CSCI E-106: Section 07

Date: 10/24/2018

Contents

Section Problems

(7.04) Reference to Grocery retailer Problem 6.9.

A large,national grocery retailer tracks productivity and costs of its facilities closely. Data were obtained from a single distribution center for a one-year period. Each data point for each variable represents one week of activity. The variables included are the number of cases shipped (X_1), the indirect costs of the total labor hours as a percentage (X_2), a qualitative predictor called holiday that is coded 1 if the week has a holiday and 0 otherwise (X_3), and the total labor hours (Y).

Please use dataset titled ***CH06PR09.txt*** *when applicable*

```
#Load the data
df_704 = read.delim(file="CH06PR09.txt", sep="",
                        header = FALSE,
                        col.names = c("laborHours",
                                      "shippedCases",
                                      "indirectCosts",
                                      "holiday"))
```

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

Solution Below

```
lmFit_704 = lm(laborHours ~ shippedCases+holiday+indirectCosts, data=df_704)

summary(lmFit_704)
```

```
##
## Call:
## lm(formula = laborHours ~ shippedCases + holiday + indirectCosts,
##     data = df_704)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -264.05 -110.73  -22.52   79.29  295.75
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.150e+03  1.956e+02  21.220  < 2e-16 ***
## shippedCases   7.871e-04  3.646e-04   2.159   0.0359 *
## holiday        6.236e+02  6.264e+01   9.954 2.94e-13 ***
## indirectCosts -1.317e+01  2.309e+01  -0.570   0.5712
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 143.3 on 48 degrees of freedom
## Multiple R-squared:  0.6883, Adjusted R-squared:  0.6689
## F-statistic: 35.34 on 3 and 48 DF,  p-value: 3.316e-12
```

```r
anovaTable = data.frame(anova(lmFit_704))
totals = c(round(sum(anovaTable[,1])), round(sum(anovaTable[,2])), "", "", "")
anovaTable = rbind(anovaTable, totals)

#add names to the table
row.names(anovaTable) = c("SSR(X1)", "SSR(X3|X1)", "SSR(X2|(X1X3))", "SSE", "Total")
colnames(anovaTable) = c("DF", "Sum Sq.", "Mean Sq.", "F-Value", "Pr(>F)")


#print our analysis of variance table
kable(anovaTable)
```

	DF	Sum Sq.	Mean Sq.	F-Value	Pr(>F)
SSR(X1)	1	136366.24332931	136366.24332931	6.64168656883043	0.0130903845875702
SSR(X3|X1)	1	2033565.34526788	2033565.34526788	99.0443331924046	2.96333996706384e-13
SSR(X2|(X1X3))	1	6674.58808589154	6674.58808589154	0.325084280099203	0.571227380591661
SSE	48	985529.746393846	20531.8697165385	NA	NA
Total	51	3162136			

SSR(X1) = Sum of squares associated with X1

SSR(X3|X1) = Sum of squares associated with X3 given X1

SSR(X2|(X1X3)) = Sum of squares into extra sums of squares associated with X2 given X1 and X3

b. Test whether X_2 can be dropped from the regression model given that X_1, and X_3 are retained. Use the F* test statistic and $\alpha = .05$. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

Solution Below

```r
ssr = as.numeric(anovaTable[3,2])
sse = as.numeric(anovaTable[4,2])

fStar = (ssr/1) / (sse/as.numeric(anovaTable[4,1]))
print(fStar)
```

```
## [1] 0.3250843
```

```r
#alpha is given
alpha = 0.05

# df from Summary above in a
db = qf(1-alpha, 1, 48)
print(db)
```

```
## [1] 4.042652
```

ANALYSIS

Hypotheses:

$H_0 : \beta_2 = 0$

$H_a : \beta_2 \neq 0$

Decision Rules:

If $F^* \leq 4.0426521$, conclude H_0

If $F^* > 4.0426521$, conclude H_a

Conclusion:

Since our test statistic, $F^* = 0.3250843$, and $0.3250843 \leq 4.0426521$, we conclude H_0.

c. Does $SSR(X_1) + SSR(X_2|X_1)$ equal $SSR(X_2) + SSR(X_1|X_2)$ here? Must this always be the case? (Does our sum of squares associated with x1 plus sum of squares associated with x2 given x1 equal sum of squares associated with x2 plus sum of squares associated with x1 given x2?)

