hw6

Yuhao Wang, yw3204

11/3/2018

3.14

100-2-2-1=95

a.

```
# load data
pla <- read.table("CH01PR22.txt")
names(pla) <- c("Y", "X")

# calculate SSE(F), page 122
library(plyr)
f1 <- function(df) {
  return(sum((df$Y - mean(df$Y))^2))
}
sse_f <- sum(as.numeric(daply(pla, "X", f1)))

# calculate SSE(R)
lm1 <- lm(Y~X, pla)
b <- coef(lm1)

f2 <- function(df) {
  return(sum((df$Y - b[1] - b[2]*df$X)^2))
}
sse_r <- sum(as.numeric(daply(pla, "X", f2)))

# calculate degree of freedom and F-value
df_r <- nrow(pla)-length(b)
df_f <- nrow(pla)-length(unique(pla$X))
F_star <- (sse_r - sse_f)/(df_r-df_f) * df_f / sse_f

alpha <- 0.01
F_ref <- qf(1-alpha, df_r-df_f, df_f)
F_star
```

```
## [1] 0.8236893
```

```
F_ref
```

```
## [1] 6.926608
```

The alternative is $\mathbb{E}(Y) \neq \beta_0 + \beta_1 * X$. The decision rule is when F^* is greater than $F(0.99, 2, 12)$, we accept the alternative. Otherwise, we accept the null. And since under this occation F^* is smaller than $F(0.99, 2, 12)$, we conclude H_0.

b.

Since the distribution of the estimate of the response variable is somewhat influenced by the number of observations at each predictor level, holding the number equal indicates that the estimate distribution is at

the same precision level.

-2: The advantage is that equal number of observations under each X level can make error smaller at each level. Also, we can tell whether there is an outlier at each X level. The estimation could become more robust.
-2: The disadvantage is that in the variance of prediction error in linear regression is not constant through all X.

c.

No, it doesn't. We may transform the data or including other variables.

7.7

-1: The linear assumption between response and predictor is appropriate from the F-test.

a.

```
cp <- read.table("CH06PR18.txt")
names(cp) <- c("Y", "X1", "X2", "X3", "X4")

lm1 <- lm(Y~X4, cp)
lm2 <- lm(Y~X1+X4, cp)
lm3 <- lm(Y~X1+X2+X4, cp)
lm4 <- lm(Y~X1+X2+X3+X4, cp)
#anova(lm1)
#anova(lm2)
#anova(lm3)
#anova(lm4)
```

According to the anova analysis in R, we build the table below.

source of variation	SS	df	MS
$SSR(X_1, X_2, X_3, X_4)$	138.33	4	34.58
$SSR(X_4)$	67.78	1	67.78
$SSR(X_1\|X_4)$	42.27	1	42.27
$SSR(X_2\|X_1, X_4)$	27.86	1	27.86
$SSR(X_3\|X_1, X_2, X_4)$	0.42	1	0.42
$SSE(X_1, X_2, X_3, X_4)$	98.23	76	1.29
SST	236.56	80	2.96

b.

```
F_3 <- 0.42 / 1.29
F_3_ref <- qf(0.99, 1, 76)

F_3
```

```
## [1] 0.3255814
```

```
F_3_ref
```

```
## [1] 6.980578
```

```
# p-value
1 - pf(F_3, 1, 76)
```

```
## [1] 0.5699558
```

The alternative is $H_A : \beta_3 \neq 0$. The decision rule is when when the calculated F-statistic is greater than

$F(0.99, 1, 76)$, we reject the null. And in this occasion, we accept the null based on the result above. The P-value is 0.57.

