Sean Anselmo Assignment 2

ASSIGNMENT 2

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

Deadline: March 17 , 2024, by 11:59 pm. Submit to Gradescope.ca


```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

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.

```
tires=read.csv("https://raw.githubusercontent.com/DanikaLipman/DATA603/main/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.

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

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

(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.

(e) From the model in part (d), report the adjusted-R^2 value from the model selected and interpret its value.

(f) Predict the average tread wear per 160 km in the percentage of tread thickness for a car with type A with the average speed of 100 km/hour from the model selected in part (d) with 95% confidence.

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 **Mental-Health.csv**.

Answer the following questions

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

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

(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].

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

(e) From part (d), write the final model with sub-models for predicting the treatment effectiveness. Please ensure you substitute all regression coefficients to the models.

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

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

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 the 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 traffic control, PCTTRAF.

```
FLAG2 <- read.delim("~/603/Ass2/FLAG2.txt")
library(olsrr)
library(leaps)
```

The data are saved in the file named **FLAG2.txt**. Answer the following questions:

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

```
fullmodelFLAG = lm(LOWBID~ DOTEST+STATUS+DISTRICT+NUMIDS+DAYSEST+RDLNGTH+PCTASPH+PCTBASE+PCTTRAF+PCTSTRU

StepWiseFLAG1=ols_step_both_p(fullmodelFLAG,p_enter = 0.05, p_remove = 0.1, details=TRUE)
```

```
## Stepwise Selection Method
## -------------------------
##
## Candidate Terms:
##
## 1. DOTEST
## 2. STATUS
## 3. DISTRICT
## 4. NUMIDS
## 5. DAYSEST
## 6. RDLNGTH
## 7. PCTASPH
## 8. PCTBASE
## 9. PCTTRAF
## 10. PCTSTRUC
## 11. PCTMOBIL
##
##
## Step    => 0
## Model   => LOWBID ~ 1
## R2      => 0
##
## Initiating stepwise selection...
##
## Step      => 1
## Selected  => DOTEST
## Model     => LOWBID ~ DOTEST
## R2        => 0.975
```

```
##
## Step      => 2
## Selected  => STATUS
## Model     => LOWBID ~ DOTEST + STATUS
## R2        => 0.976
##
## Step      => 3
## Selected  => NUMIDS
## Model     => LOWBID ~ DOTEST + STATUS + NUMIDS
## R2        => 0.976
##
##
## No more variables to be added or removed.
```

ANSWER TO Q3a The significant independant variables found from the stepwise selection method are:

Department of Transportation (DOT) engineer's estimate of fair contract price ($), DOTEST. Status of contract (1 if fixed, 0 if competitive), STATUS
Number of bidders on contract, NUMIDS.

The model would be:

$$LOWBIDS = \beta_0 + \beta_1 DOTEST + \beta_2 STATUS + \beta_3 NUMIDS$$

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

```
StepWiseFLAG2=ols_step_forward_p(fullmodelFLAG,p_val=0.05, details=TRUE)
```

```
## Forward Selection Method
## ------------------------
##
## Candidate Terms:
##
## 1. DOTEST
## 2. STATUS
## 3. DISTRICT
## 4. NUMIDS
## 5. DAYSEST
## 6. RDLNGTH
## 7. PCTASPH
## 8. PCTBASE
## 9. PCTTRAF
## 10. PCTSTRUC
## 11. PCTMOBIL
##
##
## Step   => 0
## Model  => LOWBID ~ 1
## R2     => 0
##
## Initiating stepwise selection...
```

```
##
##                     Selection Metrics Table
## ---------------------------------------------------------------
## Predictor    Pr(>|t|)    R-Squared    Adj. R-Squared      AIC
## ---------------------------------------------------------------
## DOTEST        0.00000        0.975             0.975    7812.664
## DAYSEST       0.00000        0.644             0.643    8553.080
## NUMIDS        0.00000        0.101             0.097    8811.470
## PCTBASE       0.00012        0.052             0.049    8826.199
## PCTASPH       0.00015        0.051             0.047    8826.515
## PCTSTRUC      0.00031        0.046             0.043    8827.933
## DISTRICT      0.00042        0.044             0.041    8828.518
## STATUS        0.00895        0.024             0.021    8834.186
## PCTMOBIL      0.05384        0.013             0.010    8837.329
## PCTTRAF       0.07064        0.012             0.008    8837.782
## RDLNGTH       0.24828        0.005             0.001    8839.735
## ---------------------------------------------------------------
##
## Step      => 1
## Selected  => DOTEST
## Model     => LOWBID ~ DOTEST
## R2        => 0.975
##
##                     Selection Metrics Table
## ---------------------------------------------------------------
## Predictor    Pr(>|t|)    R-Squared    Adj. R-Squared      AIC
## ---------------------------------------------------------------
## STATUS        0.00041        0.976             0.976    7802.016
## NUMIDS        0.00069        0.976             0.976    7803.010
## PCTMOBIL      0.07665        0.975             0.975    7811.490
## DAYSEST       0.16263        0.975             0.975    7812.690
## PCTASPH       0.25278        0.975             0.975    7813.340
## DISTRICT      0.39253        0.975             0.975    7813.924
## PCTTRAF       0.61943        0.975             0.975    7814.415
## PCTSTRUC      0.69233        0.975             0.975    7814.506
## RDLNGTH       0.94193        0.975             0.975    7814.659
## PCTBASE       0.99710        0.975             0.975    7814.664
## ---------------------------------------------------------------
##
## Step      => 2
## Selected  => STATUS
## Model     => LOWBID ~ DOTEST + STATUS
## R2        => 0.976
##
##                     Selection Metrics Table
## ---------------------------------------------------------------
## Predictor    Pr(>|t|)    R-Squared    Adj. R-Squared      AIC
## ---------------------------------------------------------------
## NUMIDS        0.04241        0.976             0.976    7799.830
## DAYSEST       0.08699        0.976             0.976    7801.039
## PCTMOBIL      0.09348        0.976             0.976    7801.157
## PCTASPH       0.16940        0.976             0.976    7802.097
## DISTRICT      0.42927        0.976             0.976    7803.381
## PCTBASE       0.45380        0.976             0.976    7803.446
```

```
## PCTSTRUC      0.46494        0.976             0.976    7803.474
## PCTTRAF       0.71766        0.976             0.976    7803.883
## RDLNGTH       0.87116        0.976             0.976    7803.990
## ---------------------------------------------------------------
##
## Step      => 3
## Selected  => NUMIDS
## Model     => LOWBID ~ DOTEST + STATUS + NUMIDS
## R2        => 0.976
##
##                     Selection Metrics Table
## ---------------------------------------------------------------
## Predictor    Pr(>|t|)    R-Squared    Adj. R-Squared      AIC
## ---------------------------------------------------------------
## DAYSEST       0.08415        0.977             0.976    7798.787
## PCTASPH       0.10365        0.977             0.976    7799.128
## PCTMOBIL      0.12290        0.977             0.976    7799.402
## PCTBASE       0.14803        0.977             0.976    7799.696
## PCTSTRUC      0.30125        0.976             0.976    7800.740
## DISTRICT      0.33105        0.976             0.976    7800.866
## PCTTRAF       0.58374        0.976             0.976    7801.524
## RDLNGTH       0.95081        0.976             0.976    7801.826
## ---------------------------------------------------------------
##
##
## No more variables to be added.
##
## Variables Selected:
##
## => DOTEST
## => STATUS
## => NUMIDS
```

ANSWER TO Q3b The significant independant variables found from the forward selection are:

Department of Transportation (DOT) engineer's estimate of fair contract price ($), DOTEST. Status of contract (1 if fixed, 0 if competitive), STATUS
Number of bidders on contract, NUMIDS.

The model would be:

$LOWBIDS = \beta_0 + \beta_1 DOTEST + \beta_2 STATUS + \beta_3 NUMIDS$

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

```
StepWiseFLAG3=ols_step_backward_p(fullmodelFLAG,p_val=0.05, details=TRUE)
```

```
## Backward Elimination Method
## ---------------------------
##
## Candidate Terms:
##
```

```
## 1. DOTEST
## 2. STATUS
## 3. DISTRICT
## 4. NUMIDS
## 5. DAYSEST
## 6. RDLNGTH
## 7. PCTASPH
## 8. PCTBASE
## 9. PCTTRAF
## 10. PCTSTRUC
## 11. PCTMOBIL
##
##
## Step   => 0
## Model  => LOWBID ~ DOTEST + STATUS + DISTRICT + NUMIDS + DAYSEST + RDLNGTH + PCTASPH + PCTBASE + PCT
## R2     => 0.977
##
## Initiating stepwise selection...
##
## Step     => 1
## Removed  => PCTSTRUC
## Model    => LOWBID ~ DOTEST + STATUS + DISTRICT + NUMIDS + DAYSEST + RDLNGTH + PCTASPH + PCTBASE + PC
## R2       => 0.97722
##
## Step     => 2
## Removed  => PCTTRAF
## Model    => LOWBID ~ DOTEST + STATUS + DISTRICT + NUMIDS + DAYSEST + RDLNGTH + PCTASPH + PCTBASE + PC
## R2       => 0.97719
##
## Step     => 3
## Removed  => PCTBASE
## Model    => LOWBID ~ DOTEST + STATUS + DISTRICT + NUMIDS + DAYSEST + RDLNGTH + PCTASPH + PCTMOBIL
## R2       => 0.97713
##
## Step     => 4
## Removed  => DISTRICT
## Model    => LOWBID ~ DOTEST + STATUS + NUMIDS + DAYSEST + RDLNGTH + PCTASPH + PCTMOBIL
## R2       => 0.97702
##
## Step     => 5
## Removed  => RDLNGTH
## Model    => LOWBID ~ DOTEST + STATUS + NUMIDS + DAYSEST + PCTASPH + PCTMOBIL
## R2       => 0.97693
##
## Step     => 6
## Removed  => PCTASPH
## Model    => LOWBID ~ DOTEST + STATUS + NUMIDS + DAYSEST + PCTMOBIL
## R2       => 0.97681
##
## Step     => 7
## Removed  => PCTMOBIL
## Model    => LOWBID ~ DOTEST + STATUS + NUMIDS + DAYSEST
## R2       => 0.97665
##
```

```
## Step     => 8
## Removed  => DAYSEST
## Model    => LOWBID ~ DOTEST + STATUS + NUMIDS
## R2       => 0.9764
##
##
## No more variables to be removed.
##
## Variables Removed:
##
## => PCTSTRUC
## => PCTTRAF
## => PCTBASE
## => DISTRICT
## => RDLNGTH
## => PCTASPH
## => PCTMOBIL
## => DAYSEST
```

ANSWER TO Q3c The significant independant variables found from the backward selection are:

Department of Transportation (DOT) engineer's estimate of fair contract price ($), DOTEST. Status of contract (1 if fixed, 0 if competitive), STATUS
Number of bidders on contract, NUMIDS.

The model would be:

$LOWBIDS = \beta_0 + \beta_1 DOTEST + \beta_2 STATUS + \beta_3 NUMIDS$

All the models created from the bothway selection, forward, and backward are using the same significant predictors.

(d) Test the individual t-test at $\alpha = 0.05$ to evaluate the variables in the full model. What predictors should be kept in the model based on the individual t-tests from the full model?

```
fullmodelFLAG = lm(LOWBID~ DOTEST+factor(STATUS)+factor(DISTRICT)+NUMIDS+DAYSEST+RDLNGTH+PCTASPH+PCTBASE
summary(fullmodelFLAG)
```

```
##
## Call:
## lm(formula = LOWBID ~ DOTEST + factor(STATUS) + factor(DISTRICT) +
##     NUMIDS + DAYSEST + RDLNGTH + PCTASPH + PCTBASE + PCTTRAF +
##     PCTSTRUC + PCTMOBIL, data = FLAG2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2102678   -72570    -2476    68490  1637851
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         4.639e+04  6.730e+04   0.689   0.4912
## DOTEST              9.351e-01  1.692e-02  55.250   <2e-16 ***
## factor(STATUS)1     1.053e+05  4.275e+04   2.463   0.0144 *
## factor(DISTRICT)2   7.183e+04  6.404e+04   1.122   0.2630
## factor(DISTRICT)3 -1.845e+03  2.042e+05  -0.009   0.9928
```

```
## factor(DISTRICT)4 -3.053e+05  1.369e+05  -2.230   0.0266 *
## factor(DISTRICT)5 -2.021e+04  3.807e+04  -0.531   0.5960
## NUMIDS            -2.242e+04  8.832e+03  -2.538   0.0117 *
## DAYSEST            5.029e+01  1.848e+02   0.272   0.7857
## RDLNGTH            5.589e+03  4.945e+03   1.130   0.2595
## PCTASPH           -8.093e+04  7.923e+04  -1.022   0.3079
## PCTBASE            1.921e+05  1.816e+05   1.058   0.2910
## PCTTRAF           -8.414e+04  1.419e+05  -0.593   0.5536
## PCTSTRUC           1.131e+05  1.617e+05   0.700   0.4848
## PCTMOBIL           4.380e+05  2.710e+05   1.617   0.1072
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 279100 on 264 degrees of freedom
## Multiple R-squared:  0.9778, Adjusted R-squared:  0.9766
## F-statistic: 829.1 on 14 and 264 DF,  p-value: < 2.2e-16
```

