Multiple Linear Regression

Code

ASSIGNMENT 2

First-order Model with Interaction Term (Quantitative and Qualitative Variable and Model Selection

Deadline: Nov. 25 , 2022, by 11:59 pm. Submit to Gradescope.ca

Problem 1. The file **tires.csv** provides the results of an experiment on tread wear per 160 km and the driving speed in km/hour. The researchers looked at 2 types of tires and tested 20 random sample tires. The response variable is the tread wear per 160 km in the percentage of tread thickness, and the quantitative predictor is the average speed in km/hour.

Hide

```
tires=read.csv("c:/Users/thunt/OneDrive - University of Calgary/dataset603/tires.CSV", header = TRUE)
str(tires)
```

```
'data.frame':   140 obs. of  3 variables:
 $ type: chr  "A" "A" "A" "A" ...
 $ wear: num  0.3 0.3 0.3 0.3 0.3 0.3 0.3 0.4 0.4 0.4 ...
 $ ave : int  80 80 80 80 80 80 80 88 88 88 ...
```

Answer the following questions

a. Use the individual T-test to evaluate the significant predictors from the full model at $\alpha = 0.05$ and write the estimated best fit model.

Hide

```
additivemodel<-lm(wear~factor(type)+ave,data=tires)
summary(additivemodel)
```

```
Call:
lm(formula = wear ~ factor(type) + ave, data = tires)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.092858 -0.033451 -0.000953  0.039404  0.116668 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -0.6445083  0.0525675  -12.26   <2e-16 ***
factor(type)B  0.1725006  0.0093544   18.44   <2e-16 ***
ave            0.0113094  0.0005155   21.94   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05384 on 137 degrees of freedom
Multiple R-squared:  0.8861,    Adjusted R-squared:  0.8844 
F-statistic: 532.8 on 2 and 137 DF,  p-value: < 2.2e-16
```

<!– The R command summary(additivemodel) (see R code above) shows that we can write the model as

$$\widehat{wear} = -0.6445083 + 0.1725006type + 0.0113094ave.$$

Moreover, the overall test F shows that at least one of the predictors must be related to the tread wear per 160 km in percentage of tread thickness for a car as the p-value is < 2.2e-16 <0.05. –>

b. Based on the output in (a), define the dummy variable that explains the two types of tires.

<!– It’s defined as type=0 if tire A and type=1 if tire B.

–>

c. From the best fit model in part (a), interpret all possible regression coefficient estimates, $\hat{\beta}_i$.

$\widehat{\beta_1} = 0.1725006$ represents the difference in the tread wear per 160 km in percentage of tread thickness for a car between tire type A and B.

$\widehat{\beta_2} = 0.0113094$ means that the average speed increases 1 km/hour, leads to an increase in the tread wear per 160 km of tread thickness by 0.0113094 %.

–>

d. From the best fit model in part (a), you can improve this model by adding an interaction term(s). Evaluate whether the interaction term(s) is(are) significant to be added in the model at $\alpha = 0.05$. Summarize Which model would you suggest using for predicting y?

Hide

```
interacmodel<-lm(wear~factor(type)+ave+factor(type)*ave,data=tires)
summary(interacmodel)
```

!– Output from summary(interacmodel) shows that the interaction is significant at α = .05 (p-value < 2.2e-16).

Hence, the best fit model would be

$$\widehat{wear} = -0.3888744 - 1.0800050 type + 0.0087833 ave + 0.0119840 type \times ave$$

as it fits the data better than the additive model in part (a)) with the R^2_{Adj}=0.96 and RMSE= 0.03169

–>

e. From the model in part (d), report the adjusted-R2 value from the model selected and interpret its value.

<!– R^2_{Adj}=0.96 means that 96% of the variation in Y can be explained by a type of tires and average speed. The rest 4% can be explained by other predictors.

–>

f. Predict the average tread wear per 160 km in percentage of tread thickness for a car with type A with the average speed 100 km/hour from the model selected in part (d).

Hide

```
tires=read.csv("c:/Users/thunt/OneDrive - University of Calgary/dataset603/tires.CSV", header = TRUE)
fav_stats(tires$ave)
```

min	Q1	median	Q3	max	mean	sd	n	missing
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>
80	101.5	105.5	109.5	113	103.3	9.105062	140	0

1 row

Hide

```
interacmodel<-lm(wear~factor(type)+ave+factor(type)*ave,data=tires)
newdata = data.frame(type="A", ave=100)
predict(interacmodel,newdata,interval="predict")
```

<!– With 95% confidence interval, for a car that has type A with an average speed 100 km/hour, the average tread wear per 160 km of tread thickness is between 0.4263475 % to 0.5525725 %. We obtained this result from the command predict(interacmodel,newdata,interval="predict") (see R code above).

–>

Problem 2. A team of mental health researchers wishes to compare three methods (A, B, and C) of treating severe depression. They would also like to study the relationship between age and treatment effectiveness as well as the interaction (if any) between age and treatment.

Each member of a simple random sample of 36 patients, comparable with respect to diagnosis and severity of depression, was randomly assigned to receive treatment A, B, or C. The data are given in **MentalHealth.csv**.

Answer the following questions

a. Which is the dependent variable (the response variable)?

<!–It's treatment effectiveness called EFFECT in the dataset.

–>

b. What are the independent variables (the predictors)?

<!– They are treatment methods (A, B and C) and age.

