DATA 603 Assignment 2

Michael Ellsworth

November 22nd, 2019

Problem 1

The amount of water used by the production facilities of a plant varies. Observations on water usage and other possibility related variables, were collected for 249 months. The data are given in water.csv file. The explanatory variables are:

- *TEMP = average monthly temperature(degree celsius)*
- *PROD = amount of production(10cubic)*
- *DAYS = number of operationing day in the month*
- *HOUR = number of hours shut down for maintenance*

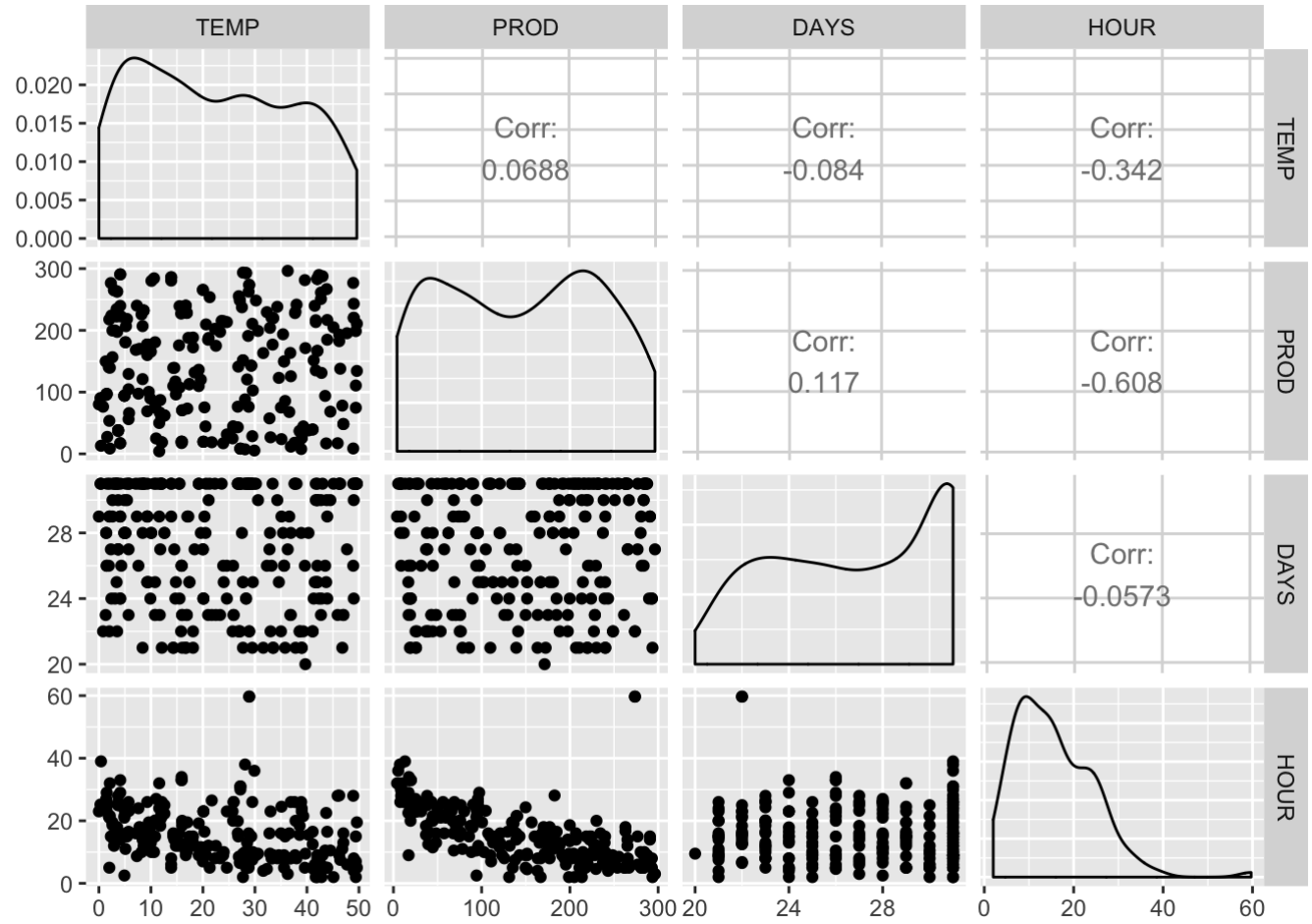
The response variable is USAGE = monthly water usage (gallons/minute). From Exercise 1 and 2, assume that the best fitted model is:

$$\widehat{USAGE} = \hat{\beta}_0 + \hat{\beta}_1 TEMP + \hat{\beta}_2 HOUR + \hat{\beta}_3 PROD * TEMP + \hat{\beta}_4 PROD * HOUR$$

a

Many researchers avoid the problems of multicollinearity by always omitting all but one of the "redundant" variables from the model. By checking all pairwise combinations of predictors in scatterplots and using the VIF function, do you detect any high correlation (r > 0.8) between predictors? Does there appear to be any problem with multicollinearity assumption?

```
full_additive_water <- lm(data = water, USAGE ~ TEMP + PROD + DAYS + HOUR)
best_model_water <- lm(data = water, USAGE ~ TEMP + HOUR + PROD * TEMP + PROD * HOUR)
ggpairs(data = water, columns = c("TEMP", "PROD", "DAYS", "HOUR"))
```



Based on all of the predictor scatterplots, there does not appear to be any high correlation between the predictors. All the predictors have $r < 0.8$. Additionally, the data is quite scattered in each of the plots without any reasonable observable pattern.

```
vif(full_additive_water)
```

```
##     TEMP     PROD     DAYS     HOUR
## 1.184136 1.657098 1.023041 1.855499
```

This conclusion is supported by calculating the VIF of each variable. Each of the predictor variables have a VIF between 1 and 5 which suggests moderate collinearity but it is not severe enough to correct the model. There does not appear to be any significant multicollinearity present in the model.

b

Conduct a test for heteroscedasticity (non constant variance) and plot a residual plot. Does there appear to be any problem with homoscedasticity assumption?

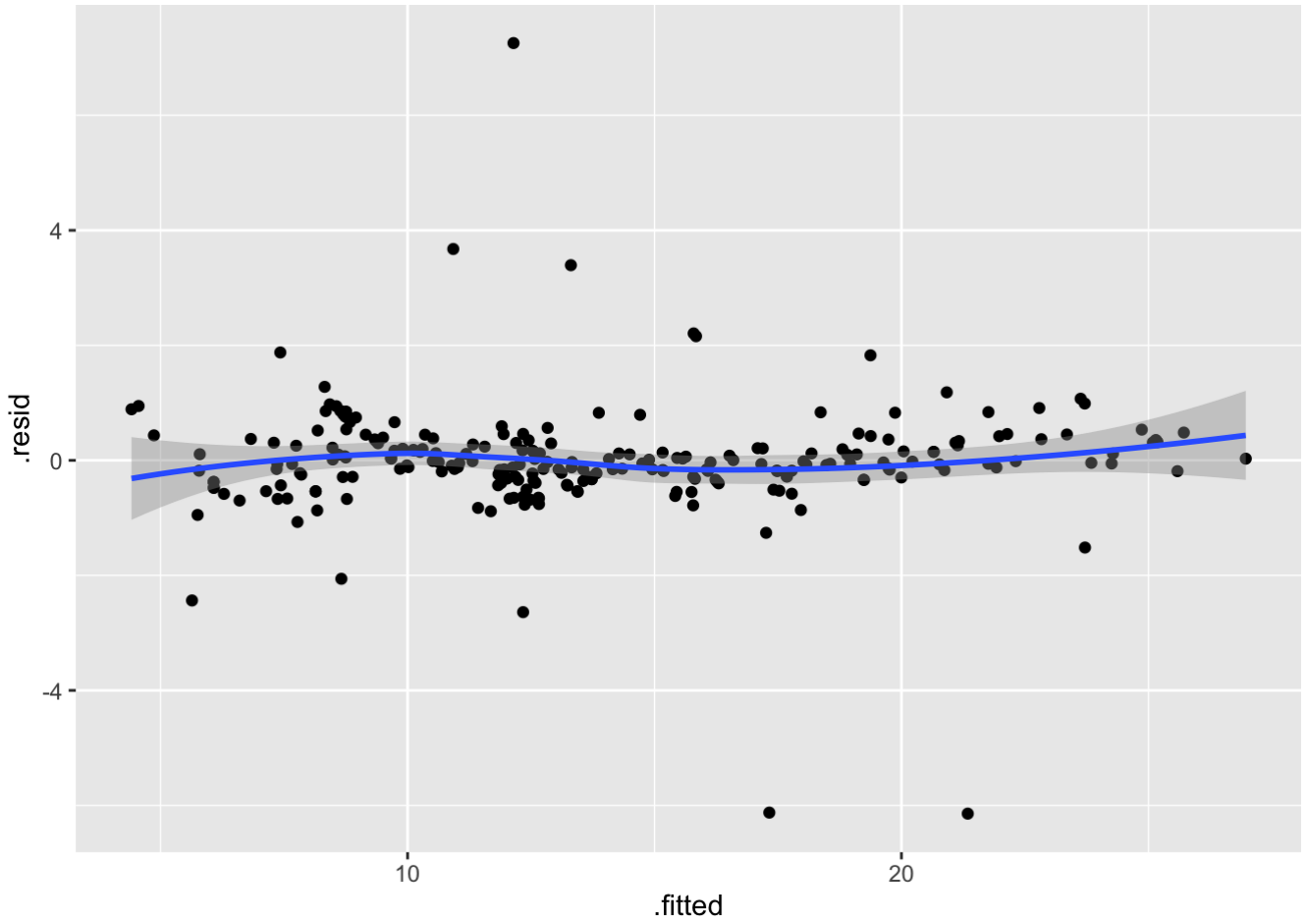
```
bptest(best_model_water)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  best_model_water
## BP = 2.0057, df = 5, p-value = 0.8484
```