ANSWER TO Q3d The significant independent variables found from the individual t-test at $\alpha = 0.05$ are:

Department of Transportation (DOT) engineer's estimate of fair contract price ($), DOTEST. Status of contract (1 if fixed, 0 if competitive), STATUS
District (1, 2, 3, 4, or 5) in which the construction project is located, DISTRICT. Number of bidders on contract, NUMIDS.

The model would be:

$LOWBIDS = \beta_0 + \beta_1 DOTEST + \beta_2 STATUS + \beta_3 DISTRICT + \beta_4 NUMIDS$

The individual t test examination at $\alpha = 0.05$ yielded the same significant predictors as the stepwise selection procedure, with the DISTRICT factor predictor included as well.

(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.

ANSWER TO Q3e The independent variables that are consistently selected are:

Department of Transportation (DOT) engineer's estimate of fair contract price ($), DOTEST. Status of contract (1 if fixed, 0 if competitive), STATUS
Number of bidders on contract, NUMIDS.

Possible additive models could be:

$LOWBIDS = \beta_0 + \beta_1 DOTEST + \beta_2 STATUS + \beta_3 NUMIDS$ $LOWBIDS = \beta_0 + \beta_1 DOTEST + \beta_2 STATUS + \beta_3 NUMIDS + \beta_4 DISTRICT$

(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.

```
partmodelFLAG = lm(LOWBID~ DOTEST+factor(STATUS)+factor(DISTRICT)+NUMIDS ,data=FLAG2)
summary(partmodelFLAG)
```

```
##
## Call:
```

```
## lm(formula = LOWBID ~ DOTEST + factor(STATUS) + factor(DISTRICT) +
##     NUMIDS, data = FLAG2)
##
## 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 *
## 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
## NUMIDS            -1.736e+04  8.255e+03  -2.103   0.0364 *
## ---
## 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
```

ANSWER TO Q3f To find the absolute difference in price between District 1 and District 4 while everything else is held constant, we have to look at the β values. Specifically, we look at the β_0 value for District 1, and the β_5 value associated with District 4.

β_0 = 6.05x10^4 β_5 = -3.165x10^5

Absolute Difference = District 4 - District 1

Absolute Difference = $(-3.165\times10 + \beta_0)$- β_0

Absolute Difference = 3.165x10^5

The absolute difference in price between District 1 and District 4 is $316,500 less for District 4.

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

ANSWER TO Q3g To find the absolute difference in price between District 2 and District 5 while everything else is held constant, we have to look at the β values. Specifically, we look at the β_3 value for District 1, and the β_6 value associated with District 4.

β_3 = 6.05x10^4 β_6 = -3.165x10^5

Absolute Difference = District 5 - District 2

Absolute Difference = (7.100e+04 + β_0)- (-1.415e+04 + β_0)

Absolute Difference = 7.100e+04 + 1.415e+04

The absolute difference in price between District 2 and District 5 is $85,150 less for District 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.

```
firstOrder = lm(LOWBID~ DOTEST+factor(STATUS)+factor(DISTRICT)+NUMIDS ,data=FLAG2)
firstOrderint = lm(LOWBID~ (DOTEST+factor(STATUS)+factor(DISTRICT)+NUMIDS)^2 ,data=FLAG2)
summary(firstOrderint)
```

```
##
## Call:
## lm(formula = LOWBID ~ (DOTEST + factor(STATUS) + factor(DISTRICT) +
##     NUMIDS)^2, data = FLAG2)
##
## 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
```

```
firstOrderintFinal = lm(LOWBID ~ DOTEST+factor(STATUS)+factor(DISTRICT)+NUMIDS+DOTEST:factor(STATUS)+DO
summary(firstOrderintFinal)
```