–>

c. Draw a scatter diagram of the sample data with EFFECT on the y-axis and AGE on the x-axis using different symbols/colors for each of the three treatments. Briefly summarize the visualization. [Hint: Check MLR part II under Interaction Effect in MLR with both Quantitative and Qualitative Variable models].

Hide

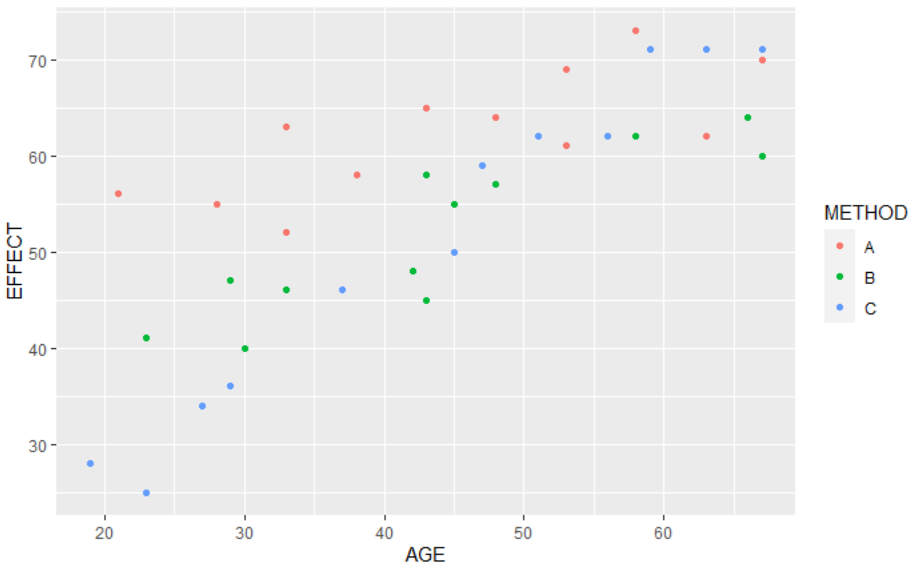
```
library(ggplot2)
Healthdata=read.csv("c:/Users/thunt/OneDrive - University of Calgary/dataset603/MentalHealth.CSV", header = TRUE)
head(Healthdata,5)
```

	EFFECT <int>	AGE <int>	METHOD <chr>
1	56	21	A
2	41	23	B
3	40	30	B
4	28	19	C
5	55	28	A

5 rows

Hide

```
ggplot(data=Healthdata,mapping=aes(x=AGE,y=EFFECT,colour=METHOD))+geom_point()
```



<!– The scatter plot shows that there may be a positive relationship between AGE and treatment effectiveness.Treatment A effectiveness seems to perform better than both treatments B and C.

–>

d. Is there any interaction between age and treatment? Test the hypothesis at $\alpha = 0.05$.

Hide

```
intermodel=lm(EFFECT~AGE+factor(METHOD)+AGE*factor(METHOD),data = Healthdata)
summary(intermodel)
```

```
Call:
lm(formula = EFFECT ~ AGE + factor(METHOD) + AGE * factor(METHOD),
    data = Healthdata)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4366 -2.7637  0.1887  2.9075  6.5634

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          47.51559    3.82523  12.422 2.34e-13 ***
AGE                   0.33051    0.08149   4.056 0.000328 ***
factor(METHOD)B     -18.59739    5.41573  -3.434 0.001759 **
factor(METHOD)C     -41.30421    5.08453  -8.124 4.56e-09 ***
AGE:factor(METHOD)B   0.19318    0.11660   1.657 0.108001
AGE:factor(METHOD)C   0.70288    0.10896   6.451 3.98e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.925 on 30 degrees of freedom
Multiple R-squared:  0.9143,	Adjusted R-squared:  0.9001
F-statistic: 64.04 on 5 and 30 DF,  p-value: 4.264e-15
```

<!– From summary(intermodel) (see R code above), we can conclude that at least one interaction term is significant between a treatment and age.

–>

e. From part (d), write the final model for predicting the treatment effectiveness.

<!–

$$TreatmentEffectiveness_i = \begin{cases} \beta_0 + \beta_1 Age + \epsilon & \text{if the person received treatment A} \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_4) Age + \epsilon & \text{if the person received treatment B} \\ (\beta_0 + \beta_3) + (\beta_1 + \beta_5) Age + \epsilon & \text{if the person received treatment C} \end{cases}$$

Substitutes regression coefficient values into the sub-models;

$$

$$\widehat{TreatmentEffectiveness_i} = \begin{cases} 47.51559 + 0.33051 Age & \text{if the person received treatment A} \\ (47.51559 - 18.59739) + (0.33051 + 0.19318) Age & \text{if the person received treatment B} \\ (47.51559 - 41.30421) + (0.33051 + 0.70288) Age & \text{if the person received treatment C} \end{cases}$$

$$

Note! As mentioned in class, we still keep all the interaction terms, although some levels are not significant.