Solution Below

```r
ssr_x1 = anova(lm(laborHours~shippedCases+indirectCosts, data=df_704))[1,2]
ssr_x2x1 = anova(lm(laborHours~shippedCases+indirectCosts, data=df_704))[2,2]
eq1_sum = round(ssr_x1+ssr_x2x1)
ssr_x1 = paste0(ssr_x1)
ssr_x2x1 = paste0(ssr_x2x1)
eq1_sum = paste0(eq1_sum)
print(eq1_sum)
```

```
## [1] "142092"
```

```r
ssr_x2 = anova(lm(laborHours~indirectCosts+shippedCases, data=df_704))[1,2]
ssr_x1x2 = anova(lm(laborHours~indirectCosts+shippedCases, data=df_704))[2,2]
eq2_sum = round(ssr_x2+ssr_x1x2)
ssr_x2 = paste0(ssr_x2)
ssr_x1x2 = paste0(ssr_x1x2)
eq2_sum = paste0(eq2_sum)
print(eq2_sum)
```

```
## [1] "142092"
```

ANALYSIS

We can calculate this mathematically to see if $SSR(X_1) + SSR(X_2|X_1) = SSR(X_2) + SSR(X_1|X_2)$.

Equation 1: $SSR(X_1) + SSR(X_2|X_1)$

$$SSR(X_1) + SSR(X_2|X_1)$$
$$136366.24332931 + 5725.9218096369 = 142092$$

Equation 2: $SSR(X_2) + SSR(X_1|X_2)$

$$SSR(X_2) + SSR(X_1|X_2)$$
$$11394.9229160115 + 130697.242222935 = 142092$$

Combining equation 1 and equation 2:

$$SSR(X_1) + SSR(X_2|X_1) = SSR(X_2) + SSR(X_1|X_2)$$
$$142092 = 142092$$

As a result, we see that $SSR(X_1) + SSR(X_2|X_1) = SSR(X_2) + SSR(X_1|X_2)$. It will always be the case where the expressions are equivalent because of the inherent symmetry of the models.

(7.38) Projects. Reference to SENIC data set in Appendix C.1.

> *The primary objective of the Study on the Efficacy of Nosocomial Infection Control (SENIC Project) was to determine whether infection surveillance and control programs have reduced the rates of nosocomial (hospital-acquired) infection in United States hospitals. This data set consists of a random sample of 113 hospitals selected from the original 338 hospitals surveyed. Each line of the dataset has an identification number and provides information on 11 variables for a single hospital. The data presented here are for the 1975-76 study period.*

*Please use dataset titled **APPENC01.txt** when applicable*

For predicting the average length of stay of patients in a hospital (Y), it has been decided to include age (X_1) and infection risk (X_2) as predictor variables. The question now is whether an additional predictor variable would be helpful in the model and, if so, which variable would be most helpful. Assume that a first-order multiple regression model is appropriate.

a. For each of the following variables, calculate the coefficient of partial determination given that X_1 and X_2 are included in the model: routine culturing ratio (X_3), average daily census (X_4), number of nurses (X_5), and available facilities and services (X_6).

Solution Below

```r
# Y: Average length of stay of patients in hospital
# X1: Age
# X2: Infection Risk
# X3: Routine Culturing Ratio
# X4: Average Daily Census
# X5: Number of Nurses
# X6: Available Facilities & Services

df_738 = read.delim(file="APPENC01.txt", sep="", header = FALSE)[,c(2,3,4,5,10,11,12)]
colnames(df_738) = c("Y","X1","X2","X3","X4","X5","X6")
```

```r
r2_X3 = anova(lm(Y~X1+X2+X3,df_738))[3,2]/sum(anova(lm(Y~X1+X2+X3,df_738))[3:4,2])

r2_X4 = anova(lm(Y~X1+X2+X4,df_738))[3,2]/sum(anova(lm(Y~X1+X2+X4,df_738))[3:4,2])

r2_X5 = anova(lm(Y~X1+X2+X5,df_738))[3,2]/sum(anova(lm(Y~X1+X2+X5,df_738))[3:4,2])

r2_X6 = anova(lm(Y~X1+X2+X6,df_738))[3,2]/sum(anova(lm(Y~X1+X2+X6,df_738))[3:4,2])
```

ANALYSIS

$R^2_{3|12} = 0.0116729$

$R^2_{4|12} = 0.1362033$

$R^2_{5|12} = 0.0373663$

$R^2_{6|12} = 0.0363888$

b. On the basis of the results in part (a), which of the four additional predictor variables is best? Is the extra sum of squares associated with this variable larger than those for the other three variables?

Solution Below

ANALYSIS

Based on the results from part (a), it looks like the fourth addition of X_4 (which is our second calculation above) would be the best. The extra sum of squares is associated with X_4 and it is larger than the other three variables.

c. Using the $F*$ test statistic, test whether or not the variable determined to be best in part (b) is helpful in the regression model when X_1 and X_2 are included in the model; use $\alpha = .05$. State the alternatives, decision rule, and conclusion. Would the $F*$ test statistics for the other three potential predictor variables be as large as the one here? Discuss.

