Assignment Two

Maria Delgado

2024-03-14

```
library(olsrr)

## Warning: package 'olsrr' was built under R version 4.3.3

## 
## Attaching package: 'olsrr'

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

library(ggplot2)

                                     ##QUESTION ONE

##A
tires <- read.csv("C:/Users/camil/OneDrive/Desktop/Data 603/Assignment Two/ti
res.csv")
head(tires)

##   type wear ave
## 1    A  0.3  80
## 2    A  0.3  80
## 3    A  0.3  80
## 4    A  0.3  80
## 5    A  0.3  80
## 6    A  0.3  80

full_model_tires = lm(wear~factor(type)+ave, data = tires)
summary(full_model_tires)

## 
## 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

#both predictors, type and ave, have p-values much smaller than the typical s
ignificance level of 0.05. This indicates that both predictors are statistica
lly significant in explaining the variability in the response variable, wear.
#SubEquations:
#When type is A: Wear_hat=-0.64450834+0.01130937×aveWear_hat=-0.64450834+0.01
130937×ave
#When type is B: Wear_hat=(-0.64450834+0.17250064)+0.01130937×aveWear_hat=(-0
.64450834+0.17250064)+0.01130937×ave
#sub-equations are:
#Wear_hat = -0.64450834 + 0.01130937 * ave, when type is A
#Wear_hat = -0.4720077 + 0.01130937 * ave, when type is B

##B
#The variable "type" is represented as a dummy variable, which is a categoric
al predictor with two levels. Specifically, when the type is "A", the dummy v
ariable factor(type) takes a value of 0, and when the type is "B", the dummy
variable factor(type) takes a value of 1.

##C

#Beta 0: The estimated average wear when both the type of tire and average sp
eed are at their reference levels.
#Beta 1: The estimated average difference in wear between type A and type B t
ires.
#Beta 2: The estimated average percentage change in wear for a one-unit incre
ase in average speed (per 1 km/h), holding all other variables constant

##D
tires_interaction_model = lm(wear~factor(type)+ave+factor(type)*ave, data = t
ires)
summary(tires_interaction_model)

##
## Call:
## lm(formula = wear ~ factor(type) + ave + factor(type) * ave,
##     data = tires)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.070158 -0.016493 -0.003643  0.024086  0.063703
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -0.3888744  0.0347705  -11.18   <2e-16 ***
## factor(type)B     -1.0800050  0.0779442  -13.86   <2e-16 ***
## ave                0.0087833  0.0003415   25.72   <2e-16 ***
## factor(type)B:ave  0.0119840  0.0007439   16.11   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03169 on 136 degrees of freedom
## Multiple R-squared:  0.9608, Adjusted R-squared:   0.96
## F-statistic:  1112 on 3 and 136 DF,  p-value: < 2.2e-16
```

```
#The individual coefficients test (t-test) indicates that all variables, including the introduced interaction term, are statistically significant. Their corresponding t-values substantially exceed the significance level of 5%. Additionally, the p-values associated with each coefficient are extremely small, providing strong evidence against the null hypothesis that the respective coefficients are equal to zero. We can confidently reject the null hypothesis and accept the alternative, suggesting that the beta coefficient for the interaction term is not equal to zero.

#New estimated first order model:
#General equation:
#Wear_hat =  -0.388874431 - 1.080004985 * factor(type) + 0.008783344 * ave + 0.011984013 * factor(type) * ave
#Sub-models:
#Wear_hat =  -0.388874431 + 0.008783344 * ave #0 when type is A
#Wear_hat =  -1.468879 + 0.02076736 * ave  #1 when type is B

cat("The base model has adjusted r-squared =",
    summary(full_model_tires)$adj.r.squared,
    "\n",
    "The model with the interaction term has adjusted r-squared =",
    summary(tires_interaction_model)$adj.r.squared,
    "\n\n")
```

```
## The base model has adjusted r-squared = 0.8844236
##  The model with the interaction term has adjusted r-squared = 0.9599663
```

```
## The base model has adjusted r-squared = 0.8844236
## The model with the interaction term has adjusted r-squared = 0.9599663
cat("The base model has RMSE =",
    sigma(full_model_tires),
    "\n",
    "Model with the interaction term has RMSE =",
    sigma(tires_interaction_model))
```

```
## The base model has RMSE = 0.05383824
##  Model with the interaction term has RMSE = 0.03168614
```

```
## The base model has RMSE = 0.05383824
## The model with the interaction term has RMSE = 0.03168614
```

```
#Based on these results, I would recommend for the model with the interaction
term. This choice is supported by the fact that the adjusted R-squared of thi
s model is higher than that of the base model, indicating a better fit. Addit
ionally, the lower RMSE of the model with the interaction term further streng
thens its suitability for prediction purposes.

##E
summary(tires_interaction_model)$adj.r.squared

## [1] 0.9599663

#The adjusted R-squared value of 0.9599663 indicates that approximately 96% o
f the variation in the response variable, water tread wear, is explained by t
he model, after accounting for the number of predictors in the model.

##F
# From this equation: Wear_hat =  -0.388874431 + 0.008783344 * ave
wear_predicted = -0.388874431 + 0.008783344 * 100
#The average tread wear per 160 km is predicted to be approximately 0.48946 p
ercentage points of tread thickness for a car with type A and an average spee
d of 100 km/hour."

                                      ##QUESTION TWO

MentalHealth <- read.csv("C:/Users/camil/OneDrive/Desktop/Data 603/Assignment
Two/MentalHealth.csv")
head(MentalHealth)

##   EFFECT AGE METHOD
## 1     56  21      A
## 2     41  23      B
## 3     40  30      B
## 4     28  19      C
## 5     55  28      A
## 6     25  23      C

##Dependent/response variable is EFFECT.

##B The independent variables (the predictors) are AGE and METHOD

##C
Health_model = lm(EFFECT~AGE+factor(METHOD),data=MentalHealth)
summary(Health_model)

##
## Call:
## lm(formula = EFFECT ~ AGE + factor(METHOD), data = MentalHealth)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.5732  -3.3922   0.9829   3.9613   9.5062
##
```

```
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        32.54335    3.58105   9.088 2.23e-10 ***
## AGE                 0.66446    0.06978   9.522 7.42e-11 ***
## factor(METHOD)B    -9.80758    2.46471  -3.979 0.000371 ***
## factor(METHOD)C   -10.25276    2.46542  -4.159 0.000224 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.035 on 32 degrees of freedom
## Multiple R-squared:  0.784,  Adjusted R-squared:  0.7637
## F-statistic: 38.71 on 3 and 32 DF,  p-value: 9.287e-11
```

```r
#General Equation: EFFECT = 32.5433481 + 0.6644606 * AGE - 9.8075777 * factor(METHOD)B - 10.2527575 * factor(METHOD)C

#Sub Equations:
coefficients <- coef(Health_model)

sub_eq_A <- paste("EFFECT =", coefficients[1], "+", coefficients[2], "* AGE")
sub_eq_B <- paste("EFFECT =", coefficients[1] + coefficients[3], "+", coefficients[2], "* AGE")
sub_eq_C <- paste("EFFECT =", coefficients[1] + coefficients[4], "+", coefficients[2], "* AGE")
print(sub_eq_A)
```

```
## [1] "EFFECT = 32.5433481144758 + 0.66446063685184 * AGE"
```