–>

f. Interpret the effect of treatment from the sub-models in part e)

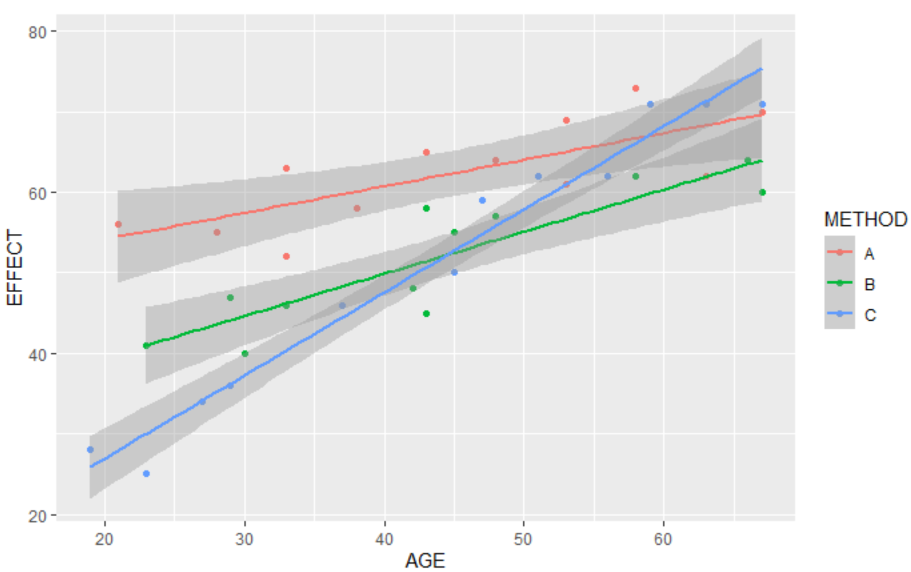
The output shows that the slope of AGE for patients with treatment C is 0.33051+0.70288=1.03339 and with treatment B is 0.52369 and with treatment A is 0.33051, suggesting that older patients are associated with better treatment effectiveness for treatment C as compared to treatment A. i.e treatment C are better than treatment A and B for older patients. Treatment A and B do not differ greatly with respect to their slopes, but their y intercepts are considerably different.

–>

g. Plot the three regression lines on the scatter diagram obtained in part (c). May one have the same conclusion as in part (f)?

Hide

```
ggplot(data=Healthdata,mapping= aes(x=AGE,y=EFFECT,colour=METHOD))+geom_point()+geom_smooth(method='lm')
```



<!–The Figure above contains the scatter diagram of the original data along with the regression lines for the three treatments. Visual inspection shows that treatment A and B do not differ greatly with respect to their slopes, but their y-intercepts are considerably different.

The graph suggests that treatment C is better than B and A for older patients and worst for younger patients. We have the same conclusions with question e.

–>

Problem 3. **Collusive bidding in road construction**. Road construction contracts in the state of Florida are awarded on the basis of competitive, sealed bids; the contractor who submits the lowest bid price wins the contract. During the 1980s, the Office of the Florida Attorney General (FLAG) suspected numerous contractors of practicing bid collusion (i.e., setting the winning bid price above the fair, or competitive, price in order to increase proect margin). By comparing the bid prices (and other important bid variables) of the fixed (or rigged) contracts to the competitively bid contracts, FLAG was able to establish invaluable benchmarks for detecting future bid-rigging. FLAG collected data for 279 road construction contracts. For each contract, the following variables shown below were measured and are only considered for this problem.

1. Price of contract ($) bid by lowest bidder, LOWBID.
2. Department of Transportation (DOT) engineer's estimate of fair contract price ($), DOTEST.
3. Status of contract (1 if fixed, 0 if competitive), STATUS

4. District (1, 2, 3, 4, or 5) in which construction project is located, DISTRICT.
5. Number of bidders on contract, NUMIDS.
6. Estimated number of days to complete work, DAYSEST.
7. Length of road project (miles), RDLNGTH.
8. Percentage of costs allocated to liquid asphalt, PCTASPH.
9. Percentage of costs allocated to base material, PCTBASE.
10. Percentage of costs allocated to excavation, PCTEXCAV.
11. Percentage of costs allocated to mobilization, PCTMOBIL.
12. Percentage of costs allocated to structures, PCTSTRUC.
13. Percentage of costs allocated to trafic control, PCTTRAF.

The data are saved in the file named **FLAG2.txt**

a. Consider building a model for the low-bid price (Y). Apply **Stepwise Regression Procedure with pent=0.05 and prem=0.1** to the data to find the independent variables most suitable for modeling Y.

Hide

```
library(olsrr)
```

```
package 拶拖olsrr拶怍 was built under R version 4.1.3Registered S3 method overwritten by 'data.table':
  method           from
  print.data.table

Attaching package: 拶拖olsrr拶怍

The following object is masked from 拶拖package:datasets拶怍:

    rivers
```

Hide

```
flag=read.table("c:/Users/thunt/OneDrive - University of Calgary/dataset603/flag2.txt", header = TRUE)
str(flag)
```

```
'data.frame':   279 obs. of  15 variables:
 $ LOWBID  : int  362916 152056 239665 1559368 144062 1187104 23665 169766 1082174 433153 ...
 $ DOTEST  : int  385963 175396 194650 1925307 252925 1573451 32538 175947 1085868 545262 ...
 $ LBERATIO: num  0.94 0.867 1.231 0.81 0.57 ...
 $ STATUS  : int  0 1 1 0 0 0 0 1 0 0 ...
 $ DISTRICT: int  1 1 1 1 1 1 5 5 1 5 ...
 $ NUMIDS  : int  3 3 3 10 8 5 7 4 6 5 ...
 $ DAYSEST : int  100 75 65 250 90 230 60 125 400 230 ...
 $ RDLNGTH : num  7.2 0 0.206 3.6 23.7 2.6 0.3 2.4 0 0 ...
 $ PCTASPH : num  0.6264 0.1533 0.0827 0.1899 0.3162 ...
 $ PCTBASE : num  0 0.0124 0 0.2988 0.2453 ...
 $ PCTEXCAV: num  0.0915 0.1422 0.1053 0.2382 0.1599 ...
 $ PCTMOBIL: num  0.01998 0.04735 0.04924 0.01154 0.00347 ...
 $ PCTSTRUC: num  0.1168 0.018 0.2226 0.1662 0.0826 ...
 $ PCTTRAF : num  0.0917 0.14764 0.06029 0.04054 0.00882 ...
 $ SUBCONT : chr  "0" "0" "0" "0" ...
```