7.10

```
lm5 <- lm(Y + 0.1*X1-0.4*X2 ~ X3+X4, cp)
anova(lm5)
```

```
## Analysis of Variance Table
##
## Response: Y + 0.1 * X1 - 0.4 * X2
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X3         1   9.205   9.205  6.5187   0.01263 *
## X4         1  31.872  31.872 22.5713 9.058e-06 ***
## Residuals 78 110.141   1.412
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
F_1_2 <- ((110.141 - 98.23)/2) / 1.29
F_1_2_ref <- qf(0.99, 2, 76)

F_1_2
```

```
## [1] 4.616667
```

```
F_1_2_ref
```

```
## [1] 4.89584
```

```
#p-value
#1-pf(F_1_2, 2, 76)
```

The alternative is $H_A : \beta_1 \neq -.1$ or $\beta_2 \neq .4$. The full model is $Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_3 + \beta_4 * X_4 + \epsilon$ and the reduced model is $Y + 0.1 * X_1 - 0.4 * X_2 = \beta_0 + \beta_3 * X_3 + \beta_4 * X_4 + \epsilon$. The decision rule is if the calculated F-statistic is greater than $F(0.99, 2, 76)$, we reject the null and based on the result above, we conclude the null.

7.16

a.

```
bp <- read.table("CH06PR05.txt")
names(bp) <- c("Y", "X1", "X2")

s_bp <- apply(bp, 2, function(x) (x-mean(x))/sd(x))

s_lm <- lm(Y~X1+X2, as.data.frame(s_bp))
s_lm
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2, data = as.data.frame(s_bp))
##
## Coefficients:
```

```
## (Intercept)           X1           X2
##  -2.220e-16    8.924e-01    3.946e-01
```

The standardised model is $Y = -2.22 * 10^{-16} + 0.89 * X_1 + 0.39 * X_2$.

b.

The standardised coefficient b_1^* stands for the standardised response changes 0.89 unit as the standardised X_1 changes 1 unit.

c.

```
b <- as.numeric(coef(s_lm)[-1])
b <- b * sd(bp$Y) / c(sd(bp$X1), sd(bp$X2))
b <- c(mean(bp$Y) - b[1]*mean(bp$X1) - b[2]*mean(bp$X2), b)
b
```

```
## [1] 37.650  4.425  4.375
```

The re-transformed coefficients are presented above and they are the same to the result from Problem 6.5.

7.24

a.

```
sp_lm <- lm(Y~X1, bp)
sp_lm
```

```
##
## Call:
## lm(formula = Y ~ X1, data = bp)
##
## Coefficients:
## (Intercept)           X1
##      50.775        4.425
```

The fitted line is $Y = 50.78 + 4.43 * X_1$.

b.

They are the same.

c.

```
#anova(sp_lm)
#anova(lm(Y~X1+X2, bp))
#anova(lm(Y~X2, bp))
```

Here, $SSR(X_1)$ is 1566.45 and $SSR(X_1|X_2)$ is 1566.45. Obviously, they are the same.

d.

We find if predictors are uncorrelated, adding or removing a subset of these predictors will not influence the others, such as the coefficients and SSR.

7.37

```
cdi <- read.table("APPENCO2.txt")
cdi <- cdi[, c(8, 5, 16, 4, 7, 9, 10)]
names(cdi) <- c("Y", "X1", "X2", "X3", "X4", "X5", "X6")

#anova(lm(Y~X1+X2, cdi))
#sse: 140967081
#anova(lm(Y~X1+X2+X3, cdi))
#sse: 136903711
#anova(lm(Y~X1+X2+X4, cdi))
#sse: 140425434
#anova(lm(Y~X1+X2+X5, cdi))
#sse: 62896949
#anova(lm(Y~X1+X2+X6, cdi))
#sse: 139934722
```

Based on the above anova analysis and the formula $\frac{SSE(X_1,X_2)-SSE(X_1,X_2,X_k)}{SSE(X_1,X_2)}, k = 3, 4, 5, 6$, we find the coefficient of partial determination respectively: 0.0288, 0.0038, 0.5538 and 0.0073.

b.

Since X_5 has the highest coefficients of partial determination, it is the best. The extra sum of squares is 78070132 and is larger than that of others.

c.

```
#nrow(cdi)
F_cdi <- (140967081-62896949) / (62896949/(440-4))
F_cdi_ref <- qf(0.99, 1, 440-4)

F_cdi
```

```
## [1] 541.1801
```

```
F_cdi_ref
```

```
## [1] 6.693358
```

The alternative is $\beta_5 \neq 0$. The rule is when the calculated F-statistic is greater than $F(0.99, 1, 396)$, we reject the null. And according to the result, we accept the alternative that we should include X_5. Unfortunately, none is as large as X_5 here.