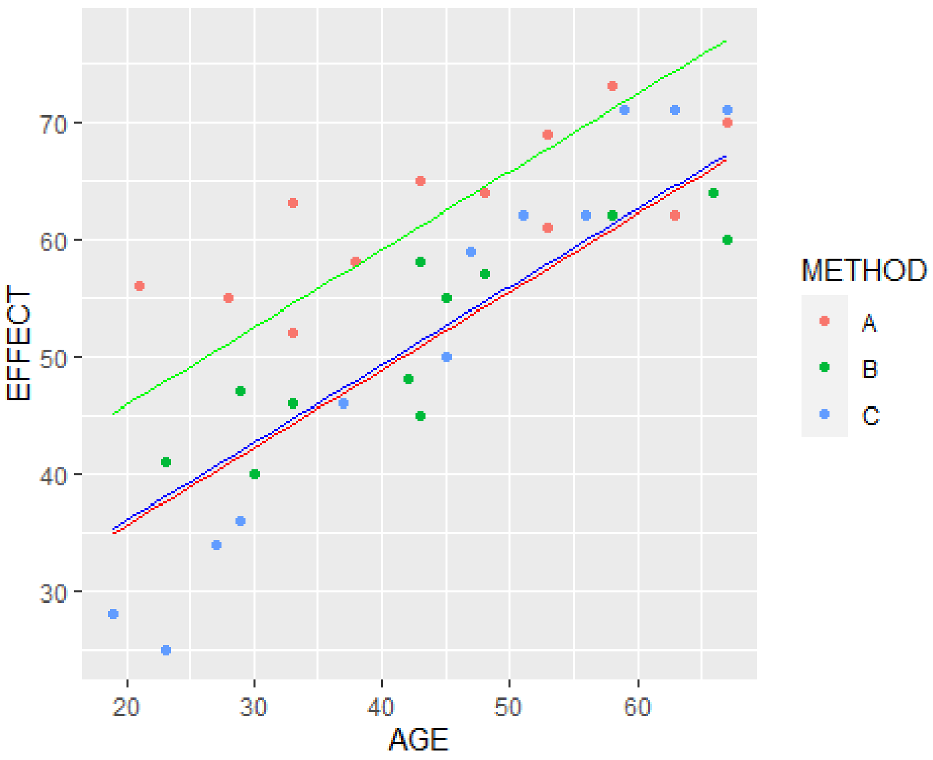
```r
print(sub_eq_B)
```

```
## [1] "EFFECT = 22.7357703649234 + 0.66446063685184 * AGE"
```

```r
print(sub_eq_C)
```

```
## [1] "EFFECT = 22.2905905772073 + 0.66446063685184 * AGE"
```

```r
effect_A <- function(x) { 32.5433481144758 + 0.66446063685184 * x }
effect_B <- function(x) { 22.7357703649234 + 0.66446063685184 * x  }
effect_C <- function(x) { 22.2905905772073 + 0.66446063685184 * x }

ggplot(data = MentalHealth, aes(x = AGE, y = EFFECT, colour = METHOD)) +
  geom_point() +
  stat_function(fun = effect_A, geom = "line", color = 'green') +
  stat_function(fun = effect_B, geom = "line", color = 'blue') +
  stat_function(fun = effect_C, geom = "line", color = 'red')
```



```
##In the plot depicted above, it's evident that three parallel lines are fitted to the data: one representing method A, another for method B, and the third for method C. These lines exhibit distinct intercepts but maintain the same slope. The parallel nature of these lines signifies that the average effect on the outcome (EFFECT) due to a one-year increase in age remains consistent across all treatment groups.

##D. Null Hypothesis (H0): The interaction term between AGE and METHOD doesn’t exist, meaning the coefficient of the interaction term is zero. Alternative Hypothesis (Ha): There is an interaction between AGE and METHOD.

Health_model_interaction = lm(EFFECT~AGE+factor(METHOD)+AGE*factor(METHOD),data=MentalHealth)
summary(Health_model_interaction)

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