Hide

```
step=ols_step_both_p(fullmodel, pent =0.05, prem =0.1, details=FALSE)
# final model
summary(step$model)
```

```
Call:
lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
    data = l)

Residuals:
    Min      1Q  Median      3Q     Max 
-10.899  -3.791  -1.548   2.810  23.609 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   7.0149     5.9767   1.174    0.244    
weight       64.6481     8.2714   7.816 3.67e-11 ***
fiber        -2.2078     0.4810  -4.590 1.86e-05 ***
fat           8.0881     0.9052   8.936 3.09e-13 ***
carbo         1.4409     0.2938   4.905 5.73e-06 ***
sugars        1.2756     0.2791   4.570 2.01e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.039 on 71 degrees of freedom
Multiple R-squared:  0.8781,    Adjusted R-squared:  0.8695 
F-statistic: 102.3 on 5 and 71 DF,  p-value: < 2.2e-16
```

<!– Using the stepwise regression, the final model contains only three variables and the model can be written as

$$\widehat{LOWBID} = 57105.97 + 0.9374269 DOTEST + 95252.39 STATUS - 15353.82 NUMIDS$$

–>

b. Consider building a model for the low-bid price (Y). Apply **Forward Regression Procedure with pent=0.05** :*ols_step_forward_p(fullmodel,pent=0.05)* to the data to find the independent variables most suitable for modeling Y.

Hide

```
forw=ols_step_forward_p(fullmodel,pent=0.05, details=FALSE)
# final model
summary(forw$model)
```

```
Call:
lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
    data = l)

Residuals:
    Min      1Q  Median      3Q     Max 
-10.899  -3.791  -1.548   2.810  23.609 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   7.0149     5.9767   1.174    0.244    
weight       64.6481     8.2714   7.816 3.67e-11 ***
fiber        -2.2078     0.4810  -4.590 1.86e-05 ***
fat           8.0881     0.9052   8.936 3.09e-13 ***
carbo         1.4409     0.2938   4.905 5.73e-06 ***
sugars        1.2756     0.2791   4.570 2.01e-05 ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.039 on 71 degrees of freedom
Multiple R-squared:  0.8781,    Adjusted R-squared:  0.8695 
F-statistic: 102.3 on 5 and 71 DF,  p-value: < 2.2e-16
```

<!– Using the forward regression procedure method, we obtained the same model as in a. with the stepwise method –>

c. Consider building a model for the low-bid price (Y). Apply **Backward Regression Procedure with prem=0.05** :*ols_step_backward_p(fullmodel,prem=0.05)* to the data to find the independent variables most suitable for modeling Y.

Hide

```
backw=ols_step_backward_p(fullmodel, prem =0.05, details=FALSE)
# final model
summary(backw$model)
```

```
Call:
lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
    data = l)

Residuals:
    Min      1Q  Median      3Q     Max 
-10.899  -3.791  -1.548   2.810  23.609 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   7.0149     5.9767   1.174    0.244    
fat           8.0881     0.9052   8.936 3.09e-13 ***
fiber        -2.2078     0.4810  -4.590 1.86e-05 ***
carbo         1.4409     0.2938   4.905 5.73e-06 ***
sugars        1.2756     0.2791   4.570 2.01e-05 ***
weight       64.6481     8.2714   7.816 3.67e-11 ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.039 on 71 degrees of freedom
Multiple R-squared:  0.8781,	Adjusted R-squared:  0.8695 
F-statistic: 102.3 on 5 and 71 DF,  p-value: < 2.2e-16
```

<!– The backward methods also gives the same model. –>

d. Test the individual t-test at $\alpha = 0.05$ to evaluate the variables in the model. What predictors should be kept in the model.