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

```
##     DOTEST:NUMIDS + NUMIDS:factor(DISTRICT), data = FLAG2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1489137   -50878      574    54016  1480203
##
## Coefficients: (1 not defined because of singularities)
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -7.343e+04  6.421e+04  -1.144   0.2538
## DOTEST                      1.102e+00  2.921e-02  37.729  < 2e-16 ***
## factor(STATUS)1             6.156e+04  4.652e+04   1.323   0.1869
## factor(DISTRICT)2           2.458e+04  1.613e+05   0.152   0.8790
## factor(DISTRICT)3           6.326e+04  3.785e+05   0.167   0.8674
## factor(DISTRICT)4          -1.531e+06  6.557e+05  -2.334   0.0203 *
## factor(DISTRICT)5           1.572e+04  7.240e+04   0.217   0.8283
## NUMIDS                      1.974e+02  1.181e+04   0.017   0.9867
## DOTEST:factor(STATUS)1      9.218e-02  3.580e-02   2.575   0.0106 *
## DOTEST:factor(DISTRICT)2    3.939e-02  5.566e-02   0.708   0.4798
## DOTEST:factor(DISTRICT)3   -1.326e-01  5.149e-01  -0.258   0.7970
## DOTEST:factor(DISTRICT)4   -2.532e-02  6.257e-02  -0.405   0.6861
## DOTEST:factor(DISTRICT)5   -1.335e-01  2.854e-02  -4.679 4.63e-06 ***
## DOTEST:NUMIDS              -1.995e-02  3.549e-03  -5.622 4.82e-08 ***
## factor(DISTRICT)2:NUMIDS    1.648e+03  2.119e+04   0.078   0.9381
## factor(DISTRICT)3:NUMIDS           NA         NA      NA       NA
## factor(DISTRICT)4:NUMIDS    1.513e+05  4.653e+04   3.252   0.0013 **
## factor(DISTRICT)5:NUMIDS    1.803e+04  1.589e+04   1.135   0.2575
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 251400 on 262 degrees of freedom
## Multiple R-squared:  0.9821, Adjusted R-squared:  0.981
## F-statistic:   898 on 16 and 262 DF,  p-value: < 2.2e-16
```

```
interactive = regsubsets(LOWBID ~ DOTEST+factor(STATUS)+factor(DISTRICT)+NUMIDS, data = FLAG2, nv=10)
summary(interactive)
```

```
## Subset selection object
## Call: regsubsets.formula(LOWBID ~ DOTEST + factor(STATUS) + factor(DISTRICT) +
##     NUMIDS, data = FLAG2, nv = 10)
## 7 Variables  (and intercept)
##                   Forced in Forced out
## DOTEST                FALSE      FALSE
## factor(STATUS)1       FALSE      FALSE
## factor(DISTRICT)2     FALSE      FALSE
## factor(DISTRICT)3     FALSE      FALSE
## factor(DISTRICT)4     FALSE      FALSE
## factor(DISTRICT)5     FALSE      FALSE
## NUMIDS                FALSE      FALSE
## 1 subsets of each size up to 7
## Selection Algorithm: exhaustive
##          DOTEST factor(STATUS)1 factor(DISTRICT)2 factor(DISTRICT)3
## 1  ( 1 ) "*"    " "             " "               " "
## 2  ( 1 ) "*"    "*"             " "               " "
## 3  ( 1 ) "*"    "*"             " "               " "
```

```
## 4  ( 1 ) "*"    "*"             " "               " "
## 5  ( 1 ) "*"    "*"             "*"               " "
## 6  ( 1 ) "*"    "*"             "*"               " "
## 7  ( 1 ) "*"    "*"             "*"               "*"
##          factor(DISTRICT)4 factor(DISTRICT)5 NUMIDS
## 1  ( 1 ) " "               " "               " "
## 2  ( 1 ) " "               " "               " "
## 3  ( 1 ) "*"               " "               " "
## 4  ( 1 ) "*"               " "               "*"
## 5  ( 1 ) "*"               " "               "*"
## 6  ( 1 ) "*"               "*"               "*"
## 7  ( 1 ) "*"               "*"               "*"
```

```
reg.summary=summary(interactive)
rsquare=c(reg.summary$rsq)
cp=c(reg.summary$cp)
AdjustedR=c(reg.summary$adjr2)
RSS=c(reg.summary$rss)
BIC=c(reg.summary$bic)
cbind(rsquare,cp,BIC,RSS,AdjustedR)
```

```
##        rsquare        cp       BIC          RSS AdjustedR
## [1,] 0.9749298 21.372688 -1017.153 2.317768e+13 0.9748393
## [2,] 0.9760410 10.236969 -1024.170 2.215041e+13 0.9758674
## [3,] 0.9766582  4.940499 -1025.820 2.157979e+13 0.9764035
## [4,] 0.9769179  3.870001 -1023.311 2.133967e+13 0.9765810
## [5,] 0.9770632  4.152627 -1019.441 2.120536e+13 0.9766431
## [6,] 0.9770758  6.003219 -1013.964 2.119368e+13 0.9765701
## [7,] 0.9770761  8.000000 -1008.336 2.119342e+13 0.9764840
```