```
anova(Health_model, Health_model_interaction)
```

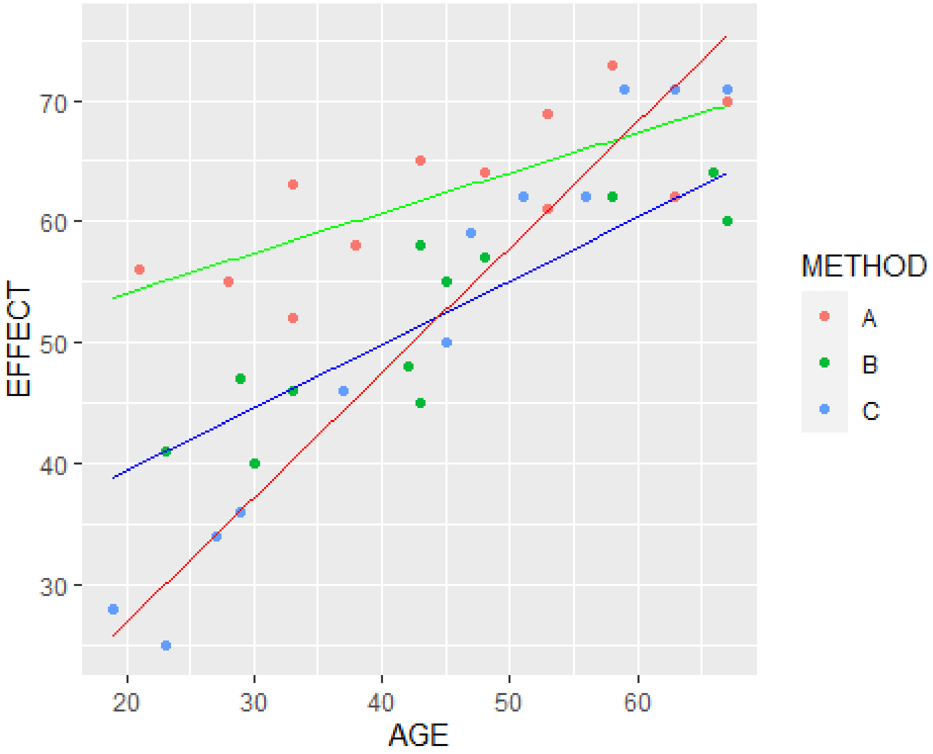
```
## Analysis of Variance Table
##
## Model 1: EFFECT ~ AGE + factor(METHOD)
## Model 2: EFFECT ~ AGE + factor(METHOD) + AGE * factor(METHOD)
##   Res.Df     RSS Df Sum of Sq      F   Pr(>F)
## 1     32 1165.57
## 2     30  462.15  2    703.43 22.831 9.41e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#The coefficients associated with the interaction terms were examined. It was
found that only the interaction term involving METHOD C was statistically sig
nificant, suggesting a potential interaction effect specifically for this tre
atment method. Patterns observed in the residuals indicated that including th
e interaction term improved the model's ability to capture variation in the d
ata, supporting the decision to include it in the analysis. Partial F-Test (P
artial ANOVA): A statistical test confirmed the necessity of including the in
teraction term in the model. The low p-value (< 0.05) provided strong evidenc
e against the null hypothesis, indicating that the interaction between age an
d treatment method significantly influences the outcome

##E
Health_model_interaction = lm(EFFECT~AGE+factor(METHOD)+AGE*factor(METHOD),da
ta=MentalHealth)
summary(Health_model_interaction)
```

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

```
#Final Model: EFFECT = 47.5155913 + 0.3305073 * AGE - 18.5973852 * factor(MET
HOD)B - 41.3042101 * factor(METHOD)C + 0.1931769 * AGE * factor(METHOD)B + 0.
7028836 * AGE * factor(METHOD)C

#Sub-equations:
#1) when METHOD is A, then factor(METHOD)B = 0, factor(METHOD)C = 0
#step1:  EFFECT = 47.5155913 + 0.3305073 * AGE - 18.5973852 * 0 - 41.3042101
* 0 + 0.1931769 * AGE * 0 + 0.7028836 * AGE * 0
#2) when METHOD is B, then factor(METHOD)B = 1, factor(METHOD)C = 0
#step1:  EFFECT = 47.5155913 + 0.3305073 * AGE - 18.5973852 * 1 - 41.3042101
* 0 + 0.1931769 * AGE * 1 + 0.7028836 * AGE * 0
#step2:  EFFECT = (47.5155913 - 18.5973852) + AGE * (0.3305073 + 0.1931769)
#3) when METHOD is C, then factor(METHOD)B = 0, factor(METHOD)C = 1
#step1:  EFFECT = 47.5155913 + 0.3305073 * AGE - 18.5973852 * 0 - 41.3042101
* 1 + 0.1931769 * AGE * 0 + 0.7028836 * AGE * 1
#step2:  EFFECT = 47.5155913 + 0.3305073 * AGE - 41.3042101 + 0.7028836 * AGE
#step3:  EFFECT = (47.5155913 - 41.3042101) + (0.3305073  + 0.7028836) * AGE

#Sub-equations:
#1) when METHOD is A
#EFFECT = 47.5155913 + 0.3305073 * AGE
#2) when METHOD is B
#EFFECT = 28.91821 + 0.5236842 * AGE
#3) when METHOD is C
#EFFECT = 6.211381 + 1.033391 * AGE