Hide

```
fullmodel<-lm(LOWBID~DOTEST+factor(STATUS)+factor(DISTRICT)+NUMIDS+DAYSEST+RDLNGTH+PCTASPH+PCTBASE+PCTEXCAV+PCTMOBIL+PCTSTRU
C+PCTTRAF , data =flag)
summary(fullmodel)
```

```

Call:
lm(formula = LOWBID ~ DOTEST + factor(STATUS) + factor(DISTRICT) +
    NUMIDS + DAYSEST + RDLNGTH + PCTASPH + PCTBASE + PCTEXCAV +
    PCTMOBIL + PCTSTRUC + PCTTRAF, data = flag)

Residuals:
     Min       1Q   Median       3Q      Max
-2061552   -76832     3703    68246  1592629

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        7.623e+04  6.916e+04   1.102   0.2714
DOTEST             9.362e-01  1.687e-02  55.494   <2e-16
factor(STATUS)1    1.089e+05  4.263e+04   2.554   0.0112
factor(DISTRICT)2  7.773e+04  6.388e+04   1.217   0.2248
factor(DISTRICT)3  2.960e+04  2.042e+05   0.145   0.8849
factor(DISTRICT)4 -2.729e+05  1.377e+05  -1.982   0.0485
factor(DISTRICT)5 -2.420e+04  3.799e+04  -0.637   0.5248
NUMIDS            -2.243e+04  8.797e+03  -2.550   0.0114
DAYSEST            8.030e+01  1.848e+02   0.434   0.6643
RDLNGTH            5.669e+03  4.926e+03   1.151   0.2509
PCTASPH           -1.022e+05  7.985e+04  -1.281   0.2015
PCTBASE            2.516e+05  1.840e+05   1.367   0.1727
PCTEXCAV          -2.824e+05  1.610e+05  -1.754   0.0805
PCTMOBIL           3.322e+05  2.765e+05   1.201   0.2308
PCTSTRUC           1.459e+05  1.621e+05   0.900   0.3690
PCTTRAF           -1.002e+05  1.416e+05  -0.707   0.4800

(Intercept)
DOTEST            ***
factor(STATUS)1   *
factor(DISTRICT)2
factor(DISTRICT)3
factor(DISTRICT)4 *
factor(DISTRICT)5
NUMIDS            *
DAYSEST
RDLNGTH
PCTASPH
PCTBASE
PCTEXCAV          .
PCTMOBIL
PCTSTRUC
PCTTRAF
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 278000 on 263 degrees of freedom
Multiple R-squared:  0.978, Adjusted R-squared:  0.9768
F-statistic: 780.2 on 15 and 263 DF,  p-value: < 2.2e-16
```

<!– By testing all individual predictors from the 12 predictors in the full model, we found that DOTEST, STATUS, DISTRICT and NUMIDS predictorshould be added to the model at α=0.05. –>

e. Compare the results, parts (a)-(d). Which independent variables consistently are selected as the ‘‘best’’ predictors for the model? Write all possible additive model(s) for predicting Y. Note! proposing more than one model is acceptable.

<!– We selected consistently those three variables from parts a-d: DOTEST, STATUS and NUMIDS. However, the individual t-test method selectsin addition DISTRICT. Hence, we are considering two models: the model with and without DISTRICT.

–>

Hide

```
firstordermodel1<-lm(LOWBID~DOTEST+factor(STATUS)+NUMIDS,data=flag)
summary(firstordermodel1)
```

```
Call:
lm(formula = LOWBID ~ DOTEST + factor(STATUS) + NUMIDS, data = flag)

Residuals:
     Min       1Q   Median       3Q      Max 
-2127947   -62934    -7025    59043  1665603 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      5.711e+04  4.582e+04   1.246   0.2137    
DOTEST           9.374e-01  9.280e-03 101.011   <2e-16 ***
factor(STATUS)1  9.525e+04  4.196e+04   2.270   0.0240 *  
NUMIDS          -1.535e+04  7.530e+03  -2.039   0.0424 *  
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 281700 on 275 degrees of freedom
Multiple R-squared:  0.9764,    Adjusted R-squared:  0.9761 
F-statistic:  3792 on 3 and 275 DF,  p-value: < 2.2e-16
```

Hide

```
firstordermodel2<-lm(LOWBID~DOTEST+ factor(STATUS)+NUMIDS+factor(DISTRICT),data=flag)
summary(firstordermodel2)
```

```
Call:
lm(formula = LOWBID ~ DOTEST + factor(STATUS) + NUMIDS + factor(DISTRICT), 
    data = flag)

Residuals:
     Min       1Q   Median       3Q      Max 
-2160166   -66952    -6042    55358  1625579 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        6.050e+04  5.197e+04   1.164   0.2454
DOTEST             9.447e-01  1.002e-02  94.258   <2e-16
factor(STATUS)1    9.991e+04  4.189e+04   2.385   0.0178
NUMIDS            -1.736e+04  8.255e+03  -2.103   0.0364
factor(DISTRICT)2  7.100e+04  6.316e+04   1.124   0.2619
factor(DISTRICT)3  1.156e+04  2.038e+05   0.057   0.9548
factor(DISTRICT)4 -3.165e+05  1.336e+05  -2.370   0.0185
factor(DISTRICT)5 -1.415e+04  3.733e+04  -0.379   0.7049
                     
(Intercept)          
DOTEST            ***
factor(STATUS)1   *  
NUMIDS            *  
factor(DISTRICT)2    
factor(DISTRICT)3    
factor(DISTRICT)4 *  
factor(DISTRICT)5    
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 279700 on 271 degrees of freedom
Multiple R-squared:  0.9771,    Adjusted R-squared:  0.9765 
F-statistic:  1650 on 7 and 271 DF,  p-value: < 2.2e-16
```

<!– DISTRICT is still significant in the second model, which has an adjusted R^2 of 0.9765. Model 1 has an adjusted R^2 of 0.9761, slightly inferiorto the second model. We would prefer the second model but the first model is still acceptable as the difference between the adjusted R^2's is negligible. Hence both models can be written as:

Model 1:

$$\widehat{LOWBID} = 57105.97 + 0.9374DOTEST + 95252.39STATUS - 15353.82NUMIDS$$

Model 2:

$$\widehat{LOWBID} = 60498.36 + 0.9447DOTEST + 99908.89STATUS - 17361.3NUMIDS + 70997.36DISTRICT_2 + 11563.79DISTRICT_3 - 3$$

f. Assume that your model selected in part (e) contains the following predictors: DOTEST,STATUS,,NUMBIDS, and DISTRICT. Calculate the absolute difference in average contact bid price (by the lowest bidder) between District 1 and 4, when other predictors are held as a constant