Solution Below

```r
(anova_738X4 = anova(lm(Y~X1+X2+X4,df_738)))
```

```
## Analysis of Variance Table
##
## Response: Y
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## X1          1  14.604  14.604   6.623   0.01141 *
## X2          1 116.356 116.356  52.768 5.928e-11 ***
## X4          1  37.899  37.899  17.187 6.722e-05 ***
## Residuals 109 240.352   2.205
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
ssr = anova_738X4[3,2]
sse = anova_738X4[4,2]

fStar = (ssr/1) / (sse/(anova_738X4[4,1]))
print(fStar)
```

```
## [1] 17.1871
```

```r
alpha = 0.05

db <- qf(1-alpha,1,(anova_738X4[4,1]))
print(db)
```

```
## [1] 3.928195
```

ANALYSIS

Hypotheses:

$H_0 : \beta_4 = 0$

$H_a : \beta_4 \neq 0$

Decision Rules:

If $F^* \leq 3.9281951$, conclude H_0

If $F^* > 3.9281951$, conclude H_a

Conclusion:

Since our test statistic, $F^* = 17.187105$, and $17.187105 > 3.9281951$, we conclude H_a.

Recall the current model already includes X_1 and X_2, which are variables representing age and infection risk, respectively. Adding X_4 (Average Daily Census) to the current model would contribute more predictive power to predict the average length of hospital stay of patients.

In addition, the F^* test statistics for the other three potential predictor variables would not be as large as the one obtained for X_4 since the SSR values for the other variables would be smaller.

(8.21) In a regression analysis of on-the-job head injuries of warehouse laborers caused by fulling objects, Y is a measure of severity of the injury, X_1 is an index inflecting both the weight of the object and the distance it fell, and X_2 and X_3 are indicator variables for nature of head protection worn at the time of the accident, coded as follows:

Type of Prediction	X_2	X_3
Hard Hat	1	0
Bump Cap	0	1
None	0	0

The response function to be used in the study is $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$.

a. Develop the response function for each type of protection category.

Solution Below

Protection Category	Response Function
Hard Hat	$E\{Y\} = (\beta_0 + \beta_2) + \beta_1 X_1$
Bump Cap	$E\{Y\} = (\beta_0 + \beta_3) + \beta_1 X_1$
None	$E\{Y\} = \beta_0 + \beta_1 X_1$

ANALYSIS

The response function used in the study implies that the regression of protection on head injuries is linear, with the same slope for all types of protections. The coefficients (β_2, β_3) indicate how much lower or higher the response functions for the protections models are than the no-protection category (e.g., 'None'). Thus, β_2 and β_3 measures the differential effects of the qualitative variable class. Differential effects of one qualitative variable on the intercept depend on the particular class of the other qualitative variable.

b. For each of the following questions, specify the alternatives H_0 and H_a for the appropriate test: (1) With X_1 fixed, does wearing a bump cap reduce the expected severity of injury as compared with wearing no protection? (2) With X_1 fixed, is the expected severity of injury the same when wearing a hard hat as when wearing a bump cap?

Solution Below

1. With X_1 fixed, does wearing a bump cap reduce the expected severity of injury as compared with wearing no protection? Null and alternative hypotheses as follows:

$$H_0 : \beta_3 \geq 0 H_a : \beta_3 < 0$$

2. With X_1 fixed, is the expected severity of injury the same when wearing a hard hat as when wearing a bump cap? Null and alternative hypotheses as follows:

$$H_0 : \beta_2 = \beta_3 H_a : \beta_2 \neq \beta_3$$

(8.38) Projects. Reference to SENIC data set in Appendix C.1.

The primary objective of the Study on the Efficacy of Nosocomial Infection Control (SENIC Project) was to determine whether infection surveillance and control programs have reduced the rates of nosocomial (hospital-acquired) infection in United States hospitals. This data set consists of a random sample of 113 hospitals selected from the original 338 hospitals surveyed. Each line of the dataset has an identification number and provides information on 11 variables for a single hospital. The data presented here are for the 1975-76 study period.

```
# Y: Number of Nurses
# X: Available Facilities & Services

df_838 = read.table(file='APPENC01.txt', sep='', header=FALSE)[,c(11,12)]
colnames(df_838) = c('Y','X')
```

Second-order regression model (8.2) is to be fitted for relating number of nurses (Y) to available facilities and services (X).

a. Fit the second-order regression model. Plot the residuals against the fitted values. How well does the second-order model appear to fit the data?