##F
#For the treatment_c sub-model: On average, for every one-year increase in ag
e, the EFFECT for patients receiving treatment C is expected to increase by 1
.03339 units.
#For the treatment_b sub-model: On average, for every one-year increase in ag
e, the EFFECT for patients receiving treatment B is expected to increase by 0
.52369 units.
#For the treatment_a sub-model: On average, for every one-year increase in ag
e, the EFFECT for patients receiving treatment A is expected to increase by 0
.33051 units.
```

```
##G
effect_inter_A = function (x){47.5155913 + 0.3305073 * x}
effect_inter_B = function (x){28.91821 + 0.5236842 * x}
effect_inter_C = function (x){6.211381 + 1.033391 * x}

ggplot(data=MentalHealth,mapping= aes(x=AGE,y=EFFECT,colour=METHOD)) +
  geom_point() +
  stat_function(fun=effect_inter_A,geom="line",color='green') +
  stat_function(fun=effect_inter_B,geom="line",color='blue') +
  stat_function(fun=effect_inter_C,geom="line",color='red')
```



```
#It's notable that these regression lines exhibit distinct intercepts, repres
ented by the coefficients for METHOD C compared to those for METHODS A and B.
Moreover, they also showcase varying slopes, indicating differences in the im
pact of AGE on EFFECT across the different treatment groups. This suggests th
e potential for changes in AGE to influence the EFFECT of patients undergoing
different treatment methods. It's worth highlighting that the most substantia
l increase in EFFECT is observed for treatment C compared to treatments A and
B.

                                           ##QUESTION THREE

##A
library(readr)
data <- read_delim("FLAG2.txt", delim = "\t")
```

```
## Rows: 279 Columns: 15
## ── Column specification ────────────────────────────────────────────────────────
──────
## Delimiter: "\t"
## chr  (1): SUBCONT
## dbl (14): LOWBID, DOTEST, LBERATIO, STATUS, DISTRICT, NUMIDS, DAYSEST, RDL
NG...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this m
essage.
```

```
data$STATUS <- factor(data$STATUS)
data$DISTRICT <- factor(data$DISTRICT)

full_model <- lm(LOWBID ~ DOTEST + STATUS + DISTRICT + NUMIDS + DAYSEST + RDL
NGTH +
                   PCTASPH + PCTBASE + PCTEXCAV + PCTMOBIL + PCTSTRUC + PCTTR
AF, data = data)
stepmod <- ols_step_both_p(full_model, p_enter = 0.05, p_remove = 0.1, detail
s = TRUE)
```

```
## Stepwise Selection Method
## -------------------------
##
## Candidate Terms:
##
## 1. DOTEST
## 2. STATUS
## 3. DISTRICT
## 4. NUMIDS
## 5. DAYSEST
## 6. RDLNGTH
## 7. PCTASPH
## 8. PCTBASE
## 9. PCTEXCAV
## 10. PCTMOBIL
## 11. PCTSTRUC
## 12. PCTTRAF
##
##
## 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.
```

```
summary(stepmod$model)
```

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## 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 ***
## STATUS1      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
```

```
#The valid model is: LOWBID_hat = 5.710597e+04 + 9.374269e-01 * DOTEST + 9.52
5239e+04 * factor(STATUS)1 - 1.535382e+04 * NUMIDS

##B
forward_model = ols_step_forward_p(full_model, p_val = 0.05, details = FALSE)
summary(forward_model$model)
```