Based on the Breusch-Pagan test for Heteroscedasticity, we cannot reject the null hypothesis and we can say that heteroscedasticity does not exist.

```
best_model_water %>%
  ggplot(aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_smooth()
```

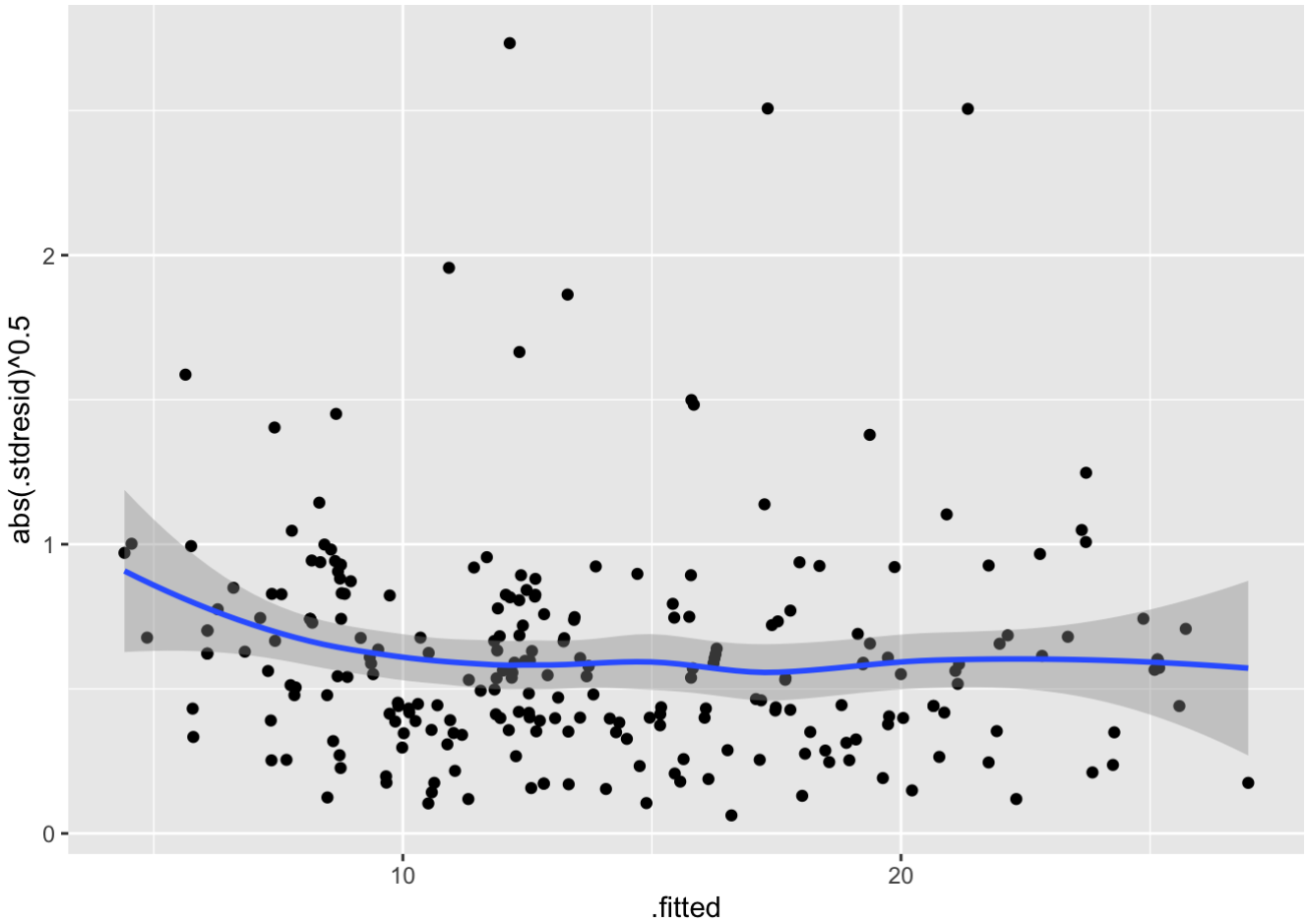
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Checking the residual versus fitted values plot, we can confirm that heteroscedasticity does not exist as the residuals appear to fall along a straight line on 0.

```
best_model_water %>%
  ggplot(aes(x = .fitted, y = abs(.stdresid)**0.5)) +
  geom_point() +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

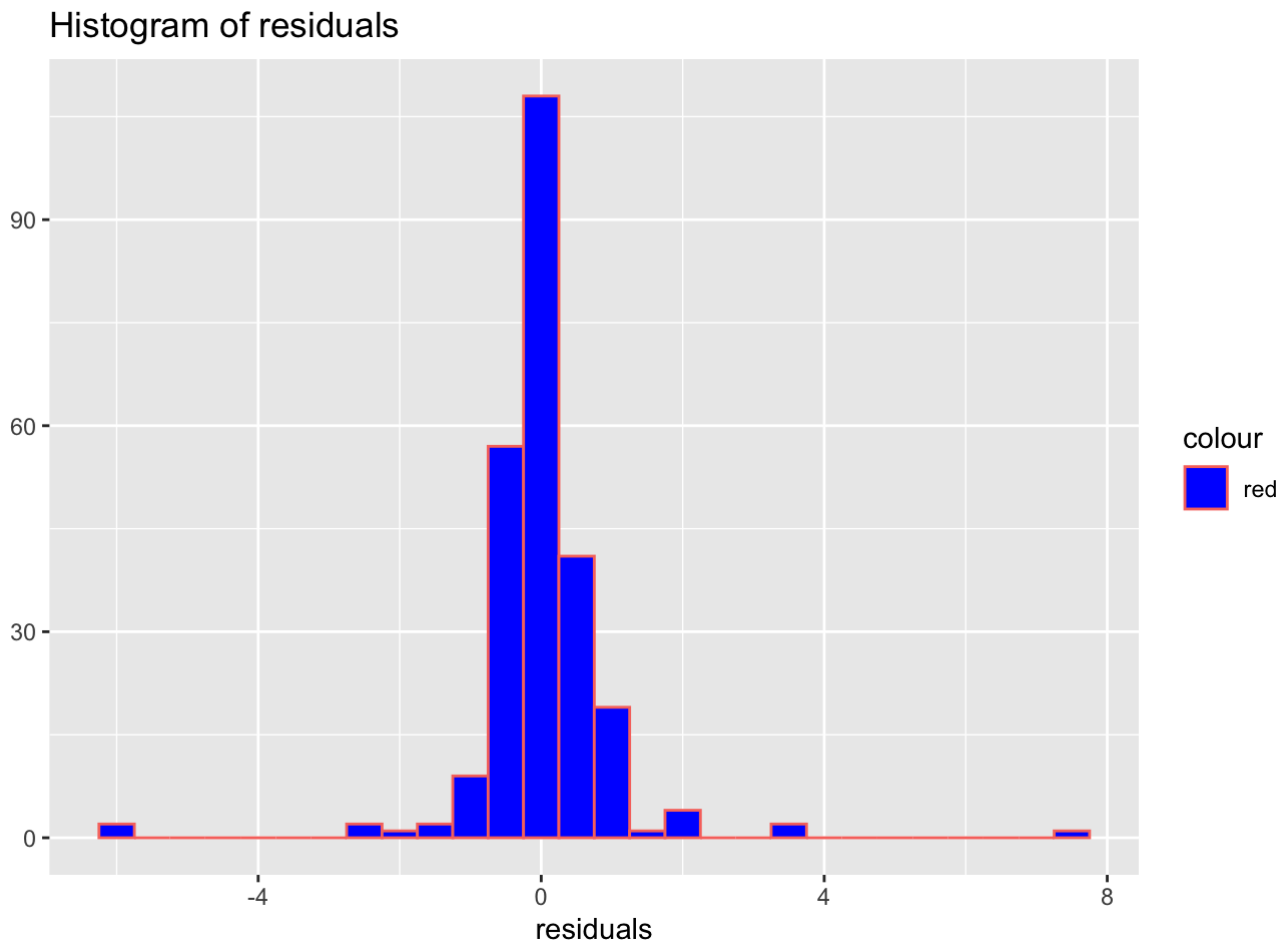


Again, from the scale-location plot, we can confirm heteroscedasticity.

C

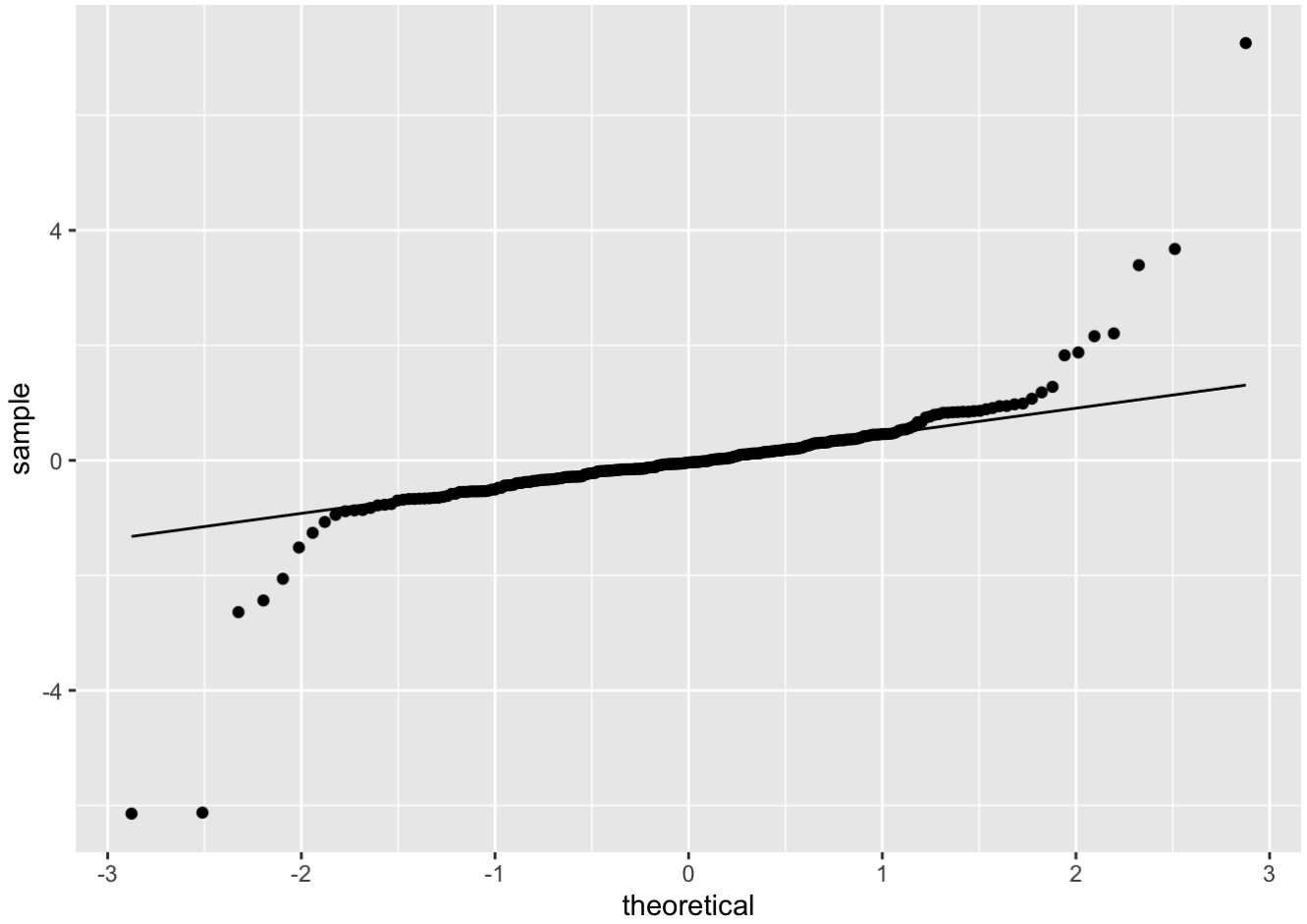
Provide a histogram for residuals, a normal Q-Q plot, and the Shapiro-Wilk test. Does there appear to be any problem with normality assumption?