Solution Below

Recall that the second-order regression model (8.2) is the following:

$$Y = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2 + \epsilon_i$$

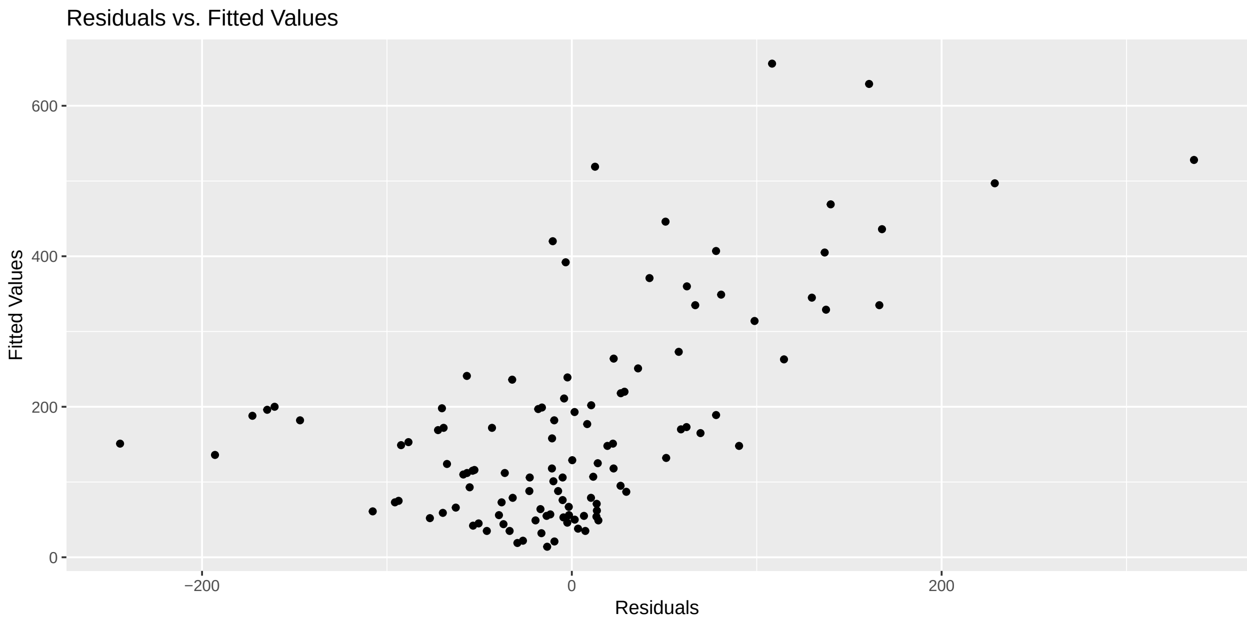
where $x_i = X_i - \bar{X}$. Since X and X^2 will be highly correlated, centering the predictor variable often reduces the multicollinearity substantially and tends to avoid computational difficulties.

```
# Center the predictor
df_838$x = df_838$X - mean(df_838$X) # generate quadratic variable

# Create the model
lm_838 = lm(Y ~ x + I(x^2), data=df_838)
summary(lm_838)
```

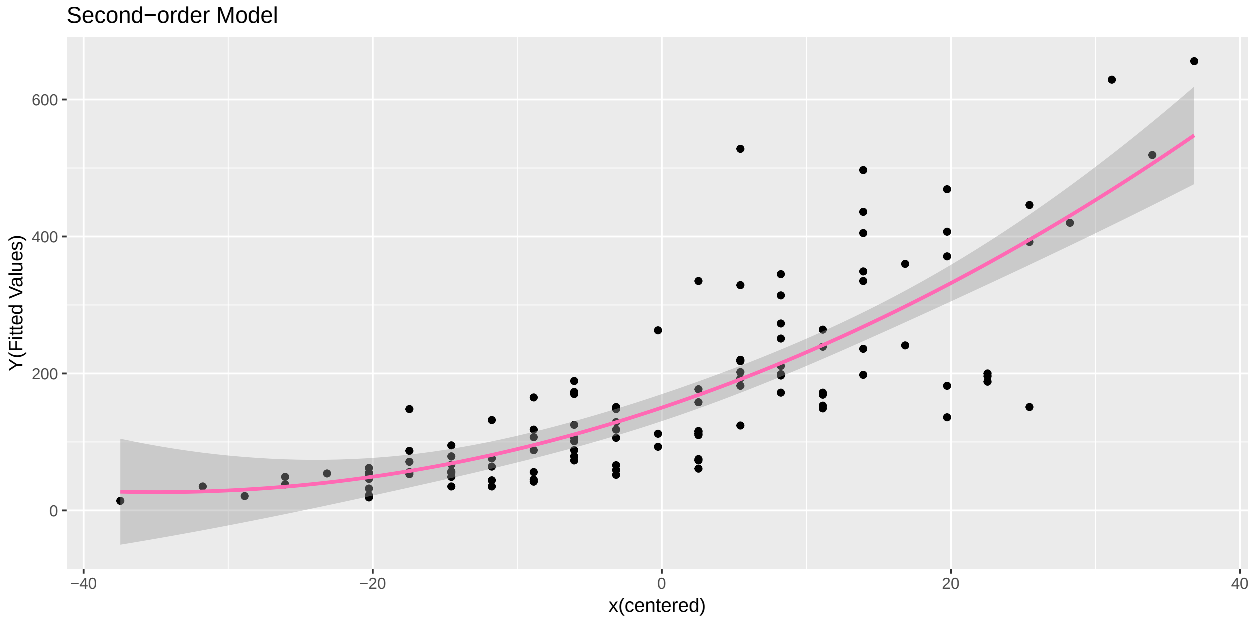
```
##
## Call:
## lm(formula = Y ~ x + I(x^2), data = df_838)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -244.32  -39.42   -4.55   26.48  336.48
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 150.07915    9.94139  15.096  < 2e-16 ***
## x             7.06617    0.51253  13.787  < 2e-16 ***
## I(x^2)        0.10116    0.02723   3.716  0.00032 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 82.31 on 110 degrees of freedom
## Multiple R-squared:  0.6569, Adjusted R-squared:  0.6507
## F-statistic: 105.3 on 2 and 110 DF,  p-value: < 2.2e-16
```

```
ggplot(mapping = aes(lm_838$residuals, df_838$Y)) +
    geom_point() +
    labs(title="Residuals vs. Fitted Values", x="Residuals", y="Fitted Values")
```




```
ggplot(mapping = aes(df_838$x, df_838$Y)) +
    geom_point() +
```

```r
  geom_smooth(method = 'lm', formula=y ~ poly(x,2), col='hotpink') +
  labs(title="Second-order Model", x="x(centered)", y="Y(Fitted Values)")
```