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

```
## 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 ***
## STATUS1      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
```

```
#LOWBID_hat = 5.710597e+04 + 9.374269e-01 * DOTEST + 9.525239e+04 * factor(ST
ATUS)1 - 1.535382e+04 * NUMIDS

##C
backward_model = ols_step_backward_p(full_model, 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. PCTEXCAV 
## 10. PCTMOBIL 
## 11. PCTSTRUC 
## 12. PCTTRAF 
## 
## 
## Step   => 0 
## Model  => LOWBID ~ DOTEST + STATUS + DISTRICT + NUMIDS + DAYSEST + RDLNGTH 
+ PCTASPH + PCTBASE + PCTEXCAV + PCTMOBIL + PCTSTRUC + PCTTRAF 
## R2     => 0.978 
## 
## Initiating stepwise selection... 
## 
```

```
## Step      => 1
## Removed   => DAYSEST
## Model     => LOWBID ~ DOTEST + STATUS + DISTRICT + NUMIDS + RDLNGTH + PCTAS
PH + PCTBASE + PCTEXCAV + PCTMOBIL + PCTSTRUC + PCTTRAF
## R2        => 0.978
##
## Step      => 2
## Removed   => PCTTRAF
## Model     => LOWBID ~ DOTEST + STATUS + DISTRICT + NUMIDS + RDLNGTH + PCTAS
PH + PCTBASE + PCTEXCAV + PCTMOBIL + PCTSTRUC
## R2        => 0.97796
##
## Step      => 3
## Removed   => PCTSTRUC
## Model     => LOWBID ~ DOTEST + STATUS + DISTRICT + NUMIDS + RDLNGTH + PCTAS
PH + PCTBASE + PCTEXCAV + PCTMOBIL
## R2        => 0.97785
##
## Step      => 4
## Removed   => RDLNGTH
## Model     => LOWBID ~ DOTEST + STATUS + DISTRICT + NUMIDS + PCTASPH + PCTBA
SE + PCTEXCAV + PCTMOBIL
## R2        => 0.97774
##
## Step      => 5
## Removed   => PCTMOBIL
## Model     => LOWBID ~ DOTEST + STATUS + DISTRICT + NUMIDS + PCTASPH + PCTBA
SE + PCTEXCAV
## R2        => 0.97764
##
## Step      => 6
## Removed   => DISTRICT
## Model     => LOWBID ~ DOTEST + STATUS + NUMIDS + PCTASPH + PCTBASE + PCTEXC
AV
## R2        => 0.97702
##
## Step      => 7
## Removed   => PCTBASE
## Model     => LOWBID ~ DOTEST + STATUS + NUMIDS + PCTASPH + PCTEXCAV
## R2        => 0.97687
##
## Step      => 8
## Removed   => PCTEXCAV
## Model     => LOWBID ~ DOTEST + STATUS + NUMIDS + PCTASPH
## R2        => 0.97663
##
## Step      => 9
## Removed   => PCTASPH
## Model     => LOWBID ~ DOTEST + STATUS + NUMIDS
## R2        => 0.9764
```

```
##
##
## No more variables to be removed.
##
## Variables Removed:
##
## => DAYSEST
## => PCTTRAF
## => PCTSTRUC
## => RDLNGTH
## => PCTMOBIL
## => DISTRICT
## => PCTBASE
## => PCTEXCAV
## => PCTASPH
```

```
summary(backward_model$model)
```

```
##
## Call:
## lm(formula = paste(response, "~", paste(c(include, cterms), collapse = " +
")),
##     data = l)
##
## 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 ***
## STATUS1      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
```

```
#The valid model is: LOWBID_hat = 5.711e+04  + 9.374e-01 * DOTEST + 9.525e+04
* factor(STATUS)1 - 1.535e+04 * NUMIDS

##D
summary(full_model)
```

```
##
## Call:
## lm(formula = LOWBID ~ DOTEST + STATUS + DISTRICT + NUMIDS + DAYSEST +
##     RDLNGTH + PCTASPH + PCTBASE + PCTEXCAV + PCTMOBIL + PCTSTRUC +
##     PCTTRAF, data = data)
##
## 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 ***
## STATUS1      1.089e+05  4.263e+04   2.554   0.0112 *
## DISTRICT2    7.773e+04  6.388e+04   1.217   0.2248
## DISTRICT3    2.960e+04  2.042e+05   0.145   0.8849
## DISTRICT4   -2.729e+05  1.377e+05  -1.982   0.0485 *
## DISTRICT5   -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
## ---
## 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
```

```
#The results of the individual t-tests indicate that the predictors DOTEST, S
TATUS, and NUMIDS exhibit statistical significance at the 5% significance lev
el, implying they should be retained in the model. Additionally, one of the f
our levels of the DISTRICT variable demonstrates significance. Thus, I am emp
loying a partial ANOVA test to determine whether to retain it.
#The null hypothesis (H0) is that the beta coefficient for DISTRICT is zero,
while the alternative hypothesis (Ha) suggests it is non-zero.
anova(full_model, lm(LOWBID~DOTEST+factor(STATUS)+NUMIDS+DAYSEST+RDLNGTH+PCTA
SPH+PCTBASE+PCTEXCAV+PCTMOBIL+PCTSTRUC+PCTTRAF, data = data))
```

```
## Analysis of Variance Table
##
## Model 1: LOWBID ~ DOTEST + STATUS + DISTRICT + NUMIDS + DAYSEST + RDLNGTH
+
##     PCTASPH + PCTBASE + PCTEXCAV + PCTMOBIL + PCTSTRUC + PCTTRAF
## Model 2: LOWBID ~ DOTEST + factor(STATUS) + NUMIDS + DAYSEST + RDLNGTH +
##     PCTASPH + PCTBASE + PCTEXCAV + PCTMOBIL + PCTSTRUC + PCTTRAF
##   Res.Df        RSS Df   Sum of Sq      F Pr(>F)
## 1    263 2.0321e+13
## 2    267 2.0865e+13 -4 -5.4447e+11 1.7617 0.1369

#Since I did not find sufficient evidence to reject the null hypothesis that
the coefficient for DISTRICT is zero, I have opted not to incorporate it into
the model.
#With valid predictors:
model_valid = lm(LOWBID~DOTEST+factor(STATUS)+NUMIDS, data = data)
summary(model_valid)

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

#The valid model: LOWBID_hat = 5.711e+04  + 9.374e-01 * DOTEST + 9.525e+04 *
factor(STATUS)1 - 1.535e+04 * NUMIDS

##E
#Stepwise: LOWBID_hat = 5.710597e+04 + 9.374269e-01 * DOTEST + 9.525239e+04 *
factor(STATUS)1 - 1.535382e+04 * NUMIDS
#Forward: LOWBID_hat = 5.710597e+04 + 9.374269e-01 * DOTEST + 9.525239e+04 *
factor(STATUS)1 - 1.535382e+04 * NUMIDS
#Backward: LOWBID_hat = 5.711e+04 + 9.374e-01 * DOTEST + 9.525e+04 * factor(S
TATUS)1 - 1.535e+04 * NUMIDS
#Modelpartd: LOWBID_hat = 5.711e+04 + 9.374e-01 * DOTEST + 9.525e+04 * factor
(STATUS)1 - 1.535e+04 * NUMIDS
#DOTEST, STATUS and NUMIDS appear in all four models
#My proposed model is:LOWBID_hat = 5.710597e+04 + 9.374269e-01 * DOTEST + 9.5
25239e+04 * factor(STATUS)1 - 1.535382e+04 * NUMIDS

##F
model_f = lm(LOWBID~DOTEST+factor(STATUS)+factor(DISTRICT)+NUMIDS, data = dat
a)
summary(model_f, data = data)
```

