Spring 2023 STAT 707 Chapter 7 Homework

Reina Li

Chapter 7: 7.5, 7.6, 7.9, 7.18, 7.30, 7.34

7.5. Refer to Patient satisfaction Problem 6.15.

```
# Import data set
df <- read.table("CH06PR15.txt")
# Set column names
colnames(df) <- c("Y", "X1", "X2", "X3")
```

a. Obtain the analysis of variance table that decomposes the regression sum of squares into extra sum of squares associated with X_2; with X_1, given X_2; and with X_3, given X_2 and X_1.

```
# Fit regression model
model <- lm(Y ~ X2 + X1 + X3, data = df)
# ANOVA table
anova(model)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X2         1 4860.3  4860.3 48.0439 1.822e-08 ***
## X1         1 3896.0  3896.0 38.5126 2.008e-07 ***
## X3         1  364.2   364.2  3.5997   0.06468 .
## Residuals 42 4248.8   101.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# SSR(X1,X2,X3)
SSR <- sum(anova(model)[1:3,2]); SSR
```

```
## [1] 9120.464
```

```
# MSR(X1,X2,X3)
MSR <- SSR / 3; MSR
```

```
## [1] 3040.155
```

```r
# SSE(X1,X2,X3)
SSE <- anova(model)[4,2]; SSE
```

```
## [1] 4248.841
```

```r
# MSE(X1,X2,X3)
MSE <- anova(model)[4,3]; MSE
```

```
## [1] 101.1629
```

b. Test whether X_3 can be dropped from the regression model given that X_1 and X_2 are retained. Use the F^* test statistic and level of significance .025. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

```r
# Full model : lm(Y ~ X2 + X1 + X3, data = df)
# Fit reduced model without X3: Y_i = B0 + B1*X_i1 + B2*X_i2 + e_i
reduced_model <- lm(Y ~ X2 + X1, data = df)
anova(reduced_model, model)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X2 + X1
## Model 2: Y ~ X2 + X1 + X3
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     43 4613.0
## 2     42 4248.8  1    364.16 3.5997 0.06468 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Critical value F
alpha <- 0.025
qf(1-alpha, 1, 42)
```

```
## [1] 5.403859
```

Alternatives:
$H_0 : \beta_3 = 0$
$H_a : \beta_3 \neq 0$

Test statistic:
$F^* = 3.5997$
$F(0.975,1,42) = 5.403859$

Decision rule:
If $F^* \leq F(0.975,1,42)$, conclude H_0
If $F^* > F(0.975,1,42)$, conclude H_a

Conclusion:
$F^* \leq F(0.975,1,42)$
Conclude H_0. Fail to reject the null hypothesis H_0. X_3 can be dropped from the regression model that already contains X_1 and X_2.

P-value: 0.06468

7.6. Refer to Patient satisfaction Problem 6.15. Test whether both X_2 and X_3 can be dropped from the regression model given that X_1 is retained. Use $\alpha = .025$. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

```
# Full model : lm(Y ~ X2 + X1 + X3, data = df)
# Fit reduced model without X2 and X3: Y_i = B0 + B1*X_i1 + e_i
reduced_model <- lm(Y ~ X1, data = df)
anova(reduced_model, model)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X1
## Model 2: Y ~ X2 + X1 + X3
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     44 5093.9
## 2     42 4248.8  2    845.07 4.1768 0.02216 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Critical value F
alpha <- 0.025
qf(1-alpha, 2, 42)
```

```
## [1] 4.03271
```

Alternatives:
$H_0 : \beta_2 = \beta_3 = 0$
H_a : not both β_2 and β_3 equal zero

Test statistic:
$F^* = 4.1768$
$F(0.975,2,42) = 4.03271$

Decision rule:
If $F^* \leq F(0.975,2,42)$, conclude H_0
If $F^* > F(0.975,2,42)$, conclude H_a

Conclusion:
$F^* > F(0.975,2,42)$
Conclude H_a. Reject the null hypothesis H_0. Don't drop X_2 and X_3 from the regression model that already contains X_1.