ANALYSIS

Residuals appear to be relatively small for smaller values of Y.

Quadratic model appears to fit the data well and follows the trend of the data. R^2 indicates roughly 66% of the data is explained by the model.

b. Obtain R^2 for the second-order regression model. Also obtain the coefficient of simple determination for the first-order regression model. Has the addition of the quadratic term in the regression model substantially increased the coefficient of determination?

Solution Below

```r
rSquare = summary(lm_838)$r.squared
rSquare = paste0(signif(rSquare, digits=4))
rSquaure_simp = summary(lm(Y~X, data=df_838))$r.squared
rSquare_simp = paste0(signif(rSquaure_simp, digits=4))
```

ANALYSIS

The R^2 for the second-order regression model (AKA coefficient of multiple determination) is 0.6569 and the coefficient of simple determinate is 0.6139. We see that the coefficient of multiple determination is a slightly higher, which suggests that the quadratic term increased the proportion of the variance in the data.

c. Test whether the quadratic term can be dropped from the regression model; use $\alpha = .01$. State the alternatives, decision rule, and conclusion.

Solution Below

```r
(anova = pureErrorAnova(lm_838))
```

```
## Analysis of Variance Table
##
```

```
## Response: Y
##                Df  Sum Sq Mean Sq  F value    Pr(>F)
## x               1 1333486 1333486 237.3688 < 2.2e-16 ***
## I(x^2)          1   93533   93533  16.6495 9.939e-05 ***
## Residuals     110  745204    6775
##  Lack of fit   23  256457   11150   1.9848   0.01223 *
##  Pure Error    87  488747    5618
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
alpha = .01

SSR_x2x = anova[2,2]
SSE_xx2 = anova[3,2]

db = qf(1-alpha, 1, nrow(df_838)-3)

fStar = (SSR_x2x/1)/(SSE_xx2/(nrow(df_838)-3))
```

ANALYSIS

Hypotheses:

$H_0 : \beta_{11} = 0$

$H_a : \beta_{11} \neq 0$

Decision Rules:

If $F^* \leq 6.8710278$, conclude H_0

If $F^* > 6.8710278$, conclude H_a

Conclusion:

Since our test statistic, $F^* = 13.8065048$, and $13.8065048 > 6.8710278$, we conclude H_a where we would not drop the quadratic term from the model and keep it instead.

(9.33) Case Study. Reference to Real estate sales Case Study 9.31.

The regression model identified in Case Study 9.31 is to be validated by means of the validation data set consisting of those cases not selected for the model building data set.

9.31. *Residential sales that occurred during the year 2002 were available from a city in the Midwest. Data on 522 arms-length transactions include sales price, style, finished square feet, number of bedrooms, pool, lot size, year built, air conditioning, and whether or not the lot is adjacent to a highway. The city tax assessor was interested in predicting sales price based on the demographic variable information given above. Select a random sample of 300 observations to use in the model-building data set. Develop a best subset model for predicting sales price. Justify your choice of model. Assess your model's ability to predict and discuss its use as a tool for predicting sales price.*

Data Set C.7. Real Estate Sales. Page 1353 *The city tax assessor was interested in predicting residential home sales prices in a Midwestern city as a function of various characteristics of the home and surrounding property. Data on 522 arms-length transactions were obtained for home sales during the year 2002. Each line of the data set has an identification number and provides information on 12 other variables.*

```r
df_933 = read.table(file='APPENC07.txt', sep='', header=FALSE,
                    col.names=c('id','salesPrice','sqFt','nBeds','nBaths',
                    'ac','garageSize','pool','year','quality',
                    'style','lotSize','hwy'))
```

a. Fit the regression model identified in Case Study 9.31 to the validation data set. Compare the estimated regression coefficients and their estimated standard errors with those obtained in Case Study 9.31. Also compare the error mean square and coefficients of multiple determination. Does the model fitted to the validation data set yield similar estimates as the model fitted to the model-building data set?