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

```
#Model is:
#LOWBID = 6.049836e+04 + 9.447389e-01 * DOTEST + 9.990889e+04 * factor(STATUS
)1 + 7.099736e+04 * factor(DISTRICT)2 + 1.156379e+04 * factor(DISTRICT)3 - 3.
165056e+05 * factor(DISTRICT)4 - 1.415127e+04 * factor(DISTRICT)5- 1.736130e+
04 * NUMIDS
#Given the property that when DISTRICT is 1, all factor(DISTRICT)i = 0, and w
hen DISTRICT is 4, factor(DISTRICT)4 = 1, else 0, we can find the difference
in the average contract bid price (by the lowest bidder) between District 1 a
nd 4 when other predictors are held constant.
#Difference_1_minus_4 = 0 - (- 3.165056e+05 * 1) = 316505.6
#Therefore, the difference in average contract bid price (by the lowest bidde
r) between District 1 and 4, when other predictors are held constant, is $316
,505.6.

##G
#LOWBID = 6.049836e+04 + 9.447389e-01 * DOTEST + 9.990889e+04 * factor(STATUS
)1 + 7.099736e+04 * factor(DISTRICT)2 + 1.156379e+04 * factor(DISTRICT)3 - 3.
165056e+05 * factor(DISTRICT)4 - 1.415127e+04 * factor(DISTRICT)5- 1.736130e+
04 * NUMIDS
#When DISTRICT is 2 then factor(DISTRICT)2 = 1. Also, When DISTRICT is 5 then
factor(DISTRICT)5 = 1.
#Step 1:
#Difference_5_minus_2 = 6.049836e+04 + 9.447389e-01 * DOTEST + 9.990889e+04 *
0 + 7.099736e+04 * 0 + 1.156379e+04 * 0 - 3.165056e+05 * 0 - 1.415127e+04 * 1
```

```
- 1.736130e+04 * NUMIDS -(6.049836e+04 + 9.447389e-01 * DOTEST + 9.990889e+04
* 0 + 7.099736e+04 * 1 + 1.156379e+04 * 0 - 3.165056e+05 * 0 - 1.415127e+04 *
0 - 1.736130e+04 * 0)
#Step 2:
#Difference_5_minus_2 =  - 1.415127e+04 - 7.099736e+04 = -85148.63
#The difference in average contact bid price = $85148.63

##H
interaction_model = lm(LOWBID~(DOTEST+factor(STATUS)+factor(DISTRICT)+NUMIDS)
^2, data = data)
summary(interaction_model)

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

```
interaction_model_refined = lm(LOWBID~DOTEST+factor(STATUS)+NUMIDS+DOTEST*NUM
IDS+DOTEST*factor(STATUS), data = data)
summary(interaction_model_refined)
```

```
##
## Call:
## lm(formula = LOWBID ~ DOTEST + factor(STATUS) + NUMIDS + DOTEST *
##     NUMIDS + DOTEST * factor(STATUS), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1677680   -36672     6590    38551  1496732
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -3.245e+04  5.014e+04  -0.647 0.518009
## DOTEST                   1.017e+00  2.527e-02  40.237  < 2e-16 ***
## factor(STATUS)1          3.743e+04  4.765e+04   0.785 0.432900
## NUMIDS                   1.593e+03  8.171e+03   0.195 0.845535
## DOTEST:NUMIDS           -1.206e-02  3.142e-03  -3.839 0.000154 ***
## DOTEST:factor(STATUS)1   1.106e-01  3.639e-02   3.038 0.002612 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 264800 on 273 degrees of freedom
## Multiple R-squared:  0.9793, Adjusted R-squared:  0.9789
## F-statistic:  2583 on 5 and 273 DF,  p-value: < 2.2e-16
```

```
#model with interactions is: LOWBID = -3.245094e+04 + 1.01689 * DOTEST + 3.74
2610e+04 * factor(STATUS) + 1.593291e+03 * NUMIDS - 1.206275e-02 * DOTEST * N
UMIDS + 1.105697e-01 * DOTEST * factor(STATUS)

##I
cat("part d model) has RMSE =",
    sigma(model_valid),
    "\n",
    "The model with the interaction term has RMSE =",
    sigma(interaction_model_refined))
```

```
## part d model) has RMSE = 281686.7
##  The model with the interaction term has RMSE = 264767
```

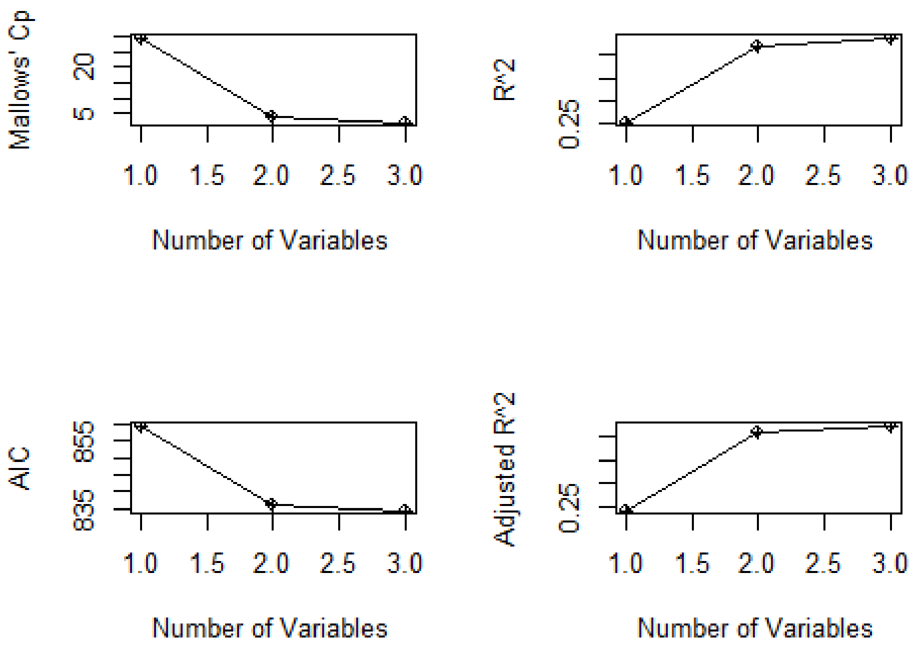
```
#The model with the interaction terms exhibits a lower RMSE as shown in the output above. This RMSE, which amounts to $264,767, represents the standard deviation of the unexplained variance in bid prices