P-value: 0.02216

7.9. Refer to Patient satisfaction Problem 6.15. Test whether $\beta_1 = -1.0$ and $\beta_2 = 0$; use $\alpha = .025$. State the alternatives, full and reduced models, decision rule, and conclusion.

```r
# Full model : lm(Y ~ X2 + X1 + X3, data = df)
# Fit reduced model: Y_i = B0 - X_i1 + B3*X_i3 + e_i
reduced_model <- lm(Y + X1 ~ X3, data = df)
anova(reduced_model, model)
```

```
## Analysis of Variance Table
##
## Response: Y + X1
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X3         1 1636.3 1636.26   16.26 0.0002162 ***
## Residuals 44 4427.7  100.63
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# SSE(R)
SSE_R <- 4427.7; SSE_R
```

```
## [1] 4427.7
```

```r
# df(R)
df_R <- 44; df_R
```

```
## [1] 44
```

```r
# SSE(F)
SSE_F <- SSE; SSE_F
```

```
## [1] 4248.841
```

```r
# df(F)
df_F <- 42; df_F
```

```
## [1] 42
```

```r
# Test statistic F* = (SSR(R)-SSE(F) / df(R)-df(F)) / SSE(F) / df(F)
test_stat <- ((SSE_R - SSE_F) / (df_R - df_F)) / (SSE_F / df_F); test_stat
```

```
## [1] 0.8840166
```

```r
# Critical value F
alpha <- 0.025
qf(1-alpha, 2, 42)
```

```
## [1] 4.03271
```

Alternatives:
$H_0 : \beta_1 = -1.0, \beta_2 = 0$
H_a : not both equalities in H_0 hold

Test statistic:
$F^* = 0.8840166$
$F(0.975,2,42) = 4.03271$

Decision rule:
If $F^* \leq F(0.975,2,42)$, conclude H_0
If $F^* > F(0.975,2,42)$, conclude H_a

Conclusion:
$F^* \leq F(0.975,2,42)$
Conclude H_0. Fail to reject the null hypothesis H_0.

7.18. Refer to Patient satisfaction Problem 6.15.

a. Transform the variables by means of the correlation transformation (7.44) and fit the standardized regression model (7.45).

```
# Create data frame for transformed data
transformed_df <- data.frame(matrix(ncol = 4, nrow = 46))
colnames(transformed_df) <- c("Y_star", "X1_star", "X2_star", "X3_star")
# Standardize variables
y_bar <- mean(df$Y)
s_y <- sd(df$Y)
x_1bar <- mean(df$X1)
s_x1 <- sd(df$X1)
x_2bar <- mean(df$X2)
s_x2 <- sd(df$X2)
x_3bar <- mean(df$X3)
s_x3 <- sd(df$X3)
for (i in 1:46) {
  transformed_df[i,1] <- (1/sqrt(46-1)) * ((df$Y[i]-y_bar)/s_y)
  transformed_df[i,2] <- (1/sqrt(46-1)) * ((df$X1[i]-x_1bar)/s_x1)
  transformed_df[i,3] <- (1/sqrt(46-1)) * ((df$X2[i]-x_2bar)/s_x2)
  transformed_df[i,4] <- (1/sqrt(46-1)) * ((df$X3[i]-x_3bar)/s_x3)
}
# Fit standardized regression model
sd_model <- lm(Y_star ~ X1_star + X2_star + X3_star, data = transformed_df)
summary(sd_model)
```

```
##
## Call:
## lm(formula = Y_star ~ X1_star + X2_star + X3_star, data = transformed_df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.158723 -0.055550  0.004493  0.072402  0.148411
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.485e-17  1.283e-02   0.000   1.0000
## X1_star     -5.907e-01  1.111e-01  -5.315 3.81e-06 ***
## X2_star     -1.106e-01  1.231e-01  -0.898   0.3741
## X3_star     -2.339e-01  1.233e-01  -1.897   0.0647 .
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08699 on 42 degrees of freedom
## Multiple R-squared:  0.6822, Adjusted R-squared:  0.6595
## F-statistic: 30.05 on 3 and 42 DF,  p-value: 1.542e-10
```

The estimated standardized regression function is: $\hat{Y}^* =$ -0.5907X_1^* + -0.1106X_2^* + -0.2339X_3^*

b. Calculate the coefficients of determination between all pairs of predictor variables. Do these indicate that it is meaningful here to consider the standardized regression coefficients as indicating the effect of one predictor variable when the others are held constant?

```r
# Square the correlation matrix
rsq_matrix <- (cor(transformed_df))^2; rsq_matrix
```

```
##            Y_star   X1_star   X2_star   X3_star
## Y_star  1.0000000 0.6189843 0.3635387 0.4154975
## X1_star 0.6189843 1.0000000 0.3225677 0.3245324
## X2_star 0.3635387 0.3225677 1.0000000 0.4496087
## X3_star 0.4154975 0.3245324 0.4496087 1.0000000
```

```r
# Coefficient of determination between X1* and X2*
rsq_matrix[2,3]
```

```
## [1] 0.3225677
```

```r
# Coefficient of determination between X1* and X3*
rsq_matrix[2,4]
```

```
## [1] 0.3245324
```

```r
# Coefficient of determination between X2* and X3*
rsq_matrix[3,4]
```

```
## [1] 0.4496087
```

