# DATA 603 HW 2

Kane Smith

2022-11-16

## Contents



## Problem 1

```r
# Read in CSV file
tires=read.csv("tires.csv", header = TRUE)
```

### a)

```r
# Fitting the full model
tires_full <- lm(wear~., data=tires)
summary(tires_full)
```

```
##
## Call:
## lm(formula = wear ~ ., 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 ***
## typeB        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
```

Conducting an individual t-test:

$H_0 : \beta_i = 0$

$H_a : \beta_i \neq 0$

$ i = (type_B, ave)$

Using an alpha value of 0.05, all of parameters are statistically significant by the Individual t-test. Our estimated best fit model is as follows:

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

### b)

The only categorical variable in the data set is "type".

```
levels(factor(tires$type))
```

```
## [1] "A" "B"
```

```
summary(tires_full)
```

```
##
## Call:
## lm(formula = wear ~ ., 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 ***
## typeB        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
```

There are two levels to the categorical variable "type". Type A tires and type B tires.

Based on the summary of our full model, the dummy variable is type_B and has a coefficient of 0.1725006.

### c)

$\beta_0$ (Intercept): The average tread wear per 160km of type A tires when the average speed is 0 km/hr. This value is -0.6445083%, which does not make sense for interpretation since having negative tread wear is not possible.

$\beta_{type_B}$: The average difference in tread wear per 160km between type A and type B tires. This value is 0.1725006 which means that type B tires will have 0.1725006% more tread wear on average compared to type A tires.

$\beta_0 + \beta_{type_B}$ The average tread wear per 160km of type B tires when the average speed is 0 km/hr. This value is -0.4720077% which also does not make sense for interpretation since having negative tread wear is not possible.

$\beta_{ave}$:The amount the average tread wear per 160km increases when the average speed increases by 1 km/hr. This value is 0.0113094 which means that for an increase in average speed of 1 km/hr, the average tread wear per 160km will increase by 0.0113094%.

### d)

Our best fit additive model contains all variables (type and ave). Building an interaction model with all of our variables:

```
tires_interaction <- lm(wear~(type+ave)^2, data=tires)
summary(tires_interaction)
```

```
##
## Call:
## lm(formula = wear ~ (type + ave)^2, 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 ***
## typeB       -1.0800050  0.0779442  -13.86   <2e-16 ***
## ave          0.0087833  0.0003415   25.72   <2e-16 ***
## typeB: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
```

Doing an individual t-test:

$H_0 : \beta_i = 0$

$H_a : \beta_i \neq 0$

$(i = (ave * type_B)$

We see that all terms in our model are statistically significant with a p-value <2e-16, therefore we should keep all terms in our model. The adjusted R-squared value increases from 0.8844 in our full model with no interaction terms to 0.96 in our full model with interaction terms. This means we can say that 96% of the variance in the response variable, average tread wear, can be explained with our new model with interaction terms. We see a large increase in our adjusted R-squared when adding the interaction term into our model.

Therefore, the model I would suggest for predicting y (tread wear) is: $\widehat{wear} = -0.3888744 + -1.0800050 type_B + 0.0087833 ave + 0.0119840(type_B * ave)$

### e)

```
summary(tires_interaction)
```

```
##
## Call:
## lm(formula = wear ~ (type + ave)^2, 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 ***
## typeB       -1.0800050  0.0779442  -13.86   <2e-16 ***
## ave          0.0087833  0.0003415   25.72   <2e-16 ***
## typeB: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 adjusted R-squared value from our model in d) is 0.96. This means that 96% of the variance in the response variable, average tread wear, can be explained with our new model with interaction terms.

### f)

Predicting average tread wear using the following values:

**ave:** 100 km/hr

**tire type:** Type A

```
-0.3888744 + -1.0800050*(0) + 0.0087833*(100) + 0.0119840*(0*100)
```

```
## [1] 0.4894556
```

We get a predicted average tread wear of 0.4894556% per 160km. To ensure we can trust this value, we should make sure that we are not extrapolating by checking that the value we used for ave (100 km/hr) is within the range of data we used to fit our model.