<!–

$$\widehat{LOWBID} = 60498.36 + 0.9447DOTEST + 99908.89STATUS - 17361.3NUMIDS + 70997.36DISTRICT_2 + 11563.79DISTRICT_3 - 3$$

$$\widehat{LOWBID_i} = \begin{cases} 60498.36 + 0.9447DOTEST + 99908.89STATUS - 17361.3NUMIDS & \text{if the construction project is locatec} \\ (60498.36 + 70997.36) + 0.9447DOTEST + 99908.89STATUS - 17361.3NUMIDS & \text{if the construction project is locatec} \\ (60498.36 + 11563.79) + 0.9447DOTEST + 99908.89STATUS - 17361.3NUMIDS & \text{if the construction project is locatec} \\ (60498.36 - 316505.6) + 0.9447DOTEST + 99908.89STATUS - 17361.3NUMIDS & \text{if the construction project is locatec} \\ (60498.36 - 14151.27) + 0.9447DOTEST + 99908.89STATUS - 17361.3NUMIDS & \text{if the construction project is locatec} \end{cases}$$

$$\widehat{LOWBID_i} = \begin{cases} 60498.36 + 0.9447DOTEST + 99908.89STATUS - 17361.3NUMIDS & \text{if the construction project is located from DIST} \\ 131495.7 + 0.9447DOTEST + 99908.89STATUS - 17361.3NUMIDS & \text{if the construction project is located from DIST} \\ 72062.15 + 0.9447DOTEST + 99908.89STATUS - 17361.3NUMIDS & \text{if the construction project is located from DIST} \\ -256007.2 + 0.9447DOTEST + 99908.89STATUS - 17361.3NUMIDS & \text{if the construction project is located from DIST} \\ 46347.09 + 0.9447DOTEST + 99908.89STATUS - 17361.3NUMIDS & \text{if the construction project is located from DIST} \end{cases}$$

From the sub models above, the absolute average difference in contact bid price (by the lowest bidder) between District 1 and 4 is 316,505.6 dollars which is the absolute value of $\widehat{\beta_4}$

–>

g. Assume that your model selected in part (e) contains the following predictors: DOTEST,STATUS,,NUMBIDS, and DISTRICT. Calculate the absolute difference in average contact bid price (by the lowest bidder) between District 2 and 5, when other predictors are held as a constant

<!–

From the sub models provided in part (f), the absolute average difference in contact bid price (by the lowest bidder) between District 2 and 5 is 131495.7-46347.09 = 85148.61 dollars which is the absolute value of $\widehat{\beta_2} - \widehat{\beta_5}$

–>

h. Assume that your model selected in part (e) contains the following predictors: DOTEST,STATUS,NUMBIDS, and DISTRICT. Build the first order model with interaction terms. Write the best fit model for predicting Y.

Hide

```
intermodel1<-lm(LOWBID~(DOTEST+ factor(STATUS)+factor(DISTRICT)+NUMIDS)^2,data=flag)
summary(intermodel1)
```

```

Call:
lm(formula = LOWBID ~ (DOTEST + factor(STATUS) + factor(DISTRICT) +
    NUMIDS)^2, data = flag)

Residuals:
     Min       1Q   Median       3Q      Max
-1486446   -52732     9513    46452  1477972

Coefficients: (4 not defined because of singularities)
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                      -3.353e+04  7.480e+04  -0.448  0.65434
DOTEST                            1.097e+00  2.969e-02  36.955  < 2e-16 ***
factor(STATUS)1                  -1.199e+04  1.102e+05  -0.109  0.91342
factor(DISTRICT)2                -1.215e+04  1.653e+05  -0.073  0.94147
factor(DISTRICT)3                 9.037e+04  3.802e+05   0.238  0.81229
factor(DISTRICT)4                -1.532e+06  6.568e+05  -2.332  0.02046 *
factor(DISTRICT)5                -4.438e+04  9.666e+04  -0.459  0.64655
NUMIDS                           -4.697e+03  1.273e+04  -0.369  0.71248
DOTEST:factor(STATUS)1            9.451e-02  3.673e-02   2.573  0.01063 *
DOTEST:factor(DISTRICT)2          3.988e-02  5.577e-02   0.715  0.47518
DOTEST:factor(DISTRICT)3         -1.655e-01  5.168e-01  -0.320  0.74904
DOTEST:factor(DISTRICT)4         -2.533e-02  6.268e-02  -0.404  0.68653
DOTEST:factor(DISTRICT)5         -1.330e-01  2.870e-02  -4.636 5.64e-06 ***
DOTEST:NUMIDS                    -1.934e-02  3.603e-03  -5.367 1.77e-07 ***
factor(STATUS)1:factor(DISTRICT)2        NA         NA      NA       NA
factor(STATUS)1:factor(DISTRICT)3        NA         NA      NA       NA
factor(STATUS)1:factor(DISTRICT)4        NA         NA      NA       NA
factor(STATUS)1:factor(DISTRICT)5  7.549e+04  7.891e+04   0.957  0.33964
factor(STATUS)1:NUMIDS            1.043e+04  3.188e+04   0.327  0.74370
factor(DISTRICT)2:NUMIDS          6.114e+03  2.166e+04   0.282  0.77793
factor(DISTRICT)3:NUMIDS                 NA         NA      NA       NA
factor(DISTRICT)4:NUMIDS          1.519e+05  4.661e+04   3.260  0.00126 **
factor(DISTRICT)5:NUMIDS          2.525e+04  1.798e+04   1.404  0.16148
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 251800 on 260 degrees of freedom
Multiple R-squared:  0.9822,    Adjusted R-squared:  0.9809
F-statistic: 795.6 on 18 and 260 DF,  p-value: < 2.2e-16
```