Solution Below

```r
# Prep from 9.31

# Feature Engineering
age = 2002 - df_933$year
style1 = as.numeric(df_933$style == 7)
uniform = runif(nrow(df_933))
df_933_sorted = cbind(df_933, age, style1, uniform)
df_933_sorted = as.data.frame(df_933_sorted[order(uniform),])

# Partition Train and Test sets
trainSample = as.data.frame(df_933_sorted[1:300,])
valSample = as.data.frame(df_933_sorted[301:522,])

# To find the best model, basically fit the model and iteratively delete the insignificant variables

# Recall the factor variables: garage size, quality, style

summary(lm(log(salesPrice) ~ sqFt + nBeds + nBaths + ac + factor(garageSize) + pool + age + factor(quali
```

```
##
## Call:
## lm(formula = log(salesPrice) ~ sqFt + nBeds + nBaths + ac + factor(garageSize) +
##     pool + age + factor(quality) + style1 + lotSize + hwy, data = trainSample)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.70393 -0.09919 -0.00290  0.09285  0.50450
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.181e+01  1.158e-01 101.989  < 2e-16 ***
## sqFt                2.835e-04  2.798e-05  10.130  < 2e-16 ***
## nBeds               4.287e-03  1.210e-02   0.354 0.723381
## nBaths              4.721e-02  1.569e-02   3.008 0.002863 **
## ac                  4.501e-02  3.155e-02   1.427 0.154802
## factor(garageSize)1 4.581e-02  8.940e-02   0.512 0.608780
## factor(garageSize)2 9.925e-02  8.603e-02   1.154 0.249581
## factor(garageSize)3 1.380e-01  9.145e-02   1.509 0.132351
## factor(garageSize)4 3.730e-02  1.480e-01   0.252 0.801184
## factor(garageSize)5 7.964e-02  1.893e-01   0.421 0.674262
## factor(garageSize)7 3.041e-02  1.956e-01   0.155 0.876579
## pool                8.490e-02  4.141e-02   2.050 0.041260 *
## age                -2.972e-03  7.755e-04  -3.833 0.000156 ***
```

```
## factor(quality)2    -3.013e-01  4.184e-02  -7.202 5.42e-12 ***
## factor(quality)3    -3.724e-01  5.474e-02  -6.803 6.12e-11 ***
## style1              -5.995e-02  3.022e-02  -1.984 0.048273 *
## lotSize              3.877e-06  9.443e-07   4.106 5.28e-05 ***
## hwy                 -1.221e-02  6.430e-02  -0.190 0.849568
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1633 on 282 degrees of freedom
## Multiple R-squared:  0.8531, Adjusted R-squared:  0.8443
## F-statistic: 96.35 on 17 and 282 DF,  p-value: < 2.2e-16
```

```
summary(lm(log(salesPrice) ~ sqFt + nBaths + ac + factor(garageSize) + pool + age + factor(quality) + s
```

```
##
## Call:
## lm(formula = log(salesPrice) ~ sqFt + nBaths + ac + factor(garageSize) +
##     pool + age + factor(quality) + style1 + lotSize + hwy, data = trainSample)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.70337 -0.09833 -0.00061  0.09210  0.50397
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.181e+01  1.155e-01 102.278  < 2e-16 ***
## sqFt                 2.866e-04  2.648e-05  10.823  < 2e-16 ***
## nBaths               4.861e-02  1.516e-02   3.206 0.001498 **
## ac                   4.562e-02  3.146e-02   1.450 0.148098
## factor(garageSize)1  5.057e-02  8.825e-02   0.573 0.567077
## factor(garageSize)2  1.037e-01  8.499e-02   1.220 0.223556
## factor(garageSize)3  1.418e-01  9.070e-02   1.563 0.119196
## factor(garageSize)4  4.000e-02  1.476e-01   0.271 0.786567
## factor(garageSize)5  8.585e-02  1.882e-01   0.456 0.648603
## factor(garageSize)7  2.797e-02  1.952e-01   0.143 0.886192
## pool                 8.494e-02  4.134e-02   2.055 0.040843 *
## age                 -2.979e-03  7.740e-04  -3.849 0.000147 ***
## factor(quality)2    -2.991e-01  4.132e-02  -7.240 4.25e-12 ***
## factor(quality)3    -3.701e-01  5.428e-02  -6.819 5.54e-11 ***
## style1              -6.131e-02  2.993e-02  -2.049 0.041426 *
## lotSize              3.883e-06  9.427e-07   4.120 4.99e-05 ***
## hwy                 -1.311e-02  6.415e-02  -0.204 0.838187
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.163 on 283 degrees of freedom
## Multiple R-squared:  0.8531, Adjusted R-squared:  0.8447
## F-statistic: 102.7 on 16 and 283 DF,  p-value: < 2.2e-16
```

```
summary(lm(log(salesPrice) ~ sqFt + nBaths + ac + pool + age + factor(quality) + style1 + lotSize+ hwy,
```

```
##
## Call:
## lm(formula = log(salesPrice) ~ sqFt + nBaths + ac + pool + age +
##     factor(quality) + style1 + lotSize + hwy, data = trainSample)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.70253 -0.09939 -0.00329  0.09018  0.50300
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.191e+01  8.339e-02 142.831  < 2e-16 ***
## sqFt              2.896e-04  2.558e-05  11.321  < 2e-16 ***
## nBaths            5.045e-02  1.501e-02   3.361  0.00088 ***
## ac                5.862e-02  3.043e-02   1.927  0.05499 .
## pool              8.570e-02  4.011e-02   2.137  0.03345 *
## age              -3.251e-03  7.509e-04  -4.329 2.06e-05 ***
## factor(quality)2 -3.094e-01  3.913e-02  -7.906 5.64e-14 ***
## factor(quality)3 -3.862e-01  5.204e-02  -7.422 1.30e-12 ***
## style1           -5.831e-02  2.947e-02  -1.979  0.04881 *
## lotSize           3.912e-06  9.326e-07   4.195 3.63e-05 ***
## hwy              -6.139e-03  6.380e-02  -0.096  0.92340
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.163 on 289 degrees of freedom