```
favstats(tires$ave, data=tires)[c("min","max")]
```

```
##  min max
##   80 113
```

100 km/hr is within the range of our data of 80-113 so therefore we can be sure we were not extrapolating and can trust the average tread wear that we predicted.

### Problem 2

```r
# Read in CSV file
mental_health <- read.csv("MentalHealth.csv")
```

#### a)

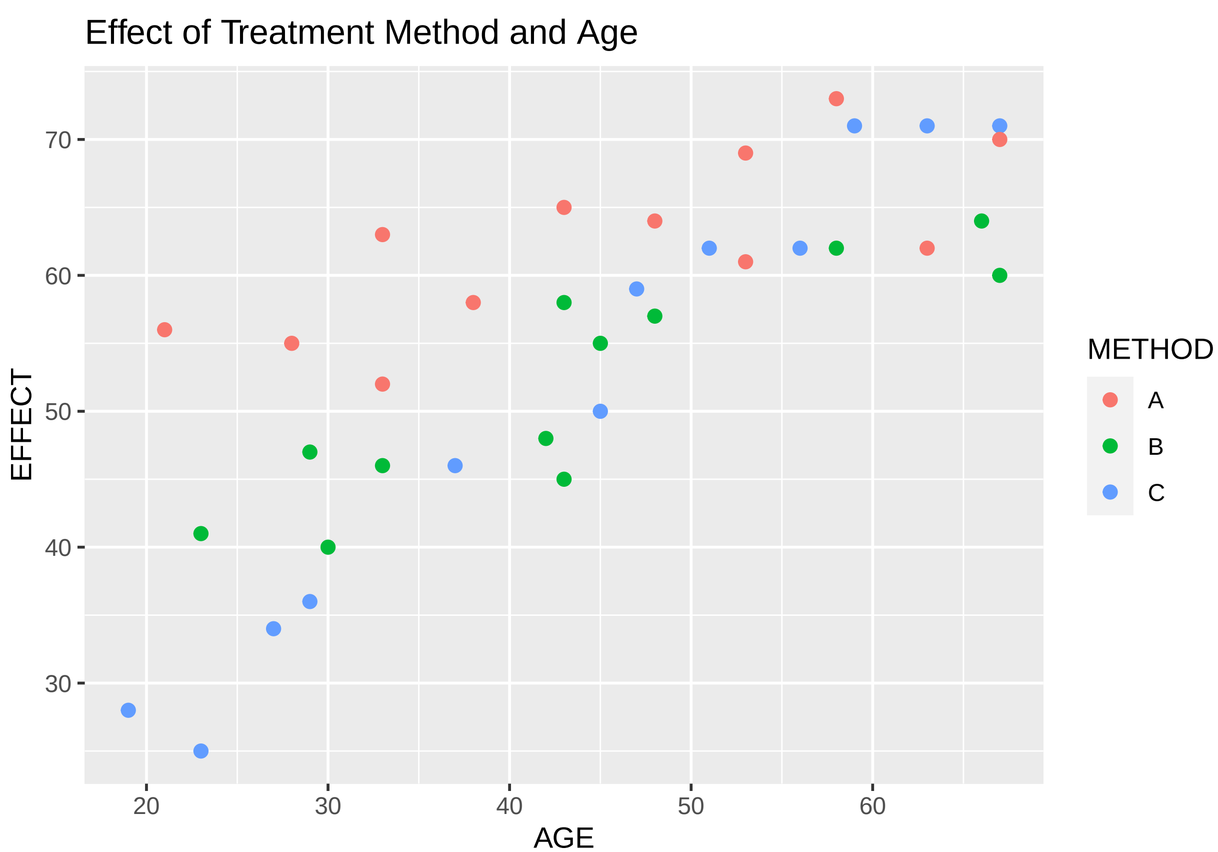
Our response/dependent variable is EFFECT, the effect of the treatment for severe depression.

#### b)

Our predictor/independent variables are AGE, the age of the patient, and METHOD, the treatment method used to treat severe depression.

#### c)

```r
ggplot(mental_health, aes(y = EFFECT, x = AGE, color = METHOD)) + geom_point(size = 2) +ggtitle("Effect
```




Based on the scatter plot, it seems like method A had the largest treatment effect on average. There also seems to be a positive relationship between the age of a patient and the treatment effect, meaning that the older a patient is, the bigger the treatment effect on average.

### d)

To check for interaction between age and treatment method, we will create an interaction model and evaluate the interaction term using the Individual t-test and partial F-test.