Yes, these do indicate that it is meaningful here to consider the standardized regression coefficients as indicating the effect of one predictor variable when the others are held constant because the predictor variables are not highly correlated.

c. Transform the estimated standardized regression coefficients by means of (7.53) back to the ones for the fitted regression model in the original variables. Verify that they are the same as the ones obtained in Problem 6.15c.

```r
cor1 <- cor(transformed_df)[2:4,2:4];
cor2 <- cor(transformed_df)[2:4,1];
b_star <- solve(cor1) %*% cor2; b_star
```

```
##                [,1]
## X1_star -0.5906664
## X2_star -0.1106149
## X3_star -0.2339312
```

```r
# b1*
b1_star <- b_star[1]; b1_star
```

```
## [1] -0.5906664
```

```r
# b2*
b2_star <- b_star[2]; b2_star
```

```
## [1] -0.1106149
```

```r
# b3*
b3_star <- b_star[3]; b3_star
```

```
## [1] -0.2339312
```

```r
# b1 = (sY/s1)b1*
b1 <- (s_y/s_x1)*b1_star; b1
```

```
## [1] -1.141612
```

```r
# b2 = (sY/s2)b2*
b2 <- (s_y/s_x2)*b2_star; b2
```

```
## [1] -0.4420043
```

```r
# b3 = (sY/s3)b3*
b3 <- (s_y/s_x3)*b3_star; b3
```

```
## [1] -13.47016
```

```r
# b0 = ybar - b1*x1bar - b2*x2bar - b3*x3bar
b0 <- y_bar - b1*x_1bar - b2*x_2bar - b3*x_3bar; b0
```

```
## [1] 158.4913
```

Standardized regression model: $\hat{Y}^* =$ -0.5907X_1^* + -0.1106X_2^* + -0.2339X_3^*

Fitted regression model in the original variables: $\hat{Y} =$ 158.4913 + -1.1416X_1 + -0.442X_2 + -13.4702X_3

7.30. Refer to Brand preference Problem 6.5.

```r
# Import data set
df <- read.table("CH06PR05.txt")
# Set column names
colnames(df) <- c("Y", "X1", "X2")
```

a. Regress Y on X_2 using simple linear regression model (2.1) and obtain the residuals.

```r
# Fit model
model_1 <- lm(Y ~ X2, data = df)
summary(model_1)
```

```
##
## Call:
## lm(formula = Y ~ X2, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.375  -7.312  -0.125   8.688  16.625
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   68.625      8.610   7.970 1.43e-06 ***
## X2             4.375      2.723   1.607     0.13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.89 on 14 degrees of freedom
## Multiple R-squared:  0.1557, Adjusted R-squared:  0.09539
## F-statistic: 2.582 on 1 and 14 DF,  p-value: 0.1304
```

```r
# Residuals
resid_1 <- summary(model_1)$residuals; resid_1
```

```
##       1       2       3       4       5       6       7       8       9      10
## -13.375 -13.125 -16.375 -10.125  -5.375  -6.125  -6.375  -3.125   5.625   2.875
##      11      12      13      14      15      16
##   8.625   6.875  10.625   8.875  16.625  13.875
```

Model: $\hat{Y} = 68.625 + 4.375X_2$

b. Regress X_1 on X_2 using simple linear regression model (2.1) and obtain the residuals.

```r
# Fit model
model_2 <- lm(X1 ~ X2, data = df)
summary(model_2)
```

```
##
```

```
## Call:
## lm(formula = X1 ~ X2, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -3.0   -1.5    0.0    1.5    3.0
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.000e+00  1.890e+00   3.704  0.00236 **
## X2          1.110e-16  5.976e-01   0.000  1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.39 on 14 degrees of freedom
## Multiple R-squared:  3.944e-32,  Adjusted R-squared:  -0.07143
## F-statistic: 5.522e-31 on 1 and 14 DF,  p-value: 1
```

```r
# Residuals
resid_2 <- summary(model_2)$residuals; resid_2
```

```
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16
## -3 -3 -3 -3 -1 -1 -1 -1  1  1  1  1  3  3  3  3
```