##J
cat("The model with the interaction term has adjusted r-squared =",
    summary(interaction_model_refined)$adj.r.squared,
    "\n\n")

## The model with the interaction term has adjusted r-squared = 0.9789205

## Approximately 97.9% of the total variation in the response variable LOWBID can be accounted for by the regression model.

                                           ##QUESTION FOUR

##A
KBI <- read.csv("C:/Users/camil/OneDrive/Desktop/Data 603/Assignment Two/KBI.csv")
colnames(KBI)

## [1] "CGAGE"    "CGINCOME" "CGDUR"    "ADL"      "MEM"      "COG"      "SOCIALSU"
## [8] "BURDEN"

base_model = lm(BURDEN~CGAGE+CGINCOME+CGDUR+ADL+MEM+COG+SOCIALSU, data=KBI)
kbi_stepmodel = ols_step_both_p(base_model, p_enter = 0.1, p_remove = 0.3, details=FALSE)
summary(kbi_stepmodel$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
```

```
##From the output: The valid model is: BURDEN = 115.5392230 + 0.5661203 * MEM
- 0.4923699 * SOCIALSU

##B
kbi_subset = ols_step_best_subset(base_model, details=TRUE)
kbi_subset

##                 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 S
ummary
## --------------------------------------------------------------------------
------------------------------------------------------------
##                        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    574.78
00    867.2849    30399.8652    310.0773    3.1340    0.7785
##   2        0.4192      0.4072        0.38     3.6101    836.1716    552.52
96    846.5923    23850.9307    245.6375    2.4842    0.6167
##   3        0.4397      0.4222      0.3865     2.1575    834.5703    551.27
13    847.5962    23249.4660    241.7415    2.4468    0.6070
##   4        0.4473      0.4241      0.3831     2.8795    835.2038    552.17
10    850.8348    23177.8870    243.2876    2.4649    0.6108
##   5        0.4511      0.4220      0.3782     4.2386    836.5114    553.71
92    854.7476    23265.4605    246.5047    2.5006    0.6189
##   6        0.4520      0.4166      0.3129     6.0981    838.3589    555.75
77    859.2003    23482.5186    251.1226    2.5510    0.6305
##   7        0.4526      0.4109      0.2989     8.0000    840.2523    557.84
08    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
```

```r
base_model = lm(BURDEN~CGAGE+CGINCOME+CGDUR+ADL+MEM+COG+SOCIALSU, data=KBI)
kbi_stepmodel = ols_step_both_p(base_model, pent = 0.1, prem = 0.3, details=FALSE)
kbisubsets_metrics <- kbi_stepmodel$metrics
print(kbisubsets_metrics)
```

```
##   step variable   method        r2    adj_r2       aic       sbc      sbic
## 1    1      MEM addition 0.2519944 0.2443617 859.4694 867.2849 574.7800
## 2    2 SOCIALSU addition 0.4191848 0.4072092 836.1716 846.5923 552.5296
## 3    3    CGDUR addition 0.4397292 0.4222207 834.5703 847.5962 551.2713
##   mallows_cp     rmse
## 1  29.707640 17.26029
## 2   3.610120 15.20949
## 3   2.157489 14.93807
```

```r
# Extracting metrics from kbisubsets_metrics dataframe
rsquare <- kbisubsets_metrics$r2
AIC <- kbisubsets_metrics$aic
AdjustedR <- kbisubsets_metrics$adj_r2

# Plotting the metrics
par(mfrow=c(2,2)) # split the plotting panel into a 2 x 2 grid
plot(kbisubsets_metrics$mallows_cp, type = "o", pch = 10, xlab = "Number of Variables", ylab = "Mallows' Cp")
plot(rsquare, type = "o", pch = 10, xlab = "Number of Variables", ylab = "R^2")
plot(AIC, type = "o", pch = 10, xlab = "Number of Variables", ylab = "AIC")
plot(AdjustedR, type = "o", pch = 10, xlab = "Number of Variables", ylab = "Adjusted R^2")
```

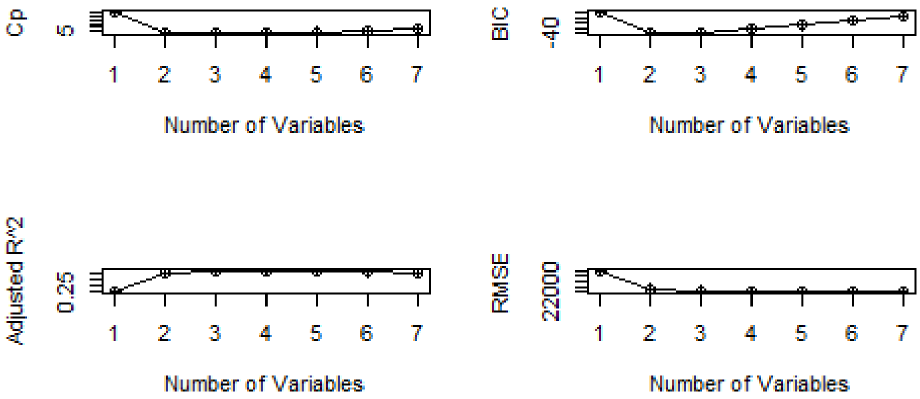