```
mental_health_interaction <- lm(EFFECT~(AGE+factor(METHOD))^2, data=mental_health)
summary(mental_health_interaction)
```

```
##
## Call:
## lm(formula = EFFECT ~ (AGE + factor(METHOD))^2, data = mental_health)
##
## 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 the summary output of our full interaction model, all terms are statistically significant except for (AGExMETHOD_B). However, we will include this term in the model since the interaction between (AGExMETHOD_C).

```
mental_health_model <- lm(EFFECT~AGE+factor(METHOD), data = mental_health)
anova(mental_health_model, mental_health_interaction)
```

```
## Analysis of Variance Table
##
## Model 1: EFFECT ~ AGE + factor(METHOD)
## Model 2: EFFECT ~ (AGE + factor(METHOD))^2
##   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
```

Doing a partial F-test, we get a p-value of 9.41e-07. Therefore we can see that adding the interaction terms to our model was significant.

### e)

$\widehat{EFFECT} = 47.51559 + 0.33051 AGE - 18.59739(METHOD_B) - 41.30421(METHOD_C) + 0.19318(AGE * METHOD_B) + 0.70288(AGE * METHOD_C)$

Sub-models:

$\widehat{EFFECT} = 47.51559 + 0.33051 AGE$, When METHOD_A is used.

$\widehat{EFFECT} = 28.9182 + 0.52369 AGE$, When METHOD_B is used.

$\widehat{EFFECT} = 6.21138 + 1.03339 AGE$, When METHOD_C is used.

### f)

We can see from part (e) that the treatment effects how effective the treatment is for different ages. Specifically:

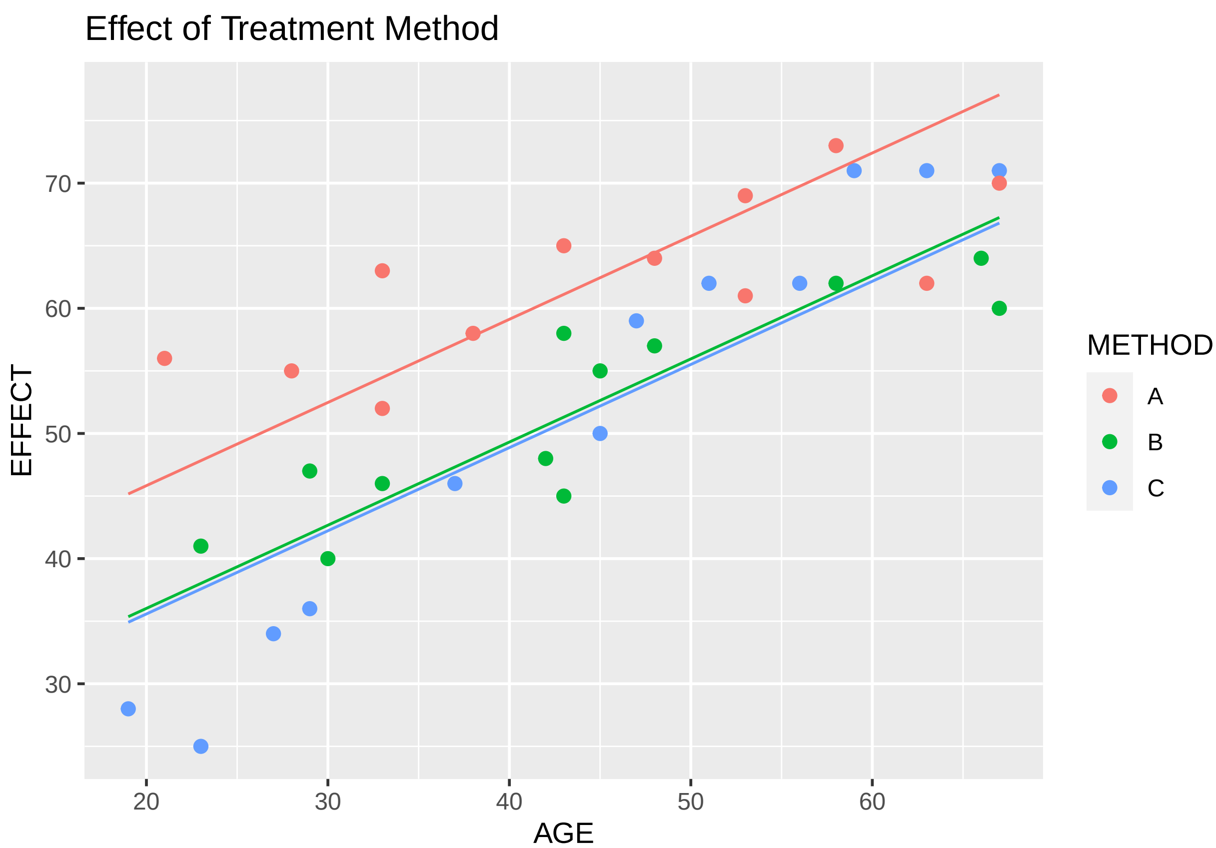
Method A: The treatment effect of METHOD_A when AGE is 0 is 47.51559. For ever 1 year increase in AGE, the treatment effect increases by 0.52369.