Hide

```
# Interaction STATUS*NUMIDS is not significant
```

Hide

```
intermodel2<-lm(LOWBID~DOTEST+factor(STATUS)+NUMIDS+factor(DISTRICT)+DOTEST*factor(STATUS)+DOTEST*factor(DISTRICT)+DOTEST*NU
MIDS,data=flag)
summary(intermodel2)
```

```
Call:
lm(formula = LOWBID ~ DOTEST + factor(STATUS) + NUMIDS + factor(DISTRICT) +
    DOTEST * factor(STATUS) + DOTEST * factor(DISTRICT) + DOTEST *
    NUMIDS, data = flag)

Residuals:
     Min       1Q   Median       3Q      Max
-1679384   -45974        0    41562  1341527

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              -7.819e+04  5.593e+04  -1.398  0.16325
DOTEST                    1.063e+00  2.686e-02  39.581  < 2e-16 ***
factor(STATUS)1           4.733e+04  4.644e+04   1.019  0.30904
NUMIDS                    5.199e+03  8.894e+03   0.585  0.55934
factor(DISTRICT)2         3.730e+03  7.596e+04   0.049  0.96087
factor(DISTRICT)3        -9.752e+03  3.766e+05  -0.026  0.97936
factor(DISTRICT)4         3.075e+05  3.146e+05   0.978  0.32919
factor(DISTRICT)5         7.612e+04  4.087e+04   1.863  0.06363 .
DOTEST:factor(STATUS)1    1.115e-01  3.571e-02   3.124  0.00198 **
DOTEST:factor(DISTRICT)2  4.399e-02  5.621e-02   0.783  0.43453
DOTEST:factor(DISTRICT)3 -7.972e-02  5.198e-01  -0.153  0.87823
DOTEST:factor(DISTRICT)4 -1.194e-01  5.526e-02  -2.160  0.03168 *
DOTEST:factor(DISTRICT)5 -1.167e-01  2.777e-02  -4.203 3.60e-05 ***
DOTEST:NUMIDS            -1.560e-02  3.203e-03  -4.871 1.91e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 255400 on 265 degrees of freedom
Multiple R-squared:  0.9813,    Adjusted R-squared:  0.9804
F-statistic:  1070 on 13 and 265 DF,  p-value: < 2.2e-16
```

<!–

We built an interaction model using selected variables in part f. The final model shows that interactions DOTESTxfactor(STATUS), DOTESTxNUMIDS and DOTESTxfactor(DISTRICT) are significant.

–>

i. Compare the RMSE from the first order model in part (e),which contained DISTRICT, with the interaction model in part (h). Interpret the result.

<!–

The RMSE for model 2 obtained in part (e) is 279700 whereas the RMSE for the interaction model in part (h) is 255400, much lower.

–>

j. Find the R^2_{adj} and interpret the result from part (h).

<!–

R^2_{adj}=0.9804, which means that 98.04% of the variation of price of contract bid by lowest bidder is explained by the model.

–>

Problem 4: An author studied family caregiving in Korea of older adults with dementia. The outcome variable, caregiver burden (BURDEN), was measured by the Korean Burden Inventory (KBI) where scores ranged from 28 to 140 with higher scores indicating higher burden. The following independent variables were reported by the researchers:

1. CGAGE: caregiver age (years)
2. CGINCOME: caregiver income (Won-Korean currency)
3. CGDUR: caregiver-duration of caregiving (month)
4. ADL: total activities of daily living where low scores indicate the elderly perform activities independently.
5. MEM: memory and behavioral problems with higher scores indicating more problems.
6. COG: cognitive impairment with lower scores indicating a greater degree of cognitive impairment.
7. SOCIALSU: total score of perceived social support (25-175, higher values indicating more support). The reported data are in file **KBI.csv**.

Answer the following questions

a. Use stepwise regression (with stepwise selection) to find the ''best'' set of predictors of caregiver burden.Report all significant predictors. *[Hint: Use pent =0.1 and prem=0.3].*

Hide

```
library(olsrr) #need to install the package olsrr
KBI=read.csv("c:/Users/thunt/OneDrive - University of Calgary/dataset603/KBI.CSV", header = TRUE)
head(KBI,5)
```

	CGAGE <int>	CGINCOME <int>	CGDUR <int>	ADL <int>	MEM <int>	COG <int>	SOCIALSU <int>	BURDEN <int>
1	41	200	12	39	4	18	119	28
2	30	120	36	52	33	9	131	68
3	41	300	60	89	17	3	141	59
4	35	350	2	57	31	7	150	91
5	37	600	48	28	35	19	142	70