## Multiple R-squared:  0.8499, Adjusted R-squared:  0.8447
## F-statistic: 163.7 on 10 and 289 DF,  p-value: < 2.2e-16
```

```r
summary(lm(log(salesPrice) ~ sqFt + nBaths + ac + age + factor(quality) + style1 + lotSize+ hwy, data=t
```

```
##
## Call:
## lm(formula = log(salesPrice) ~ sqFt + nBaths + ac + age + factor(quality) +
##     style1 + lotSize + hwy, data = trainSample)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.63620 -0.10007 -0.00372  0.08822  0.49888
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.190e+01  8.377e-02 142.066  < 2e-16 ***
## sqFt              2.948e-04  2.562e-05  11.509  < 2e-16 ***
## nBaths            5.362e-02  1.503e-02   3.568  0.00042 ***
## ac                6.063e-02  3.060e-02   1.982  0.04848 *
## age              -3.231e-03  7.555e-04  -4.276 2.58e-05 ***
## factor(quality)2 -3.105e-01  3.936e-02  -7.887 6.35e-14 ***
## factor(quality)3 -3.872e-01  5.236e-02  -7.394 1.54e-12 ***
## style1           -6.778e-02  2.931e-02  -2.312  0.02147 *
## lotSize           3.760e-06  9.355e-07   4.019 7.47e-05 ***
## hwy              -9.822e-03  6.416e-02  -0.153  0.87845
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.164 on 290 degrees of freedom
## Multiple R-squared:  0.8476, Adjusted R-squared:  0.8428
## F-statistic: 179.2 on 9 and 290 DF,  p-value: < 2.2e-16
```

```r
trainModel = lm(log(salesPrice) ~ sqFt + nBaths + ac + age + factor(quality) + style1 + lotSize+ hwy, da
```

```r
# Preliminary Analysis
#summary(df_933)
#lapply(df_933, mode)
#lapply(df_933, class)


testModel = lm(log(salesPrice) ~ sqFt + nBaths + ac + age + factor(quality) + style1 + lotSize+ hwy, da

summary(trainModel)
```

```
##
## Call:
## lm(formula = log(salesPrice) ~ sqFt + nBaths + ac + age + factor(quality) +
##     style1 + lotSize + hwy, data = trainSample)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.63620 -0.10007 -0.00372  0.08822  0.49888
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.190e+01  8.377e-02 142.066  < 2e-16 ***
## sqFt              2.948e-04  2.562e-05  11.509  < 2e-16 ***
## nBaths            5.362e-02  1.503e-02   3.568  0.00042 ***
## ac                6.063e-02  3.060e-02   1.982  0.04848 *
## age              -3.231e-03  7.555e-04  -4.276 2.58e-05 ***
## factor(quality)2 -3.105e-01  3.936e-02  -7.887 6.35e-14 ***
## factor(quality)3 -3.872e-01  5.236e-02  -7.394 1.54e-12 ***
## style1           -6.778e-02  2.931e-02  -2.312  0.02147 *
## lotSize           3.760e-06  9.355e-07   4.019 7.47e-05 ***
## hwy              -9.822e-03  6.416e-02  -0.153  0.87845
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.164 on 290 degrees of freedom
## Multiple R-squared:  0.8476, Adjusted R-squared:  0.8428
## F-statistic: 179.2 on 9 and 290 DF,  p-value: < 2.2e-16
```

```r
summary(valSample)
```

```
##        id          salesPrice          sqFt          nBeds
##  Min.   :  1.0   Min.   : 95500   Min.   :1060   Min.   :1.000
##  1st Qu.:133.2   1st Qu.:179925   1st Qu.:1700   1st Qu.:3.000
##  Median :232.0   Median :241250   Median :2156   Median :4.000
##  Mean   :254.7   Mean   :288483   Mean   :2311   Mean   :3.545
##  3rd Qu.:390.8   3rd Qu.:360000   3rd Qu.:2687   3rd Qu.:4.000
##  Max.   :522.0   Max.   :920000   Max.   :4756   Max.   :7.000
##      nBaths            ac           garageSize         pool
##  Min.   :1.000   Min.   :0.0000   Min.   :0.000   Min.   :0.00000
##  1st Qu.:2.000   1st Qu.:1.0000   1st Qu.:2.000   1st Qu.:0.00000
##  Median :3.000   Median :1.0000   Median :2.000   Median :0.00000
##  Mean   :2.743   Mean   :0.8198   Mean   :2.072   Mean   :0.07658
##  3rd Qu.:3.000   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:0.00000
```

```
##  Max.   :7.000   Max.   :1.0000   Max.   :3.000   Max.   :1.00000
##       year         quality          style          lotSize
##  Min.   :1885   Min.   :1.000   Min.   :1.000   Min.   : 5666
##  1st Qu.:1956   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:16600
##  Median :1966   Median :2.000   Median :3.000   Median :22094
##  Mean   :1967   Mean   :2.176   Mean   :3.509   Mean   :24344
##  3rd Qu.:1980   3rd Qu.:3.000   3rd Qu.:7.000   3rd Qu.:27220
##  Max.   :1998   Max.   :3.000   Max.   :7.000   Max.   :86830
##       hwy                age              style1          uniform
##  Min.   :0.00000   Min.   :  4.00   Min.   :0.0000   Min.   :0.5762
##  1st Qu.:0.00000   1st Qu.: 22.00   1st Qu.:0.0000   1st Qu.:0.6762
##  Median :0.00000   Median : 36.00   Median :0.0000   Median :0.7723
##  Mean   :0.01802   Mean   : 34.55   Mean   :0.3198   Mean   :0.7822
##  3rd Qu.:0.00000   3rd Qu.: 46.00   3rd Qu.:1.0000   3rd Qu.:0.9034
##  Max.   :1.00000   Max.   :117.00   Max.   :1.0000   Max.   :0.9999
```