Method B: The treatment effect of METHOD_B when AGE is 0 is 28.9182. For ever 1 year increase in AGE, the treatment effect increases by 0.33051.

Method C: The treatment effect of METHOD_C when AGE is 0 is 6.21138 . For ever 1 year increase in AGE, the treatment effect increases by 1.03339.

### g)

```
method_A =function(x){coef(mental_health_model)[2]*x+coef(mental_health_model)[1]}
method_B=function(x){coef(mental_health_model)[2]*x+coef(mental_health_model)[1]+coef(mental_health_mode
method_C=function(x){coef(mental_health_model)[2]*x+coef(mental_health_model)[1]+coef(mental_health_mode
ggplot(mental_health, aes(y = EFFECT, x = AGE, color = METHOD)) + geom_point(size = 2) + stat_function(
```




```
coef(mental_health_model)
```

```
##      (Intercept)              AGE factor(METHOD)B factor(METHOD)C
##       32.5433481        0.6644606       -9.8075777      -10.2527575
```

```
method_A =function(x){47.51559 + x*0.33051}
method_B=function(x){28.9182 + x*0.52369}
method_C=function(x){6.21138 + x*1.03339}
ggplot(mental_health, aes(y = EFFECT, x = AGE, color = METHOD)) + geom_point(size = 2) + stat_function(
```


From the plot with no interaction effects, we would think that METHOD_A is always the best treatment method, however when you look the plot with interactions, METHOD_B seems to be superior for patients with AGE over ~60 years. METHOD_A always seems to be superior to METHOD_C for our data.

### Problem 3

```
# Read in txt file
flag <- read.table("FLAG2.txt", header = TRUE)
flag_subset <- flag[c("LOWBID", "DOTEST", "STATUS", "DISTRICT", "NUMIDS", "DAYSEST", "RDLNGTH", "PCTASPI
```

#### a)

```
# Create full model
flag_full <- lm(LOWBID~., data=flag_subset)
# Apply step-wise regression on full model
stepmod_flag = ols_step_both_p(flag_full,pent = 0.05, prem = 0.1, details=FALSE)
summary(stepmod_flag$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 ***
## STATUS       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
```

Using step-wise regression, the final model contains the variables DOTEST, STATUS and NUMIDS.

Therefore, the final additive model using step-wise regression is: $\widehat{LOWBID} = 57110 + 0.9374(DOTEST) + 95250(STATUS_1) - 15350(NUMIDS)$

We get an adjusted R-squared of 0.9761 and and a RMSE of $281686.7.

**b)**

```r
stepmod_flag2 = ols_step_forward_p(flag_full,pent = 0.05, details=FALSE)
summary(stepmod_flag2$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 ***
## STATUS       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
```

```r
sigma(stepmod_flag2$model)
```

```
## [1] 281686.7
```

Using forward regression procedure with pent=0.05, we get the same suitable independent variables as stepwise regression procedure of DOTEST, STATUS, and NUMIDS. The final additive model from forward regression procedure is:

$\widehat{LOWBID} = 57110 + 0.9374(DOTEST) + 95250(STATUS_1) - 15350(NUMIDS)$

We get an adjusted R-squared of 0.9761 and and a RMSE of \$281686.7. This is the same model as in a).

### c)

```r
stepmod_flag3 = ols_step_backward_p(flag_full,pent = 0.05, details=FALSE)
summary(stepmod_flag3$model)
```

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2051025   -71923     4060    70625  1635136
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.102e+05  6.693e+04   1.647   0.1007
## DOTEST       9.187e-01  1.493e-02  61.523   <2e-16 ***
## STATUS       1.014e+05  4.174e+04   2.430   0.0157 *
## DISTRICT    -1.027e+04  9.089e+03  -1.130   0.2595
## NUMIDS      -1.775e+04  7.906e+03  -2.245   0.0255 *
## DAYSEST      2.493e+02  1.612e+02   1.547   0.1230
## PCTASPH     -9.097e+04  6.581e+04  -1.382   0.1680
## PCTBASE      2.277e+05  1.776e+05   1.282   0.2009
## PCTEXCAV    -3.192e+05  1.528e+05  -2.089   0.0376 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 278700 on 270 degrees of freedom
## Multiple R-squared:  0.9773, Adjusted R-squared:  0.9766
## F-statistic:  1454 on 8 and 270 DF,  p-value: < 2.2e-16
```

