $\mathrm{X_5}$ is helpful to the model.

No, F-statistics for other three variables are not as large as it for number of hospita beds since their coefficients of partial determination are smaller.