```
qplot(residuals(best_model_water),
      geom = "histogram",
      binwidth = 0.5,
      main = "Histogram of residuals",
      xlab = "residuals",
      color="red",
      fill=I("blue"))
```



Based on the results from the histogram, there does appear to be some values that do no fall a normal distribution. Specifically, the residuals that appear between 3 and 8.

```
water %>%
  ggplot(aes(sample = best_model_water$residuals)) +
  stat_qq() +
  stat_qq_line()
```



Additionally, as the theoretical values increase in absolute value, the data does not appear to be normal.

```
shapiro.test(residuals(best_model_water))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(best_model_water)
## W = 0.67655, p-value < 2.2e-16
```

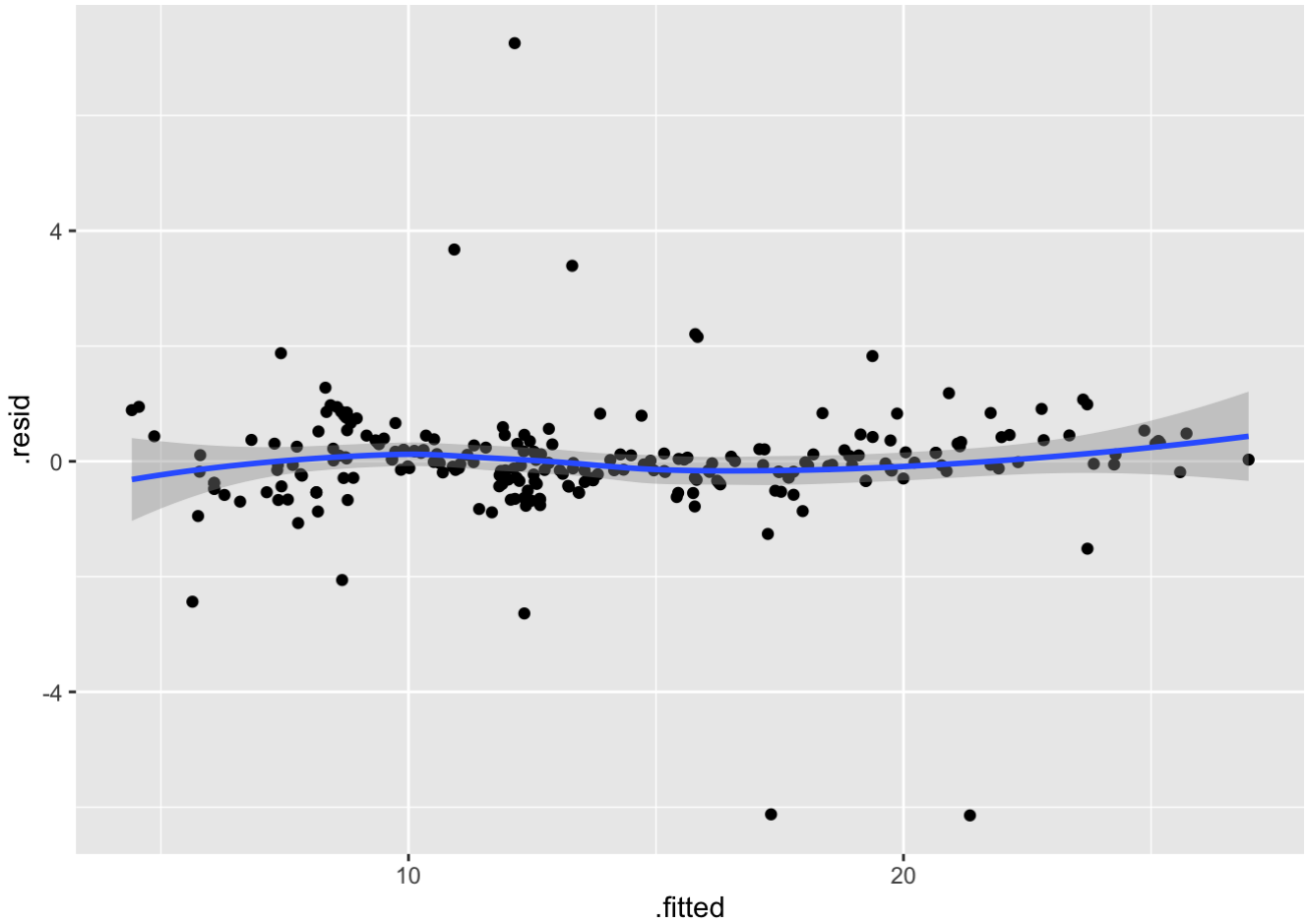
This is confirmed with the Sahpiro-Wilk normality test as the P-value is less than 0.05 meaning we can reject the null hypothesis that the residuals are normally distributed.

d

Plot the residuals vs predicted value $\hat{Y}$ plot, do you detect any patterns? Does there appear to be any problem with linearity assumption?