ANSWER FOR Q3 h The best fit model for predicting Y includes interaction terms:

$Y = \beta_0 + \beta_1 DOTEST + \beta_2 STATUS + \beta_3 NUMIDS + \beta_4 DISTRICT + \beta_1\beta_2 DOTEST * STATUS + \beta_1\beta_4 DOTEST * DISTRICT + \beta_1\beta_3 DOTEST * NUMIDS + \beta_3\beta_4 NUMIDS * DISTRICT$

(i) Compare the RSE from the first-order model in part (d) with the interaction model in part (h). Interpret the result.

ANSWER FOR Q3 i The RSE from the interactive model is 251400, and the RSE from the additive model is 279700. The RSE for the interactive model is lower, which indicates the interactive model makes esitamtes that more closely fit the actual data.

The interactive model having a lower RSE than the additive indicates it is the better fit model.

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

ANSWER FOR Q3 j The R^2_{adj} from the interactive model is 0.981, and the R^2_{adj} from the additive model is 0.9765. This means for our interactive model, 98.1% of the variance is explained by the model, whereas in the additive model only 97.65% of the variance is explained by the model.

The interactive model having a higher R^2_{adj} than the additive indicates it is the better fit 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 the file **KBI.csv**.

Answer the following questions:

```
KBI <- read.csv("~/603/Ass2/KBI.csv")
```

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

```
library(olsrr)

#First is stepwise
fullmodel = lm(BURDEN~CGAGE+CGINCOME+CGDUR+ADL+MEM+COG+SOCIALSU, data = KBI)
StepWise=ols_step_both_p(fullmodel,p_enter = 0.1, p_remove = 0.3, details=TRUE)
```

```
## Stepwise Selection Method
## -------------------------
##
## Candidate Terms:
##
## 1. CGAGE
## 2. CGINCOME
## 3. CGDUR
## 4. ADL
## 5. MEM
## 6. COG
## 7. SOCIALSU
##
##
## Step   => 0
## Model  => BURDEN ~ 1
## R2     => 0
##
## Initiating stepwise selection...
##
## Step      => 1
## Selected  => MEM
## Model     => BURDEN ~ MEM
## R2        => 0.252
```

```
##
## Step      => 2
## Selected  => SOCIALSU
## Model     => BURDEN ~ MEM + SOCIALSU
## R2        => 0.419
##
## Step      => 3
## Selected  => CGDUR
## Model     => BURDEN ~ MEM + SOCIALSU + CGDUR
## R2        => 0.44
##
##
## No more variables to be added or removed.
```

Answer to Question 4 a Our stepwise regression with stepwise selection has yielded significant predictors in:

1. MEM (Memory)
2. SOCIALSU (Social support)
3. CGDUR (Caregiver duration)

The best set of predictors are: MEM+SOCIALSU+CGDUR. This model has an Adjusted R^2 equal to 0.422. This model has an AIC of 834.570.

(b) Use all-possible-regressions-selection to find the ''best" predictors of caregiver burden. Which model would you pick based on AIC? Adjusted R^2? Report all significant predictors.

```
#All Possible Selection
AllPossible=ols_step_all_possible(fullmodel,p_enter = 0.1, p_remove = 0.3, details=TRUE)

min_adjr <- min(AllPossible$result$adjr)
model_with_min_adjr <- AllPossible$result[AllPossible$result$adjr == min_adjr, ]

min_aic <- min(AllPossible$result$aic)
model_with_min_aic <- AllPossible$result[AllPossible$result$aic == min_aic, ]
```

Answer to Question 4 b The significant predictors and model based on Adjusted R^2:

1. MEM (Memory)
2. SOCIALSU (Social support)
3. CGDUR (Caregiver duration)
4. ADL (Total Daily Living Activities)

The all possible selection based on Adjusted R^2 chose the significant predictors of MEM+SOCIALSU+CGDUR+ADL.This model has an Adjusted R^2 equal to 0.424063346 This selection included ADL (Daily Living Activities) that was not included in the stepwise selection method.