```
library("leaps")

## Warning: package 'leaps' was built under R version 4.3.3

best.subset<-regsubsets(BURDEN~CGAGE+CGINCOME+CGDUR+ADL+MEM+COG+SOCIALSU, dat
a= KBI, nv=10 )
summary(best.subset)

## Subset selection object
## Call: regsubsets.formula(BURDEN ~ CGAGE + CGINCOME + CGDUR + ADL +
##     MEM + COG + SOCIALSU, data = KBI, nv = 10)
## 7 Variables  (and intercept)
##          Forced in Forced out
## CGAGE        FALSE      FALSE
## CGINCOME     FALSE      FALSE
## CGDUR        FALSE      FALSE
## ADL          FALSE      FALSE
## MEM          FALSE      FALSE
## COG          FALSE      FALSE
## SOCIALSU     FALSE      FALSE
## 1 subsets of each size up to 7
## Selection Algorithm: exhaustive
##          CGAGE CGINCOME CGDUR ADL MEM COG SOCIALSU
## 1  ( 1 ) " "   " "      " "   " " "*" " " " "
## 2  ( 1 ) " "   " "      " "   " " "*" " " "*"
## 3  ( 1 ) " "   " "      "*"   " " "*" " " "*"
## 4  ( 1 ) " "   " "      "*"   "*" "*" " " "*"
## 5  ( 1 ) "*"   " "      "*"   "*" "*" " " "*"
## 6  ( 1 ) "*"   "*"      "*"   "*" "*" " " "*"
## 7  ( 1 ) "*"   "*"      "*"   "*" "*" "*" "*"
```

```
reg.summary = summary(best.subset)
rsquare = c(reg.summary$rsq)
cp = c(reg.summary$cp)
AdjustedR = c(reg.summary$adjr2)
RMSE = c(reg.summary$rss)
BIC = c(reg.summary$bic)
cbind(rsquare,cp,BIC,RMSE,AdjustedR)
```

```
##        rsquare        cp       BIC     RMSE AdjustedR
## [1,] 0.2519944 29.707640 -19.82415 29791.75 0.2443617
## [2,] 0.4191848  3.610120 -40.51675 23132.85 0.4072092
## [3,] 0.4397292  2.157489 -39.51282 22314.60 0.4222207
## [4,] 0.4473335  2.879523 -36.27420 22011.73 0.4240633
## [5,] 0.4511470  4.238638 -32.36144 21859.85 0.4219527
## [6,] 0.4519831  6.098124 -27.90873 21826.55 0.4166272
## [7,] 0.4525670  8.000000 -23.41016 21803.29 0.4109145
```

```
par(mfrow=c(3,2)) # split the plotting panel into a 3 x 2 grid
plot(reg.summary$cp,type = "o",pch=10, xlab="Number of Variables",ylab= "Cp")
plot(reg.summary$bic,type = "o",pch=10, xlab="Number of Variables",ylab= "BIC
")
plot(reg.summary$adjr2,type = "o",pch=10, xlab="Number of Variables",ylab= "A
djusted R^2")
plot(reg.summary$rss,type = "o",pch=10, xlab="Number of Variables",ylab= "RMS
E")
```



```
#Based on the displayed results, in Model 3, the following observations can b
e made: Cp exhibits the lowest value, while the Adjusted R-squared metric ran
ks as the second best. Additionally, AIC reflects the lowest value. However,
RMSE does not demonstrate the best performance. Model 3 includes the predicto
rs CGDUR, MEM, and SOCIALSU.

##C
#After evaluating the methods, MEM and SOCIALSU emerged as significant variab
les at a 5% significance level in the stepwise model outlined in part a). Con
sidering metrics like adj R-squared, Bc, and AIC, I opted for CGDUR, MEM, and
SOCIALSU. Notably, MEM and SOCIALSU are common factors across both assessment
s.
#Interaction Model:Null Hypothesis (H0): The beta coefficients of interaction
terms are zero, indicating no interaction between variables. Alternative Hypo
thesis (Ha): The beta coefficients of interaction terms are not zero, suggest
ing an interaction between variables.

kbi_interaction_model = lm(BURDEN~(CGDUR+MEM+SOCIALSU)^2, data = KBI)
summary(kbi_interaction_model)

##
## 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

#Based on the provided output, there are no interaction terms that are statistically significant at the 5% significance level. Consequently, we fail to reject the null hypothesis, indicating the absence of interactions among the primary predictors.

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

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

#CGDUR falls within a marginal area. I've chosen to exclude it from the model.

##Final Model:
summary(lm(BURDEN~MEM+SOCIALSU, data = KBI))
```

```
## 
## Call:
## lm(formula = BURDEN ~ MEM + SOCIALSU, data = KBI)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -33.884 -11.173  -0.331   8.723  35.091 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 116.07291   12.52448   9.268 5.12e-15 ***
## MEM           0.59941    0.10207   5.872 6.02e-08 ***
## SOCIALSU     -0.47552    0.08999  -5.284 7.76e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 15.44 on 97 degrees of freedom
## Multiple R-squared:  0.4192, Adjusted R-squared:  0.4072 
## F-statistic:    35 on 2 and 97 DF,  p-value: 3.596e-12

#BURDEN = 116.07291 + 0.59941 * MEM - 0.47552 * SOCIALSU.
```