```
best_model_water %>%
  ggplot(aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



There does not appear to be any pattern in the plot above meaning there is not a problem with the linearity assumption. The model is suitable without any quadratic terms.

e

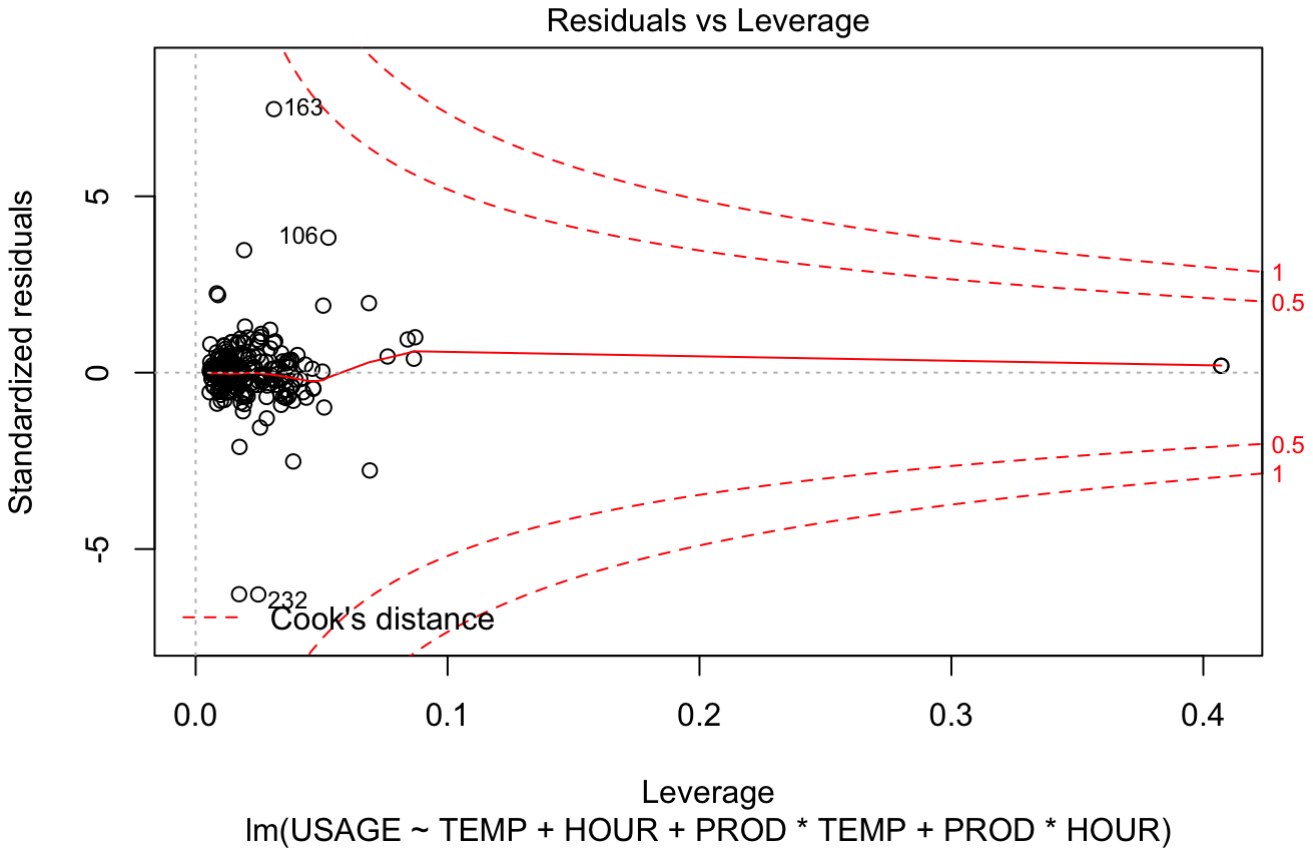
Do you detect any outliers by using Cook's distance measure (using cooks.distance() > 1) and Residual vs Leverage plot?

```
water[cooks.distance(best_model_water) > 1, ]
```

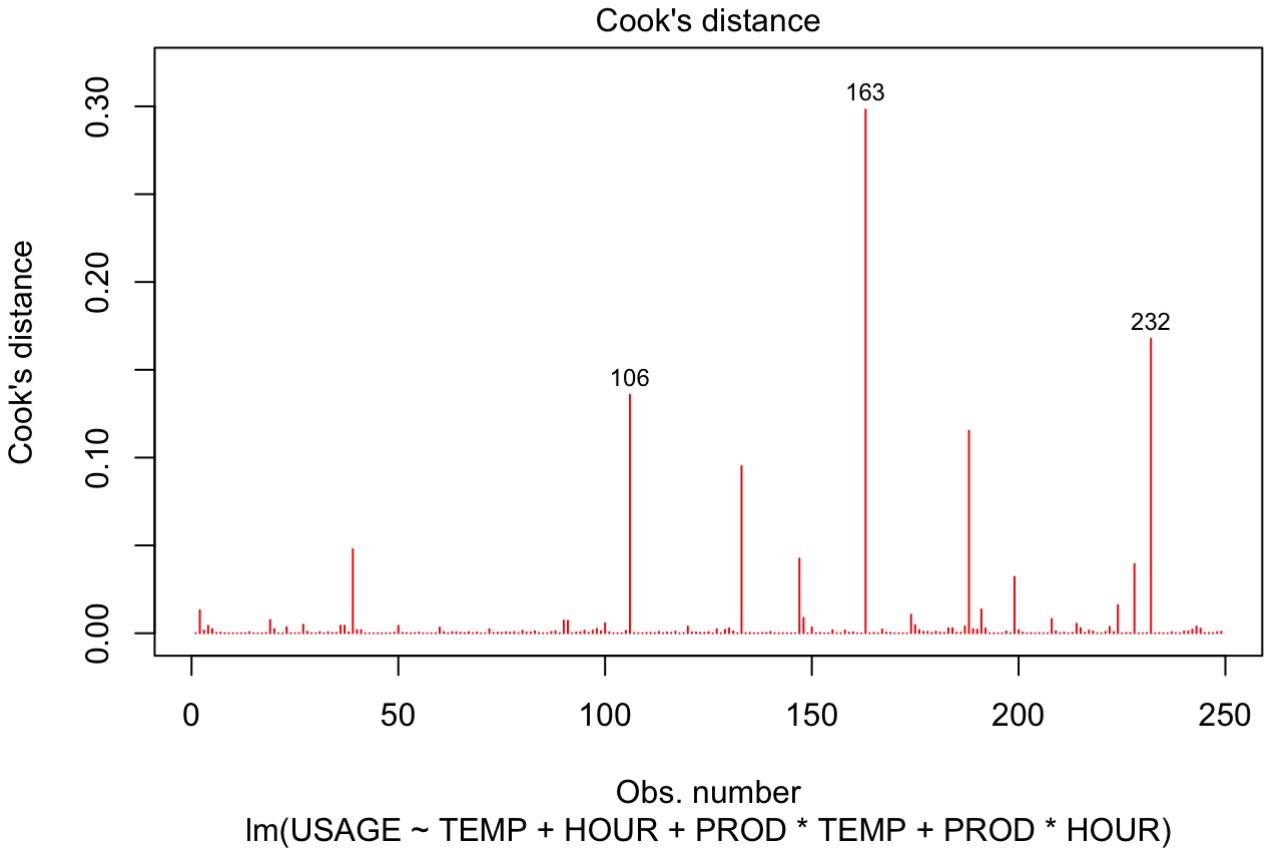
0 rows

By measuring cooks distance with a cutoff of 1, there are no outliers detected.

```
plot(best_model_water, which = 5)
```




```
plot(best_model_water, pch = 18, col = "red", which = 4)
```



This is confirmed by the plots above. The highest cook's distance appears to be observation number 163 with a cook's distance of ~0.3.

f

From part a-e, determine whether your model meets the assumptions of the analysis.

Based on the questions above, the following assumptions are true:

- No significant multicollinearity
- No heteroscedasticity
- Linearity assumption
- No significant outliers

The assumption that the residuals follow a normal distribution is not valid. The residuals follow an s-shaped pattern which indicates that the residuals have excessive skewness or there are either too many or too few large errors in both directions.

Problem 2 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 project margin). By comparing the bid prices (and other important bid variables) of the fixed (or rigged) contracts to the competitively bid contracts, FLAG was able to establish invaluable benchmarks for detecting future bid-rigging. FLAG collected data for 279 road construction contracts. For each contract, the following variables shown below were measured and are only considered for this problem:

1. *Price of contract ($) bid by lowest bidder, LOWBID.*
2. *Department of Transportation (DOT) engineer's estimate of fair contract price ($), DOTEST.*
3. *Status of contract (1 if fixed, 0 if competitive), STATUS*
4. *District (1, 2, 3, 4, or 5) in which construction project is located, DISTRICT.*
5. *Number of bidders on contract, NUMIDS.*
6. *Estimated number of days to complete work, DAYSEST.*
7. *Length of road project (miles), RDLNGTH.*
8. *Percentage of costs allocated to liquid asphalt, PCTASPH.*
9. *Percentage of costs allocated to base material, PCTBASE.*
10. *Percentage of costs allocated to excavation, PCTEXCAV.*
11. *Percentage of costs allocated to mobilization, PCTMOBIL.*
12. *Percentage of costs allocated to structures, PCTSTRUC.*
13. *Percentage of costs allocated to trafic control, PCTTRAF.*