```r
sigma(stepmod_flag3$model)
```

```
## [1] 278672.6
```

Using backward regression procedure with pent=0.05, we get the statistically significant independent variables of DOTEST, STATUS, NUMIDS and PCTEXCAV.

Re-running the regression with just these variables

```r
flag_c <- lm(LOWBID~DOTEST+factor(STATUS)+NUMIDS+PCTEXCAV, data=flag_subset)
summary(flag_c)
```

```
##
## Call:
## lm(formula = LOWBID ~ DOTEST + factor(STATUS) + NUMIDS + PCTEXCAV,
##     data = flag_subset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2113731   -70600    -7934    64085  1639128
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.868e+04  4.683e+04   1.467   0.1436
## DOTEST           9.401e-01  9.554e-03  98.404   <2e-16 ***
## factor(STATUS)1  9.624e+04  4.194e+04   2.295   0.0225 *
## NUMIDS          -1.380e+04  7.639e+03  -1.807   0.0719 .
## PCTEXCAV        -1.717e+05  1.457e+05  -1.178   0.2397
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 281500 on 274 degrees of freedom
## Multiple R-squared:  0.9765, Adjusted R-squared:  0.9762
## F-statistic:  2848 on 4 and 274 DF,  p-value: < 2.2e-16
```

Now PCTEXCAV is insignficant and NUMIDS is in the grey zone. We will remove PCTEXCAV and re-run the model one more time:

```r
flag_c <- lm(LOWBID~DOTEST+factor(STATUS)+NUMIDS, data=flag_subset)
summary(flag_c)
```

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

Everything is now significant, so we will stop here for creating an additive model. The significant predictors are DOTEST, STATUS, and NUMIDS.

The final model to be used for prediction would be the following:

$\widehat{LOWBID} = 57110 + 0.9374(DOTEST) + 95250(STATUS_1) - 15350(NUMIDS)$

We get an adjusted R-squared of 0.9761 and and a RMSE of 281686.7. This is the same model as in a) and b).

### d)

```r
summary(lm(LOWBID~., data=flag_subset))
```

```
##
## Call:
## lm(formula = LOWBID ~ ., data = flag_subset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2039770   -74426     7712    75746  1632765
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.064e+05  7.251e+04   1.467   0.1435
## DOTEST       9.207e-01  1.560e-02  59.013   <2e-16 ***
## STATUS       1.037e+05  4.269e+04   2.430   0.0158 *
## DISTRICT    -1.143e+04  9.252e+03  -1.235   0.2178
## NUMIDS      -1.968e+04  8.121e+03  -2.423   0.0160 *
## DAYSEST      1.866e+02  1.791e+02   1.042   0.2983
## RDLNGTH      5.593e+03  4.942e+03   1.132   0.2588
## PCTASPH     -1.162e+05  7.968e+04  -1.458   0.1460
## PCTBASE      2.351e+05  1.842e+05   1.276   0.2031
## PCTEXCAV    -3.103e+05  1.593e+05  -1.948   0.0524 .
## PCTMOBIL     2.601e+05  2.761e+05   0.942   0.3470
## PCTSTRUC     1.039e+05  1.615e+05   0.643   0.5209
## PCTTRAF     -1.067e+05  1.423e+05  -0.750   0.4540
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 279300 on 266 degrees of freedom
## Multiple R-squared:  0.9776, Adjusted R-squared:  0.9765
## F-statistic: 965.6 on 12 and 266 DF,  p-value: < 2.2e-16
```

Conducting individual t-tests with an alpha of 0.05:

$H_0 : \beta_i = 0$

$H_a : \beta_i \neq 0$

$(i = DOTEST, STATUS, DISTRICT, NUMIDS, DAYSEST, RDLNGTH, PCTASPH, PCTBASE, PCTEXCAV, PC$

We get the following statistically significant variables: DOTEST, STATUS,and NUMIDS. Therefore the model that we would propose for predictive purposes should contain these variables.

Re-running the regression with just these variables:

```r
flag_d <- lm(LOWBID~DOTEST+factor(STATUS)+NUMIDS, data=flag_subset)
summary(flag_d)
```

```
##
## Call:
## lm(formula = LOWBID ~ DOTEST + factor(STATUS) + NUMIDS, data = flag_subset)
##
## 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 final model I would recommend for predicting the lowest bid would be the following:

$\widehat{LOWBID} = 57110 + 0.9374(DOTEST) + 95250(STATUS_1) - 15350(NUMBIDS)$

**e)**

The independent variables that consistently are selected throughout the procedures in (a)-(d) are DOTEST, STATUS and NUMIDS. In fact, all additive model from (a)-(d) are exactly the same. Therefore the only possible additive model from (a)-(d) is:

$\widehat{LOWBID} = 57110 + 0.9374(DOTEST) + 95250(STATUS_1) - 15350(NUMBIDS)$

**f)**

```r
flag_f <- lm(LOWBID~DOTEST+factor(STATUS)+NUMIDS+factor(DISTRICT), data = flag_subset)
summary(flag_f)
```

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

The absolute difference between district 1 and district 4 is the regression coefficient for district 4. This is because district one is the default in our model. From the output summary, we get a difference in the average contract bid price between district 1 and 4 of -316505.6188.

### g)

```r
flag_g <- lm(LOWBID~DOTEST+factor(STATUS)+NUMIDS+factor(DISTRICT), data = flag_subset)
summary(flag_g)
```

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

The difference in average bid price from the lowest bidder from district 2 and district 5 is the coefficient of district 2 minus the coefficient of district 5. Therefore we get 71000 - (-14150) = 85150. The difference in the average contract bid of district 2 and district 5 is 85150.

**h)**

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

```
##
## Call:
## lm(formula = LOWBID ~ (DOTEST + factor(STATUS) + NUMIDS + factor(DISTRICT))^2,
##     data = flag_subset)
##
## 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
## NUMIDS                             -4.697e+03  1.273e+04  -0.369  0.71248
## 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
## DOTEST:factor(STATUS)1              9.451e-02  3.673e-02   2.573  0.01063 *
## DOTEST:NUMIDS                      -1.934e-02  3.603e-03  -5.367 1.77e-07 ***
## 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 ***
## factor(STATUS)1:NUMIDS              1.043e+04  3.188e+04   0.327  0.74370
## 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
## NUMIDS:factor(DISTRICT)2            6.114e+03  2.166e+04   0.282  0.77793
## NUMIDS:factor(DISTRICT)3                   NA         NA      NA       NA
## NUMIDS:factor(DISTRICT)4            1.519e+05  4.661e+04   3.260  0.00126 **
## NUMIDS:factor(DISTRICT)5            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
```

Based on the above output, the following terms are significant: DOTEST, DISTRICT, (DOTEST x STATUS), (DOTEST x NUMIDS), (DOTEST x DISTRICT) and (NUMIDS x DISTRICT).

Because there are significant interaction with NUMIDS and STATUS, we should keep these variables in our model. We will remove the interaction between STATUS and DISTRICT as these terms are not significant, and then rerun our model.

I will run a partial F-test to confirm we should remove this interaction. Our null and alternative hypothesis are:

Null hypothesis: $H_0 : \beta_{STATUS*DISTRICT} = 0$

Alternative hypothesis: At least one $H_A : \beta_{STATUS*DISTRICT} \neq 0$

We will set the alpha value to 0.05.

```
anova(lm(LOWBID~DOTEST+factor(STATUS)+NUMIDS+factor(DISTRICT)+DOTEST:factor(STATUS)+DOTEST:NUMIDS+ DOTES
```

```
## Analysis of Variance Table
##
## Model 1: LOWBID ~ DOTEST + factor(STATUS) + NUMIDS + factor(DISTRICT) +
##     DOTEST:factor(STATUS) + DOTEST:NUMIDS + DOTEST:factor(DISTRICT) +
##     NUMIDS:factor(DISTRICT)
## Model 2: LOWBID ~ (DOTEST + factor(STATUS) + NUMIDS + factor(DISTRICT))^2
##   Res.Df        RSS Df  Sum of Sq      F Pr(>F)
## 1    262 1.6556e+13
## 2    260 1.6486e+13  2 6.9441e+10 0.5476  0.579
```