ANALYSIS

R^2 value for the training model was slightly higher and we see that the R^2 value for the validation model dropped. We can further analyze the variables in the model summaries. Comparing the variables, we see some notable differences. Specifically: number of baths, AC, style1 (our dummy variable), highway.

b. Calculate the mean squared prediction error (9.20) and compare it to MSE obtained from the model-building data set. Is there evidence of a substantial bias problem in MSE here?

Solution Below

```r
anova(trainModel)
```

```
## Analysis of Variance Table
##
## Response: log(salesPrice)
##                  Df Sum Sq Mean Sq   F value    Pr(>F)
## sqFt              1 37.353  37.353 1388.5146 < 2.2e-16 ***
## nBaths            1  1.555   1.555   57.7956 4.061e-13 ***
## ac                1  0.425   0.425   15.7864 8.954e-05 ***
## age               1  1.326   1.326   49.2792 1.577e-11 ***
## factor(quality)   2  2.049   1.024   38.0828 2.063e-15 ***
## style1            1  0.233   0.233    8.6556  0.003524 **
## lotSize           1  0.436   0.436   16.2257 7.187e-05 ***
## hwy               1  0.001   0.001    0.0234  0.878451
## Residuals       290  7.801   0.027
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
(MSE = anova(trainModel)[9,3])
```

```
## [1] 0.02690158
```

```r
# See MSE is ~0.033
MSE = paste0(signif(MSE, digits=4))

predsTest = predict(trainModel, valSample)

(MSPR = sum((log(valSample$salesPrice) - predsTest)^2)/(nrow(valSample)))
```

```
## [1] 0.03810536
```

```
MSPR = paste0(signif(MSPR, digits=4))
```

ANALYSIS

The MSE obtainined from the model-building set is 0.0269 and the mean squared prediction error is 0.03811. We see that the two values are fairly similar, with variations between the two being small. There is no evidence of a substantial bias problem in the MSE.