The data are saved in the file named FLAG2.txt

a

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

```
full_model_FLAG <- lm(data = FLAG, LOWBID ~.)
ols_step_both_p(full_model_FLAG, pent = 0.05, prem = 0.1)
```

```
## 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
##
## We are selecting variables based on p value...
##
## Variables Entered/Removed:
##
## - DOTEST added
## - STATUS added
## - NUMIDS added
##
## No more variables to be added/removed.
##
##
## Final Model Output
## ------------------
##
##                             Model Summary
## ----------------------------------------------------------------------
## R                       0.988       RMSE                    281686.705
## R-Squared               0.976       Coef. Var                   24.621
## Adj. R-Squared          0.976       MSE                79347399544.838
## Pred R-Squared          0.974       MAE                     137748.048
## ----------------------------------------------------------------------
##  RMSE: Root Mean Square Error
```

```
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                    ANOVA
## --------------------------------------------------------------------------
##                     Sum of
##                    Squares          DF        Mean Square       F          Sig.
## --------------------------------------------------------------------------
## Regression    9.026917e+14           3       3.008972e+14     3792.15    0.0000
## Residual      2.182053e+13         275    79347399544.838
## Total         9.245122e+14         278
## --------------------------------------------------------------------------
##
##                                      Parameter Estimates
## ------------------------------------------------------------------------------------------------------
##       model           Beta     Std. Error    Std. Beta       t        Sig          lower          upper
## ------------------------------------------------------------------------------------------------------
## (Intercept)     57105.973     45823.264                    1.246    0.214    -33102.981     147314.927
##      DOTEST         0.937         0.009        1.000     101.011    0.000         0.919          0.956
##     STATUS1     95252.389     41959.825        0.024       2.270    0.024     12649.108     177855.670
##      NUMIDS    -15353.820      7530.216       -0.023      -2.039    0.042    -30178.014       -529.626
## ------------------------------------------------------------------------------------------------------
```

```
##
##                                Stepwise Selection Summary
## ---------------------------------------------------------------------------------------
##                      Added/                   Adj.
## Step    Variable     Removed     R-Square    R-Square     C(p)         AIC          RMSE
## ---------------------------------------------------------------------------------------
##    1     DOTEST      addition       0.975       0.975    24.9740    7812.6645    289264.4939
##    2     STATUS      addition       0.976       0.976    13.6790    7802.0163    283293.3317
##    3     NUMIDS      addition       0.976       0.976    11.4100    7799.8301    281686.7046
## ---------------------------------------------------------------------------------------
```

Based on the stepwise regression procedure, the variables that are suitable for modelling 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.

b

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

```
ols_step_forward_p(full_model_FLAG, pent = 0.05)
```

```
## Forward 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
##
## We are selecting variables based on p value...
##
## Variables Entered:
##
## - DOTEST
## - STATUS
## - NUMIDS
##
## No more variables to be added.
##
## Final Model Output
## ------------------
##
##                             Model Summary
## ----------------------------------------------------------------------
## R                       0.988       RMSE                    281686.705
## R-Squared               0.976       Coef. Var                   24.621
## Adj. R-Squared          0.976       MSE                79347399544.838
## Pred R-Squared          0.974       MAE                     137748.048
## ----------------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
```

```
##  MAE: Mean Absolute Error
##
##                                    ANOVA
## --------------------------------------------------------------------------------
##                     Sum of
##                    Squares         DF        Mean Square       F          Sig.
## --------------------------------------------------------------------------------
## Regression    9.026917e+14          3       3.008972e+14     3792.15    0.0000
## Residual      2.182053e+13        275    79347399544.838
## Total         9.245122e+14        278
## --------------------------------------------------------------------------------
##
##                                      Parameter Estimates
## ----------------------------------------------------------------------------------------------------
##       model          Beta    Std. Error    Std. Beta       t        Sig          lower          upper
## ----------------------------------------------------------------------------------------------------
## (Intercept)     57105.973     45823.264                   1.246    0.214    -33102.981    147314.927
##      DOTEST         0.937         0.009        1.000    101.011    0.000         0.919         0.956
##     STATUS1     95252.389     41959.825        0.024      2.270    0.024     12649.108    177855.670
##      NUMIDS    -15353.820      7530.216       -0.023     -2.039    0.042    -30178.014      -529.626
## ----------------------------------------------------------------------------------------------------
```

```
##
##                              Selection Summary
## --------------------------------------------------------------------------------
##         Variable                  Adj.
## Step    Entered     R-Square    R-Square     C(p)         AIC          RMSE
## --------------------------------------------------------------------------------
##    1    DOTEST        0.9749      0.9748    24.9744    7812.6645    289264.4939
##    2    STATUS        0.9760      0.9759    13.6791    7802.0163    283293.3317
##    3    NUMIDS        0.9764      0.9761    11.4097    7799.8301    281686.7046
## --------------------------------------------------------------------------------
```

Based on the forward regression procedure, the variables that are suitable for modelling 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.

C

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

```
ols_step_backward_p(full_model_FLAG, prem = 0.05)
```

```
## 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
##
## We are eliminating variables based on p value...
##
## Variables Removed:
##
## - DAYSEST
## - PCTTRAF
## - PCTSTRUC
## - RDLNGTH
## - PCTMOBIL
## - PCTBASE
## - DISTRICT
## - PCTEXCAV
## - PCTASPH
##
## No more variables satisfy the condition of p value = 0.05
##
##
## Final Model Output
## ------------------
##
##                             Model Summary
## ----------------------------------------------------------------------
```

```
## R                       0.988       RMSE                    281686.705
## R-Squared               0.976       Coef. Var                   24.621
## Adj. R-Squared          0.976       MSE                  79347399544.838
## Pred R-Squared          0.974       MAE                      137748.048
## ----------------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                  ANOVA
## --------------------------------------------------------------------------------
##                     Sum of
##                    Squares         DF        Mean Square       F          Sig.
## --------------------------------------------------------------------------------
## Regression    9.026917e+14          3       3.008972e+14     3792.15    0.0000
## Residual      2.182053e+13        275    79347399544.838
## Total         9.245122e+14        278
## --------------------------------------------------------------------------------
##
##                                     Parameter Estimates
## ----------------------------------------------------------------------------------------------------
##       model          Beta    Std. Error    Std. Beta       t        Sig          lower          upper
## ----------------------------------------------------------------------------------------------------
## (Intercept)     57105.973     45823.264                    1.246    0.214    -33102.981    147314.927
##      DOTEST         0.937         0.009        1.000    101.011    0.000         0.919         0.956
##     STATUS1     95252.389     41959.825        0.024      2.270    0.024     12649.108    177855.670
##      NUMIDS    -15353.820      7530.216       -0.023     -2.039    0.042    -30178.014      -529.626
## ----------------------------------------------------------------------------------------------------
```

```
##
##
##                             Elimination Summary
## ------------------------------------------------------------------------
##         Variable                  Adj.
## Step    Removed     R-Square    R-Square     C(p)         AIC          RMSE
## ------------------------------------------------------------------------
##    1    DAYSEST        0.978      0.9768     8.1887    7802.1641    277539.3588
##    2    PCTTRAF        0.978      0.9769     6.6672    7800.6708    277266.8939
##    3    PCTSTRUC      0.9778      0.9769     6.0352    7800.1146    277462.1943
##    4    RDLNGTH       0.9777      0.9768     5.3566    7799.5022    277631.6720
##    5    PCTMOBIL      0.9776      0.9768     4.6027    7798.8044    277760.6667
##    6    PCTBASE       0.9775      0.9767     4.5648    7798.8426    278258.4309
##    7    DISTRICT      0.9769      0.9764     9.7652    7798.1973    279877.0695
##    8    PCTEXCAV      0.9766      0.9763    10.6882    7799.1284    280837.2457
##    9    PCTASPH       0.9764      0.9761    11.4097    7799.8301    281686.7046
## ------------------------------------------------------------------------
```

Based on the backward regression procedure, the variables that are suitable for modelling 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.

d

Using the full model with all predictors, test the individual t-test at $\alpha = 0.05$. What predictors should be added to the model?

```
summary(full_model_FLAG)
```

```
##
## Call:
## lm(formula = LOWBID ~ ., data = FLAG)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2061552   -76832     3703    68246  1592629
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.623e+04  6.916e+04   1.102   0.2714
## DOTEST       9.362e-01  1.687e-02  55.494   <2e-16 ***
## 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
```

Based on the summary of the full model, the following variables are significant and should be added to the model:

- 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.
- District (1, 2, 3, 4, or 5) in which construction project is located, DISTRICT

Additionally, one might look at the PCTEXCAV variable as a suitable predictor for the model as it is close to the p-value cut-off of 0.05. For now, we will include the additional variable, DISTRICT, only.

```
reduced_model_FLAG_d <- lm(data = FLAG, LOWBID ~ DOTEST + STATUS + NUMIDS + DISTRICT)
summary(reduced_model_FLAG_d)
```

```
##
## Call:
## lm(formula = LOWBID ~ DOTEST + STATUS + NUMIDS + DISTRICT, data = FLAG)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2160166   -66952    -6042    55358  1625579
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.050e+04  5.197e+04   1.164   0.2454
## DOTEST       9.447e-01  1.002e-02  94.258   <2e-16 ***
## STATUS1      9.991e+04  4.189e+04   2.385   0.0178 *
## NUMIDS      -1.736e+04  8.255e+03  -2.103   0.0364 *
## DISTRICT2    7.100e+04  6.316e+04   1.124   0.2619
## DISTRICT3    1.156e+04  2.038e+05   0.057   0.9548
## DISTRICT4   -3.165e+05  1.336e+05  -2.370   0.0185 *
## DISTRICT5   -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
```

e

Compare the results, parts a-d. Which independent variables consistently are selected as the "best" predictors for the first order model? Write the first order model for predicting Y.