Model: $\hat{X}_1 = 7$

c. Calculate the coefficient of simple correlation between the two sets of residuals and show that it equals $r_{Y1|2}$.

```r
# Coefficient of simple correlation
simple_cor <- cor(resid_1, resid_2); simple_cor
```

```
## [1] 0.9711943
```

```r
# Fit model
model <- lm(Y ~ X2 + X1, data = df)
# ANOVA table
anova(model)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X2         1  306.25  306.25  42.219 2.011e-05 ***
## X1         1 1566.45 1566.45 215.947 1.778e-09 ***
## Residuals 13   94.30    7.25
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# SSR(X1|X2)
SSR_x1x2 <- anova(model)[2,2]; SSR_x1x2
```

```
## [1] 1566.45
```

```r
SSE_x2 <- anova(model)[3,2]; SSE_x2
```

```
## [1] 94.3
```

```r
# SSR(X1|X2) = SSE(X2) - SSE(X1,X2)
# SSE(X2) = SSR(X1|X2) + SSE(X1,X2)
# R^2_{Y1|2} = SSR(X1|X2) / SSE(X2)
coeff_det <- SSR_x1x2 / (SSR_x1x2 + SSE_x2); coeff_det
```

```
## [1] 0.9432184
```

7.34. Refer to the work crew productivity example in Table 7.6.

```r
# Create data frame of data
df <- data.frame(X1 = c(4,4,4,4,6,6,6,6),
                 X2 = c(2,2,3,3,2,2,3,3),
                 Y = c(42,39,48,51,49,53,61,60))
# Create data frame for transformed data
transformed_df <- data.frame(matrix(ncol = 3, nrow = 8))
colnames(transformed_df) <- c("Y_star", "X1_star", "X2_star")
# Standardize variables
y_bar <- mean(df$Y)
s_y <- sd(df$Y)
x_1bar <- mean(df$X1)
s_x1 <- sd(df$X1)
x_2bar <- mean(df$X2)
s_x2 <- sd(df$X2)
for (i in 1:8) {
  transformed_df[i,1] <- (1/sqrt(8-1)) * ((df$Y[i]-y_bar)/s_y)
  transformed_df[i,2] <- (1/sqrt(8-1)) * ((df$X1[i]-x_1bar)/s_x1)
  transformed_df[i,3] <- (1/sqrt(8-1)) * ((df$X2[i]-x_2bar)/s_x2)
}
# Fit model
model <- lm(Y_star ~ X1_star + X2_star, data = transformed_df)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: Y_star
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X1_star    1 0.55046 0.55046  65.567 0.0004657 ***
## X2_star    1 0.40756 0.40756  48.546 0.0009366 ***
## Residuals  5 0.04198 0.00840
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

a. For the variables transformed according to (7.44), obtain:

```r
# X'X = r_{XX}
cor1 <- cor(transformed_df)[2:3,2:3]; cor1
```

(1) X'X

```
##         X1_star X2_star
## X1_star       1       0
## X2_star       0       1
```

```r
# X'Y = r_{XY}
cor2 <- cor(transformed_df)[2:3,1]; cor2
```

(2) X'Y

```
##   X1_star   X2_star
## 0.7419309 0.6384057
```

```r
# b = (X'X)^-1 * X'Y
b <- solve(cor1) %*% cor2; b
```

(3) b

```
##              [,1]
## X1_star 0.7419309
## X2_star 0.6384057
```

```r
# s^2{b} = (s*)^2 * r_{xx}^-1
anova(model)[3,3] * solve(cor1)
```

(4) $s^2\{b\}$

```
##             X1_star     X2_star
## X1_star 0.008395356 0.000000000
## X2_star 0.000000000 0.008395356
```

b. Show that the standardized regression coefficients obtained in part (a3) are related to the regression coefficients for the regression model in the original variables according to (7.53).

```r
# b1 = (sY/s1)b1*
# b1* = (s1/sY)b1
b1_star <- sd(transformed_df$X1)/sd(transformed_df$Y) * b[1]; b1_star
```

```
## [1] 0.7419309
```

```r
# b2 = (sY/s2)b2*
# b2* = (s2/sY)b2
b2_star <- sd(transformed_df$X2)/sd(transformed_df$Y) * b[2]; b2_star
```

```
## [1] 0.6384057
```