5 rows

Hide

```
mod=lm(BURDEN~., data=KBI)
stepmod=ols_step_both_p(mod, pent =0.1, prem =0.3, details=FALSE)
summary(stepmod$model)
```

```
Call:
lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
    data = l)

Residuals:
    Min      1Q  Median      3Q     Max
-32.672  -9.977   0.367   7.774  31.523

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 115.53922   12.36816   9.342 3.86e-15 ***
MEM           0.56612    0.10232   5.533 2.73e-07 ***
SOCIALSU     -0.49237    0.08930  -5.514 2.96e-07 ***
CGDUR         0.12168    0.06486   1.876   0.0637 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.25 on 96 degrees of freedom
Multiple R-squared:  0.4397,    Adjusted R-squared:  0.4222
F-statistic: 25.12 on 3 and 96 DF,  p-value: 4.433e-12
```

<!–

The stepwise regression method selected three important variables: MEM, SOCIALSU and CGDUR. The final model obtained from stepwiseis then

$$\widehat{BURDEN} = 115.539 + 0.566MEM - 0.49237SOCIALSU + 0.121CGDUR$$

–>

b. Use all-possible-regressions-selection to find the ''best'' predictors of caregiver burden (Cp, AIC, RMSE, Adjusted R^2). Report all significant predictors.

Hide

```
mod=lm(BURDEN~., data=KBI)
ks=ols_step_best_subset(mod, details=TRUE)
ks
```

```
                Best Subsets Regression
--------------------------------------------------------
Model Index    Predictors
--------------------------------------------------------
     1         MEM
     2         MEM SOCIALSU
     3         CGDUR MEM SOCIALSU
     4         CGDUR ADL MEM SOCIALSU
     5         CGAGE CGDUR ADL MEM SOCIALSU
     6         CGAGE CGINCOME CGDUR ADL MEM SOCIALSU
     7         CGAGE CGINCOME CGDUR ADL MEM COG SOCIALSU
--------------------------------------------------------

                                                    Subsets Regression Summary
------------------------------------------------------------------------------------------------------------------------------
----------
                    Adj.        Pred
Model    R-Square    R-Square    R-Square     C(p)        AIC         SBIC         SBC          MSEP          FPE         HSP
APC
------------------------------------------------------------------------------------------------------------------------------
----------
  1        0.2520      0.2444      0.2244    29.7076    859.4694    574.7800    867.2849    30399.8652    310.0773    3.1340
0.7785
  2        0.4192      0.4072        0.38     3.6101    836.1716    552.5296    846.5923    23850.9307    245.6375    2.4842
0.6167
  3        0.4397      0.4222      0.3865     2.1575    834.5703    551.2713    847.5962    23249.4660    241.7415    2.4468
0.6070
  4        0.4473      0.4241      0.3831     2.8795    835.2038    552.1710    850.8348    23177.8870    243.2876    2.4649
0.6108
  5        0.4511      0.4220      0.3782     4.2386    836.5114    553.7192    854.7476    23265.4605    246.5047    2.5006
0.6189
  6        0.4520      0.4166      0.3129     6.0981    838.3589    555.7577    859.2003    23482.5186    251.1226    2.5510
0.6305
  7        0.4526      0.4109      0.2989     8.0000    840.2523    557.8408    863.6989    23715.2744    255.9517    2.6043
0.6426
------------------------------------------------------------------------------------------------------------------------------
----------
AIC: Akaike Information Criteria
 SBIC: Sawa's Bayesian Information Criteria
 SBC: Schwarz Bayesian Criteria
 MSEP: Estimated error of prediction, assuming multivariate normality
 FPE: Final Prediction Error
 HSP: Hocking's Sp
 APC: Amemiya Prediction Criteria
```

<!–

Cp and AIC show that the model with 3 variables seems to be the best model as this model minimizes both Cp and AIC. Moreover, the $R^2_{adjusted}$ for this model is 0.4222, close to the maximum $R^2_{adjusted}$ (=.4240). This model contains the following three variables:CGDUR, MEM and SOCIALSU; same variables obtained from part (a).

–>

c. Compare the results, parts a-b. Which independent variables consistently are selected as the ‘’best’’ predictors? Build the first order model with interaction terms, evaluate which interation terms are significant to be added in the model, and conclude the the final model for the prediction.

<!– Variables CGDUR, MEM and SOCIALSU are consistently selected as the best predictors.Moreover, the adjusted for this model is 0.4222, which means that 42.22% of the variation in burden is explained by those three independent variables.

Hide

```
modinter=lm(BURDEN~(CGDUR+ MEM +SOCIALSU)^2 , data=KBI)
summary(modinter) ## No significant interactions
```

```
Call:
lm(formula = BURDEN ~ (CGDUR + MEM + SOCIALSU)^2, data = KBI)

Residuals:
    Min      1Q  Median      3Q     Max
-32.256  -9.544   0.419   7.832  35.226

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    98.094196  27.929492   3.512 0.000688 ***
CGDUR           0.350722   0.525520   0.667 0.506181
MEM             0.869719   0.790027   1.101 0.273793
SOCIALSU       -0.341339   0.210830  -1.619 0.108828
CGDUR:MEM       0.003782   0.004228   0.894 0.373411
CGDUR:SOCIALSU -0.002564   0.004042  -0.634 0.527485
MEM:SOCIALSU   -0.002998   0.006087  -0.492 0.623553
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.4 on 93 degrees of freedom
Multiple R-squared:  0.4459,    Adjusted R-squared:  0.4102
F-statistic: 12.47 on 6 and 93 DF,  p-value: 2.879e-10
```

From the output above, none of interaction terms are significant. Therefore,the final model for prediction is

$$\widehat{BURDEN} = 115.539 + 0.566MEM - 0.49237SOCIALSU + 0.121CGDUR$$