The independent variables that are consistently selected as the best predictors 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.

```
reduced_model_FLAG <- lm(data = FLAG, LOWBID ~ DOTEST + STATUS + NUMIDS)
coefficients(reduced_model_FLAG)
```

```
##   (Intercept)        DOTEST       STATUS1        NUMIDS
##  5.710597e+04  9.374269e-01  9.525239e+04 -1.535382e+04
```

$$\widehat{LOWBID} = \hat{\beta}_0 + \hat{\beta}_1 DOTEST + \hat{\beta}_2 STATUS + \hat{\beta}_3 NUMIDS$$

Where,

$$\begin{aligned}\hat{\beta}_0 &= 57105.973 \\ \hat{\beta}_1 &= 0.937 \\ \hat{\beta}_2 &= 95252.389 \\ \hat{\beta}_3 &= -15353.820\end{aligned}$$

f

Interpret the regression coefficients for each β_i

If the Depatrment of Transportation's estimate increases by $1, the price of the contract for the lowest bidder will increase by $0.937. If the status of the contract is fixed, the contract price will be $92,252.40 higher than if the contract was competitive. If the number of bidders increases by one, the constract price of the lowest bidder will decrease by $15,353.82. The lowest bidder's contract price will start at $57,105.97 if all other variables are 0.

g

Apply All Possible Regressions Selection Procedure to confirm that the independent variables in part (d) are suitable for modeling Y. Provide all three criteria value (Cp, AIC, R^2_{adj}) for the model selected.

```
ols_step_best_subset(full_model_FLAG)
```

	mindex <int>	n <int>
1	1	1
13	2	2
79	3	3
299	4	4
799	5	5
1605	6	6
2546	7	7
3367	8	8
3832	9	9
4038	10	10

1-10 of 12 rows | 1-3 of 15 columns

Previous **1** 2 Next

Based on the Subsets Regression Summary, Model 4 suggests that, with a relatively low Cp, AIC and a relatively high R^2_{adj}, the model variables from part d are suitable. Additionally, since model 7 is not the most suitable model, we can confirm that even though the PCTEXCAV variable was close to being accept in part d, we should drop it from further use.

The criteria values for the model selected would be:

- $R^2_{adj} = 0.9745$
- $Cp = 5.2934$
- $AIC = 7799.6943$

h

Build a complete first order model with interaction term. Would you suggest this model for predicting Y? Explain.

```
full_interact_FLAG <- lm(data = FLAG, LOWBID ~ (DOTEST + STATUS + NUMIDS + DISTRICT)**2)
summary(full_interact_FLAG)
```

```
##
## Call:
## lm(formula = LOWBID ~ (DOTEST + STATUS + NUMIDS + DISTRICT)^2,
##     data = FLAG)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1486446   -52732     9513    46452  1477972
##
## Coefficients: (4 not defined because of singularities)
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -3.353e+04  7.480e+04  -0.448  0.65434
## DOTEST             1.097e+00  2.969e-02  36.955  < 2e-16 ***
## STATUS1           -1.199e+04  1.102e+05  -0.109  0.91342
## NUMIDS            -4.697e+03  1.273e+04  -0.369  0.71248
## DISTRICT2         -1.215e+04  1.653e+05  -0.073  0.94147
## DISTRICT3          9.037e+04  3.802e+05   0.238  0.81229
## DISTRICT4         -1.532e+06  6.568e+05  -2.332  0.02046 *
## DISTRICT5         -4.438e+04  9.666e+04  -0.459  0.64655
## DOTEST:STATUS1     9.451e-02  3.673e-02   2.573  0.01063 *
## DOTEST:NUMIDS     -1.934e-02  3.603e-03  -5.367 1.77e-07 ***
## DOTEST:DISTRICT2   3.988e-02  5.577e-02   0.715  0.47518
## DOTEST:DISTRICT3  -1.655e-01  5.168e-01  -0.320  0.74904
## DOTEST:DISTRICT4  -2.533e-02  6.268e-02  -0.404  0.68653
## DOTEST:DISTRICT5  -1.330e-01  2.870e-02  -4.636 5.64e-06 ***
## STATUS1:NUMIDS     1.043e+04  3.188e+04   0.327  0.74370
## STATUS1:DISTRICT2         NA         NA      NA       NA
## STATUS1:DISTRICT3         NA         NA      NA       NA
## STATUS1:DISTRICT4         NA         NA      NA       NA
## STATUS1:DISTRICT5  7.549e+04  7.891e+04   0.957  0.33964
## NUMIDS:DISTRICT2   6.114e+03  2.166e+04   0.282  0.77793
## NUMIDS:DISTRICT3          NA         NA      NA       NA
## NUMIDS:DISTRICT4   1.519e+05  4.661e+04   3.260  0.00126 **
## NUMIDS:DISTRICT5   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_reduced_FLAG <- lm(data = FLAG, LOWBID ~ DOTEST + DISTRICT + STATUS + NUMIDS + DOTEST * STATUS + DOTE
ST * NUMIDS + DOTEST * DISTRICT + NUMIDS * DISTRICT)
summary(interaction_reduced_FLAG)
```

```
##
## Call:
## lm(formula = LOWBID ~ DOTEST + DISTRICT + STATUS + NUMIDS + DOTEST *
##     STATUS + DOTEST * NUMIDS + DOTEST * DISTRICT + NUMIDS * DISTRICT,
##     data = FLAG)
##
## 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 ***
## DISTRICT2         2.458e+04  1.613e+05   0.152   0.8790
## DISTRICT3         6.326e+04  3.785e+05   0.167   0.8674
## DISTRICT4        -1.531e+06  6.557e+05  -2.334   0.0203 *
## DISTRICT5         1.572e+04  7.240e+04   0.217   0.8283
## STATUS1           6.156e+04  4.652e+04   1.323   0.1869
## NUMIDS            1.974e+02  1.181e+04   0.017   0.9867
## DOTEST:STATUS1    9.218e-02  3.580e-02   2.575   0.0106 *
## DOTEST:NUMIDS    -1.995e-02  3.549e-03  -5.622 4.82e-08 ***
## DOTEST:DISTRICT2  3.939e-02  5.566e-02   0.708   0.4798
## DOTEST:DISTRICT3 -1.326e-01  5.149e-01  -0.258   0.7970
## DOTEST:DISTRICT4 -2.532e-02  6.257e-02  -0.405   0.6861
## DOTEST:DISTRICT5 -1.335e-01  2.854e-02  -4.679 4.63e-06 ***
## DISTRICT2:NUMIDS  1.648e+03  2.119e+04   0.078   0.9381
## DISTRICT3:NUMIDS         NA         NA      NA       NA
## DISTRICT4:NUMIDS  1.513e+05  4.653e+04   3.252   0.0013 **
## DISTRICT5: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
```

The first order model would be:

$$\begin{aligned}\widehat{LOWBID} =& \hat{\beta}_0 + \hat{\beta}_1 DOTEST + \\ & \hat{\beta}_2 STATUS + \hat{\beta}_3 NUMIDS + \hat{\beta}_4 DISTRICT + \\ & \hat{\beta}_5 DOTEST * STATUS + \hat{\beta}_6 DOTEST * NUMIDS + \\ & \hat{\beta}_7 DOTEST * DISTRICT + \hat{\beta}_8 NUMIDS * DISTRICT\end{aligned}$$

```
bptest(interaction_reduced_FLAG)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  interaction_reduced_FLAG
## BP = 124.75, df = 16, p-value < 2.2e-16
```

```
shapiro.test(residuals(interaction_reduced_FLAG))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(interaction_reduced_FLAG)
## W = 0.67885, p-value < 2.2e-16
```

Since both the Breusch-Pagan Test and the Shapiro-Wilk test reject the null hypothesis, meaning heteroscedasticity is present in the first order model and the sample data is also not significantly normally distributed, the first order model should not be used to predict Y.

i

Compare the RMSE from the first order model in part (d) with the interation model in part (h). Interpret the result.

```
sigma(reduced_model_FLAG_d)
```

```
## [1] 279650.7
```

```
sigma(interaction_reduced_FLAG)
```

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

```
sigma(reduced_model_FLAG) - sigma(reduced_model_FLAG_d)
```

```
## [1] 2035.996
```

The RMSE for the model in part d is 2,036 units higher than the interaction model in part h. Our model preference in this case would be the reduced interaction model from part h as we prefer models with a minimum RMSE.

j

Find the R^2_{adj} and interpret the result from part h

```
summary(interaction_reduced_FLAG)$adj.r.squared
```

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

The R^2_{adj} for the model from part h is 0.981, meaning that 98.1% of the variance in the contractor's lowest bidding price can be explained by the model in part h.