Since out p-value is 0.579 which is larger than the alpha value of 0.05, we fail to reject the null hypothesis that the coefficient of (STATUS x DISTRICT) is different from 0. Therefore we should remove it from our model.

```
flag_h_reduced <- lm(LOWBID~DOTEST+factor(STATUS)+NUMIDS+factor(DISTRICT)+DOTEST:factor(STATUS)+DOTEST:I
summary(flag_h_reduced)
```

```
##
## Call:
## lm(formula = LOWBID ~ DOTEST + factor(STATUS) + NUMIDS + factor(DISTRICT) +
##     DOTEST:factor(STATUS) + DOTEST:NUMIDS + DOTEST:factor(DISTRICT) +
##     NUMIDS:factor(DISTRICT), data = flag_subset)
##
## 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
## NUMIDS                    1.974e+02  1.181e+04   0.017   0.9867
## 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
## DOTEST:factor(STATUS)1    9.218e-02  3.580e-02   2.575   0.0106 *
## DOTEST:NUMIDS            -1.995e-02  3.549e-03  -5.622 4.82e-08 ***
## 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 ***
## NUMIDS:factor(DISTRICT)2  1.648e+03  2.119e+04   0.078   0.9381
## NUMIDS:factor(DISTRICT)3         NA         NA      NA       NA
## NUMIDS:factor(DISTRICT)4  1.513e+05  4.653e+04   3.252   0.0013 **
## NUMIDS:factor(DISTRICT)5  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
```

When we re-run our model, now the interaction between NUMIDS and DISTRICT is insignificant the rest of the interactions with dummy variables have at least one significant term, so we will keep all of them.

Now all interaction terms have at least one significant interaction with a dummy variable, so we can stop here. Our final model is the following:

$\widehat{LOWBID} = -73430 + 1.102(DOTEST) + 61560(STATUS_1) + 197.4(NUMIDS) + 24580(DISTRICT_2) + 63260(DISTRICT_3) - 1531000(DISTRICT_4) + 15720(DISTRICT_5) + 0.09218(DOTEST * STATUS_1) - 0.01995(DOTEST * NUMIDS) + 0.03939(DOTEST * DISTRICT_2) - 0.1326(DOTEST * DISTRICT_3 - 0.02532(DOTEST * DISTRICT_4) - 0.1335(DOTEST * DISTRICT_5) + 1648(NUMIDS * DISTRICT_2) + NUMIDS * DISTRICT_3 + 151300(NUMIDS * DISTRICT_4) + 18030((NUMIDS * DISTRICT_5)$

### i)

```
sigma(flag_d)
```

```
## [1] 281686.7
```

```
sigma(flag_h_reduced)
```

```
## [1] 251376.4
```

RMSE in part (d): 281686.7

RMSE in part (h): 251376.4

The RMSE from the model in part (h) is much lower than the RMSE from the model in part (d). This means that the standard deviation of the unexplained variance from model in (h) is lower than from the model in part (d). This means that the model in part (h) is superior in terms of RMSE when compared to the model in part (d).

### j)

```
summary(flag_h_reduced)$adj.r.square
```

```
## [1] 0.9809988
```

We get an adjusted R-squared value from our model in part (h) of 0.9809988 This means that 98.10% of the variance in price of the contract bid by the lowest bidder can be explained by out model. Considering the maximum value that adjusted R-squared can be is 1.00 (or 100%), this is quite a high value.

## Problem 4

```r
# Read in CSV file
kbi <- read.csv("KBI.csv")
```

### a)

```r
# Build the full model
kbi_full <- lm(BURDEN~., data=kbi)
# Do step-wise regression on full model
stepmod_kbi = ols_step_both_p(kbi_full,pent = 0.1, prem = 0.3, details=FALSE)
summary(stepmod_kbi$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
```

CGDUR is not statistically significant when using an alpha value of 0.05, so we will re-run the regression without this variable.

```r
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
```

All variables are now statistically significant. So using step-wise regression, we get the following significant predictor variables: MEM and SOCIALSU . Therefore, the model used for predicting caregiver burden would be:

$\widehat{BURDEN} = 116.07291 + 0.59941(MEM) - 0.47552(SOCIALSU)$

### b)

To do all-possible-regressions-selection, I will use the method from the "leaps" library.

```
best.subset <- regsubsets(BURDEN~., data = kbi, nv=10)
reg.summary<-summary(best.subset)
cp<-c(reg.summary$cp)
AdjustedR<-c(reg.summary$adjr2)
RMSE<-c(reg.summary$rss)
BIC<-c(reg.summary$bic)
cbind(cp,BIC,RMSE,AdjustedR)
```

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

Based on the output of our all-possible-regressions selection procedure, we should choose the model with 2 variables since it has the lowest BIC, cp is only 0.610120 away from p+1, and adjusted R-squared is only 0.0168541 lower than the highest RMSE. We can look at the summary output to see which 2 variables should be included in the model.

```
reg.summary
```

```
## Subset selection object
## Call: regsubsets.formula(BURDEN ~ ., 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 ) "*"   "*"      "*"   "*" "*" "*" "*"
```