The significant predictors and model based on AIC:

1. MEM (Memory)
2. SOCIALSU (Social support)
3. CGDUR (Caregiver duration)

This all possible selection based on AIC chose the significant predictors of MEM+SOCIALSU+CGDUR.The AIC for this model is 834.5703. This selection method found the same significant predictors than the stepwise selection method.

(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 based on these predictors, evaluate which interaction terms are significant to be added in the model, and conclude the the final model for the prediction.

```
intmodel1 = lm(BURDEN~CGDUR+MEM+SOCIALSU+MEM:SOCIALSU, data = KBI)
summary(intmodel1)
```

```
##
## Call:
## lm(formula = BURDEN ~ CGDUR + MEM + SOCIALSU + MEM:SOCIALSU,
##     data = KBI)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.408 -10.185   0.184   7.955  31.763
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  108.898122  24.226251   4.495 1.96e-05 ***
## CGDUR          0.125379   0.066181   1.894   0.0612 .
## MEM            0.802229   0.746483   1.075   0.2852
## SOCIALSU      -0.444259   0.175348  -2.534   0.0129 *
## MEM:SOCIALSU  -0.001751   0.005483  -0.319   0.7502
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.32 on 95 degrees of freedom
## Multiple R-squared:  0.4403, Adjusted R-squared:  0.4168
## F-statistic: 18.69 on 4 and 95 DF,  p-value: 2.331e-11
```

```
intmodel2 = lm(BURDEN~CGDUR+MEM+SOCIALSU+MEM:CGDUR, data = KBI)
summary(intmodel2)
```

```
##
## Call:
## lm(formula = BURDEN ~ CGDUR + MEM + SOCIALSU + MEM:CGDUR, data = KBI)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.719  -9.319   0.197   7.876  32.713
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 118.483739  13.544169   8.748 7.81e-14 ***
## CGDUR         0.055020   0.138843   0.396  0.69279
## MEM           0.506303   0.150521   3.364  0.00111 **
## SOCIALSU     -0.500376   0.090828  -5.509 3.08e-07 ***
```

```
## CGDUR:MEM     0.001980   0.003642   0.544  0.58800
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.3 on 95 degrees of freedom
## Multiple R-squared:  0.4415, Adjusted R-squared:  0.4179
## F-statistic: 18.77 on 4 and 95 DF,  p-value: 2.122e-11
```

```
intmodel3 = lm(BURDEN~CGDUR+MEM+SOCIALSU+SOCIALSU:CGDUR, data = KBI)
summary(intmodel3)
```

```
##
## Call:
## lm(formula = BURDEN ~ CGDUR + MEM + SOCIALSU + SOCIALSU:CGDUR,
##     data = KBI)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.792  -9.206  -0.281   8.051  32.003
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    108.106833  20.825496   5.191 1.19e-06 ***
## CGDUR            0.344045   0.504345   0.682  0.49680
## MEM              0.574792   0.104587   5.496 3.26e-07 ***
## SOCIALSU        -0.439034   0.149771  -2.931  0.00423 **
## CGDUR:SOCIALSU  -0.001629   0.003663  -0.445  0.65761
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.31 on 95 degrees of freedom
## Multiple R-squared:  0.4409, Adjusted R-squared:  0.4174
## F-statistic: 18.73 on 4 and 95 DF,  p-value: 2.225e-11
```

Answer to Question 4 c The independent variables that are consistently selected are MEM, SOCIALSU, and CGDUR. The Adjusted R^2 selection method also yielded ADL. The model created by the stepwise selection is the model of choice and will be proceeding with this one for interaction terms. The lowest AIC score model also agreed with the stepwise selection method.

Based on the interaction models, no interaction terms based on the selected predictors are significant. The final mode is:

Burden = MEM + SOCIALSU + CGDUR