The 2 variables that should be included in the model are: MEM and SOCIALSU.

```r
kbi_fit <- lm(BURDEN~MEM+SOCIALSU, data = kbi)
summary(kbi_fit)
```

```
##
## 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
```

Therefore our final addictive model using all-possible-regressions-selection is: $\widehat{BURDEN} = 116.07291 + 0.59941(MEM) - 0.59941(SOCIALSU)$

**c)**

Looking at (a) and (b), the variables that consistently are selected for the best model is MEM and SOCIALSU. I will use these three variables to create an interaction model.

```r
summary(lm(BURDEN~(MEM+SOCIALSU)^2, data = kbi))
```

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

```
## lm(formula = BURDEN ~ (MEM + SOCIALSU)^2, data = kbi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.911 -11.169  -0.326   8.725  35.121
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.163e+02  2.423e+01   4.801  5.8e-06 ***
## MEM           5.905e-01  7.480e-01   0.790   0.4318
## SOCIALSU     -4.773e-01  1.768e-01  -2.700   0.0082 **
## MEM:SOCIALSU  6.559e-05  5.471e-03   0.012   0.9905
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.52 on 96 degrees of freedom
## Multiple R-squared:  0.4192, Adjusted R-squared:  0.401
## F-statistic:  23.1 on 3 and 96 DF,  p-value: 2.44e-11
```

Based on our output, there are no statistically significant interactions between the variables, so we should not include any interaction terms in our first order model.

Therefore, the final model I would suggest for prediction of caregiver burden would be:

```
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
```

```
sigma(lm(BURDEN~MEM+SOCIALSU, data = kbi))
```

```
## [1] 15.44289
```

$\widehat{BURDEN} = 116.07291 + 0.59941(MEM) - 0.47552(SOCIALSU)$. This model gives an adjusted R-squared of 0.4072 and an RMSE of 15.44289.

**Session Info:**

```r
sessionInfo()
```

```
## R version 4.1.3 (2022-03-10)
## Platform: x86_64-conda-linux-gnu (64-bit)
## Running under: Ubuntu 22.04.1 LTS
##
## Matrix products: default
## BLAS/LAPACK: /opt/conda/lib/libopenblasp-r0.3.21.so
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] leaps_3.1         olsrr_0.5.3       mosaic_1.8.3      ggridges_0.5.3
##  [5] mosaicData_0.20.2 ggformula_0.10.1  ggstance_0.3.5    Matrix_1.4-1
##  [9] lattice_0.20-45   dplyr_1.0.9       ggplot2_3.3.6
##
## loaded via a namespace (and not attached):
##  [1] ggrepel_0.9.1     Rcpp_1.0.9        tidyr_1.2.0       assertthat_0.2.1
##  [5] digest_0.6.29     utf8_1.2.2        ggforce_0.3.4     R6_2.5.1
##  [9] plyr_1.8.7        backports_1.4.1   labelled_2.9.1    evaluate_0.16
## [13] highr_0.9         pillar_1.8.1      rlang_1.0.4       data.table_1.14.2
## [17] rstudioapi_0.14   car_3.1-0         goftest_1.2-3     rmarkdown_2.15
## [21] labeling_0.4.2    splines_4.1.3     readr_2.1.2       stringr_1.4.1
## [25] htmlwidgets_1.5.4 polyclip_1.10-0   munsell_0.5.0     broom_1.0.0
## [29] compiler_4.1.3    xfun_0.32         pkgconfig_2.0.3   htmltools_0.5.3
## [33] tidyselect_1.1.2  tibble_3.1.8      gridExtra_2.3     mosaicCore_0.9.0
## [37] fansi_1.0.3       tzdb_0.3.0        withr_2.5.0       MASS_7.3-58.1
## [41] grid_4.1.3        gtable_0.3.0      lifecycle_1.0.1   DBI_1.1.3
## [45] magrittr_2.0.3    scales_1.2.1      carData_3.0-5     cli_3.3.0
## [49] stringi_1.7.8     farver_2.1.1      leaflet_2.1.1     ellipsis_0.3.2
## [53] ggdendro_0.1.23   generics_0.1.3    vctrs_0.4.1       nortest_1.0-4
## [57] tools_4.1.3       forcats_0.5.2     glue_1.6.2        tweenr_2.0.1
## [61] purrr_0.3.4       hms_1.1.2         crosstalk_1.2.0   abind_1.4-5
## [65] fastmap_1.1.0     yaml_2.3.5        colorspace_2.0-3  knitr_1.39
## [69] haven_2.5.0
```