Problem 3

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 cognitiveimpairment.*
7. *SOCIALSU: total score of perceived social support (25-175, higher values indicating more support).*

The reported data are in file KBI.csv.

a

Use stepwise regression (with stepwise selection) to find the “best” set of predictors of caregiver burden. [Hint: Use pent = 0.1 and prem = 0.3].

```
full_model_KBI <- lm(data = KBI, BURDEN ~ .)
ols_step_both_p(full_model_KBI, pent = 0.1, prem = 0.3)
```

```
## Stepwise Selection Method
## --------------------------
##
## Candidate Terms:
##
## 1. CGAGE
## 2. CGINCOME
## 3. CGDUR
## 4. ADL
## 5. MEM
## 6. COG
## 7. SOCIALSU
##
## We are selecting variables based on p value...
##
## Variables Entered/Removed:
##
## - MEM added
## - SOCIALSU added
## - CGDUR added
##
## No more variables to be added/removed.
##
##
## Final Model Output
## ------------------
##
##                          Model Summary
## ----------------------------------------------------------------
## R                       0.663       RMSE                15.246
## R-Squared               0.440       Coef. Var           22.019
## Adj. R-Squared          0.422       MSE                232.444
## Pred R-Squared          0.386       MAE                 11.866
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## -------------------------------------------------------------------
```

```
##                  Sum of
##                 Squares        DF    Mean Square      F         Sig.
## ----------------------------------------------------------------------
## Regression    17513.638         3       5837.879    25.115    0.0000
## Residual      22314.602        96        232.444
## Total         39828.240        99
## ----------------------------------------------------------------------
##
##                                  Parameter Estimates
## ------------------------------------------------------------------------------------------
##       model       Beta    Std. Error    Std. Beta      t        Sig      lower      upper
## ------------------------------------------------------------------------------------------
## (Intercept)    115.539        12.368                  9.342    0.000    90.989    140.090
##         MEM      0.566         0.102        0.432     5.533    0.000     0.363      0.769
##    SOCIALSU     -0.492         0.089       -0.426    -5.514    0.000    -0.670     -0.315
##       CGDUR      0.122         0.065        0.146     1.876    0.064    -0.007      0.250
## ------------------------------------------------------------------------------------------
```

```
##
##                             Stepwise Selection Summary
## ------------------------------------------------------------------------------
##                      Added/                   Adj.
## Step    Variable     Removed     R-Square    R-Square     C(p)        AIC        RMSE
## ------------------------------------------------------------------------------
##    1       MEM       addition       0.252       0.244    29.7080    859.4694    17.4355
##    2     SOCIALSU    addition       0.419       0.407     3.6100    836.1716    15.4429
##    3      CGDUR      addition       0.440       0.422     2.1570    834.5703    15.2461
## ------------------------------------------------------------------------------
```

From the stepwise regression procedure, the following independent variables are selected as the best predictors:

- MEM: memory and behavioral problems with higher scores indicating more problems
- SOCIALSU: total score of perceived social support (25-175, higher values indicating more support)
- CGDUR: caregiver-duration of caregiving (month)

b

Use backward elimination regression to find the “best” set of predictors of caregiver burden. [Hint: Use prem = 0.1]

```
ols_step_backward_p(full_model_KBI, prem = 0.1)
```

```
## Backward Elimination Method
## --------------------------
##
## Candidate Terms:
##
## 1 . CGAGE
## 2 . CGINCOME
## 3 . CGDUR
## 4 . ADL
## 5 . MEM
## 6 . COG
## 7 . SOCIALSU
##
## We are eliminating variables based on p value...
##
## Variables Removed:
##
## - COG
## - CGINCOME
## - CGAGE
## - ADL
##
## No more variables satisfy the condition of p value = 0.1
##
##
## Final Model Output
## ------------------
##
##                          Model Summary
## ----------------------------------------------------------------
## R                       0.663       RMSE                15.246
## R-Squared               0.440       Coef. Var           22.019
## Adj. R-Squared          0.422       MSE                232.444
## Pred R-Squared          0.386       MAE                 11.866
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
```

```
## ----------------------------------------------------------------------
##                  Sum of
##                 Squares        DF    Mean Square      F         Sig.
## ----------------------------------------------------------------------
## Regression    17513.638         3       5837.879    25.115    0.0000
## Residual      22314.602        96        232.444
## Total         39828.240        99
## ----------------------------------------------------------------------
##
##                                  Parameter Estimates
## ------------------------------------------------------------------------------------------
##       model       Beta    Std. Error    Std. Beta      t        Sig      lower      upper
## ------------------------------------------------------------------------------------------
## (Intercept)    115.539        12.368                  9.342    0.000    90.989    140.090
##       CGDUR      0.122         0.065        0.146     1.876    0.064    -0.007      0.250
##         MEM      0.566         0.102        0.432     5.533    0.000     0.363      0.769
##    SOCIALSU     -0.492         0.089       -0.426    -5.514    0.000    -0.670     -0.315
## ------------------------------------------------------------------------------------------
```

```
##
##
##                           Elimination Summary
## ------------------------------------------------------------------------
##         Variable                  Adj.
## Step    Removed     R-Square    R-Square     C(p)       AIC        RMSE
## ------------------------------------------------------------------------
##    1    COG            0.452      0.4166    6.0981    838.3589    15.3197
##    2    CGINCOME      0.4511       0.422    4.2386    836.5114    15.2496
##    3    CGAGE         0.4473      0.4241    2.8795    835.2038    15.2218
##    4    ADL           0.4397      0.4222    2.1575    834.5703    15.2461
## ------------------------------------------------------------------------
```

From the backward regression procedure, the following independent variables are selected as the best predictors:

- MEM: memory and behavioral problems with higher scores indicating more problems
- SOCIALSU: total score of perceived social support (25-175, higher values indicating more support)
- CGDUR: caregiver-duration of caregiving (month)

C

Use forward elimination regression to find the “best” set of predictors of caregiver burden. [Hint: Use pent = 0.1]

```
ols_step_forward_p(full_model_KBI, pent = 0.1)
```

```
## Forward Selection Method
## --------------------------
##
## Candidate Terms:
##
## 1. CGAGE
## 2. CGINCOME
## 3. CGDUR
## 4. ADL
## 5. MEM
## 6. COG
## 7. SOCIALSU
##
## We are selecting variables based on p value...
##
## Variables Entered:
##
## - MEM
## - SOCIALSU
## - CGDUR
##
## No more variables to be added.
##
## Final Model Output
## ------------------
##
##                          Model Summary
## ----------------------------------------------------------------
## R                       0.663       RMSE                15.246
## R-Squared               0.440       Coef. Var           22.019
## Adj. R-Squared          0.422       MSE                232.444
## Pred R-Squared          0.386       MAE                 11.866
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## -------------------------------------------------------------------
##                  Sum of
```

```
##                Squares        DF    Mean Square      F         Sig.
## ----------------------------------------------------------------------
## Regression    17513.638         3       5837.879    25.115    0.0000
## Residual      22314.602        96        232.444
## Total         39828.240        99
## ----------------------------------------------------------------------
##
##                                  Parameter Estimates
## ------------------------------------------------------------------------------------------
##       model       Beta    Std. Error    Std. Beta      t        Sig      lower      upper
## ------------------------------------------------------------------------------------------
## (Intercept)    115.539        12.368                 9.342    0.000    90.989    140.090
##         MEM      0.566         0.102         0.432    5.533    0.000     0.363      0.769
##    SOCIALSU     -0.492         0.089        -0.426   -5.514    0.000    -0.670     -0.315
##       CGDUR      0.122         0.065         0.146    1.876    0.064    -0.007      0.250
## ------------------------------------------------------------------------------------------
```

```
##
##                            Selection Summary
## ------------------------------------------------------------------------
##         Variable                   Adj.
## Step    Entered     R-Square    R-Square     C(p)        AIC         RMSE
## ------------------------------------------------------------------------
##    1    MEM           0.2520      0.2444    29.7076    859.4694    17.4355
##    2    SOCIALSU      0.4192      0.4072     3.6101    836.1716    15.4429
##    3    CGDUR         0.4397      0.4222     2.1575    834.5703    15.2461
## ------------------------------------------------------------------------
```

From the forward regression procedure, the following independent variables are selected as the best predictors:

- MEM: memory and behavioral problems with higher scores indicating more problems
- SOCIALSU: total score of perceived social support (25-175, higher values indicating more support)
- CGDUR: caregiver-duration of caregiving (month)

d

Use all-possible-regressions-selection to find the “best” predictors of caregiver burden (Cp, AIC, Adjusted R^2, R^2)

```
ols_step_best_subset(full_model_KBI, details = TRUE)
```

	mindex	n	predictors	rsquare	adjr	predrsq	cp
	<int>	<int>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
5	1	1	MEM	0.2519944	0.2443617	0.2244252	29.707640
27	2	2	MEM SOCIALSU	0.4191848	0.4072092	0.3799795	3.610120
58	3	3	CGDUR MEM SOCIALSU	0.4397292	0.4222207	0.3864711	2.157489
95	4	4	CGDUR ADL MEM SOCIALSU	0.4473335	0.4240633	0.3831389	2.879523
110	5	5	CGAGE CGDUR ADL MEM SOCIALSU	0.4511470	0.4219527	0.3782262	4.238638
121	6	6	CGAGE CGINCOME CGDUR ADL MEM SOCIALSU	0.4519831	0.4166272	0.3129164	6.098124
127	7	7	CGAGE CGINCOME CGDUR ADL MEM COG SOCIALSU	0.4525670	0.4109145	0.2989164	8.000000

7 rows | 1-8 of 15 columns

The the subsets regression summary, the lowest Cp and the second highest R^2_{adj} is Model 3 which includes the following predictors:

- MEM: memory and behavioral problems with higher scores indicating more problems
- SOCIALSU: total score of perceived social support (25-175, higher values indicating more support)
- CGDUR: caregiver-duration of caregiving (month)

These predictors are considered the strongest to predict the caregiver burden.

e

Compare the results, parts a-c. Which independent variables consistently are selected as the "best" predictors? Comment on the value of the adjusted R^2. The independent variables consistently selected as the best predictors are:

- MEM: memory and behavioral problems with higher scores indicating more problems
- SOCIALSU: total score of perceived social support (25-175, higher values indicating more support)
- CGDUR: caregiver-duration of caregiving (month)

The R^2_{adj} is 0.4222 which means only 42% of the variance in the caregiver burden can be explained by this model.

f

Explain how you would use the results, parts a-c, to develop a model for caregiver burden. Check for interactions, normality and linearity assumptions. From the best predictors found in parts a-c, the interaction terms betwen those predictors should be tested as follows:

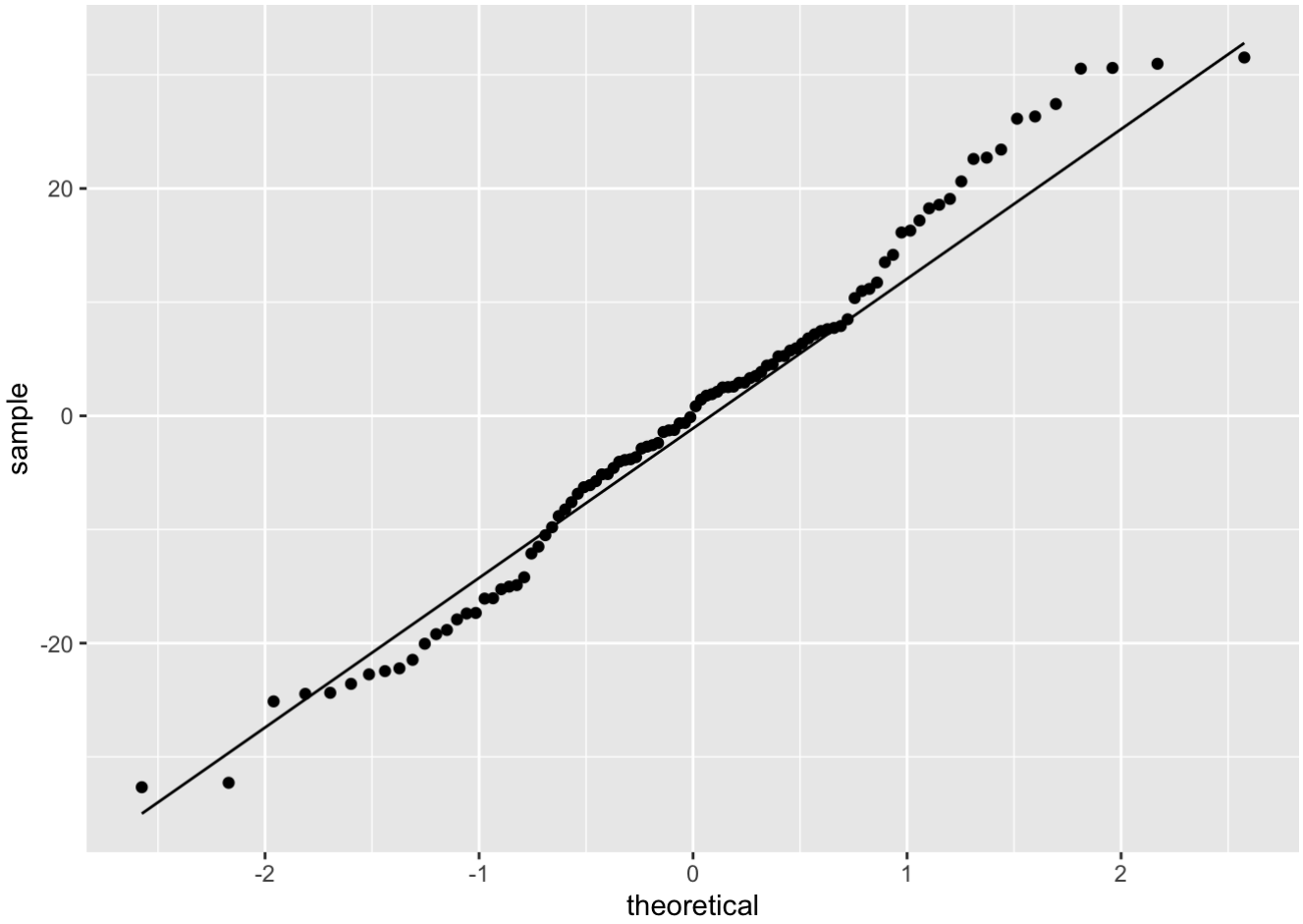
```
reduced_model_KBI <- lm(data = KBI, BURDEN ~ MEM + SOCIALSU + CGDUR)
interaction_KBI <- lm(data = KBI, BURDEN ~ (MEM + SOCIALSU + CGDUR)**2)
summary(interaction_KBI)
```

```
##
## Call:
## lm(formula = BURDEN ~ (MEM + SOCIALSU + CGDUR)^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 ***
## MEM             0.869719   0.790027   1.101 0.273793
## SOCIALSU       -0.341339   0.210830  -1.619 0.108828
## CGDUR           0.350722   0.525520   0.667 0.506181
## MEM:SOCIALSU   -0.002998   0.006087  -0.492 0.623553
## MEM:CGDUR       0.003782   0.004228   0.894 0.373411
## SOCIALSU:CGDUR -0.002564   0.004042  -0.634 0.527485
## ---
## 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
```

As there are no interactions that are significant (p-value below 0.05), interactions will not be included in this model.

Next, a qqplot and the Shapiro-Wilk test will be used to check for normality:

```
KBI %>% ggplot(aes(sample = reduced_model_KBI$residuals)) +
  stat_qq() +
  stat_qq_line()
```




```
shapiro.test(residuals(reduced_model_KBI))
```

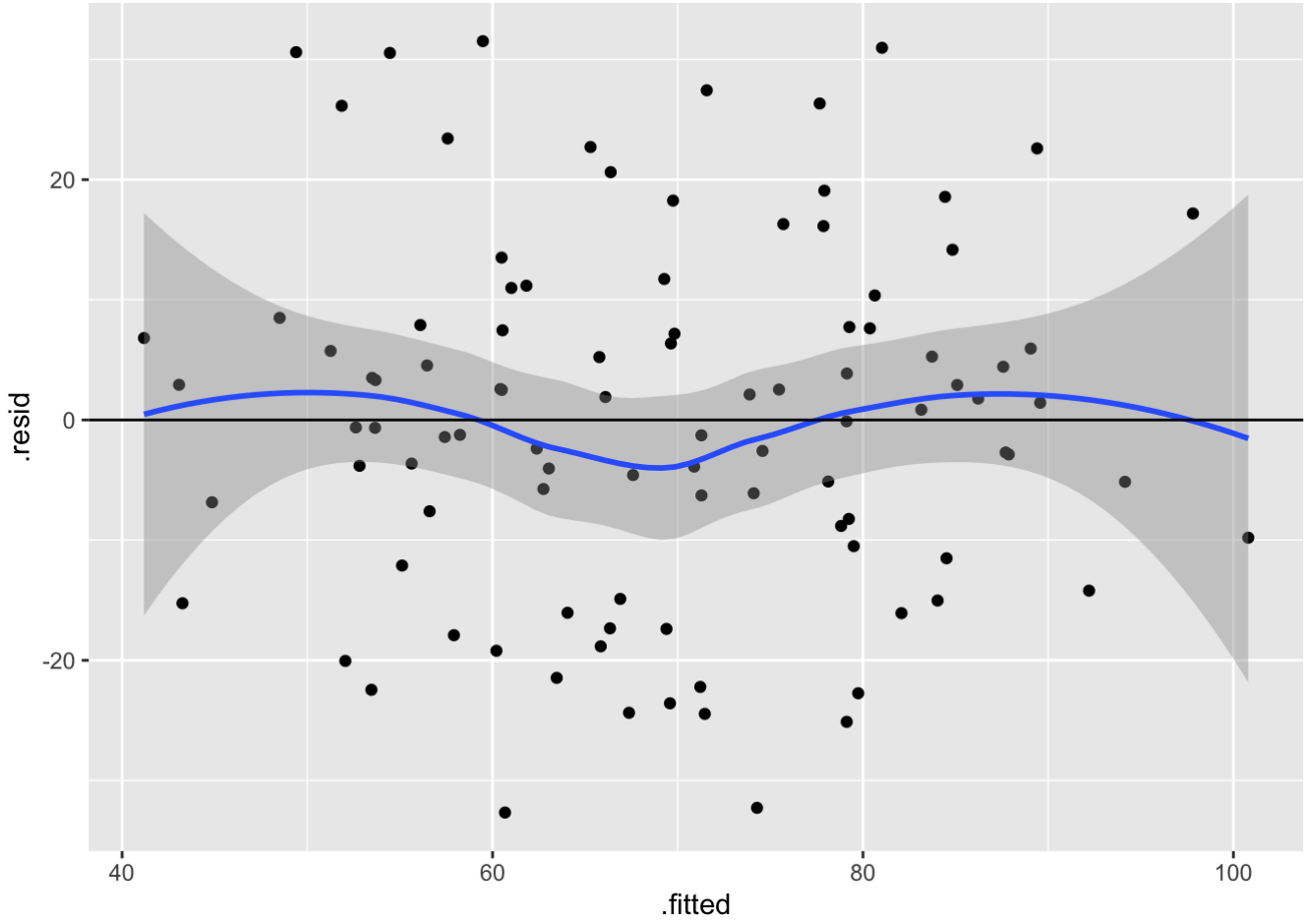
```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(reduced_model_KBI)
## W = 0.98407, p-value = 0.2716
```

Based on the qqplot, we can see the residuals of the reduced additive model follow a linear trend along the qq line. Additionally, the p-value from the Shapiro-Wilk test is greater than 0.05 which confirms that the residuals of this model follow a normal distribution.

Next, a residual plot will be used to check the linearity assumption:

```
ggplot(reduced_model_KBI, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_smooth() +
  geom_hline(yintercept = 0)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

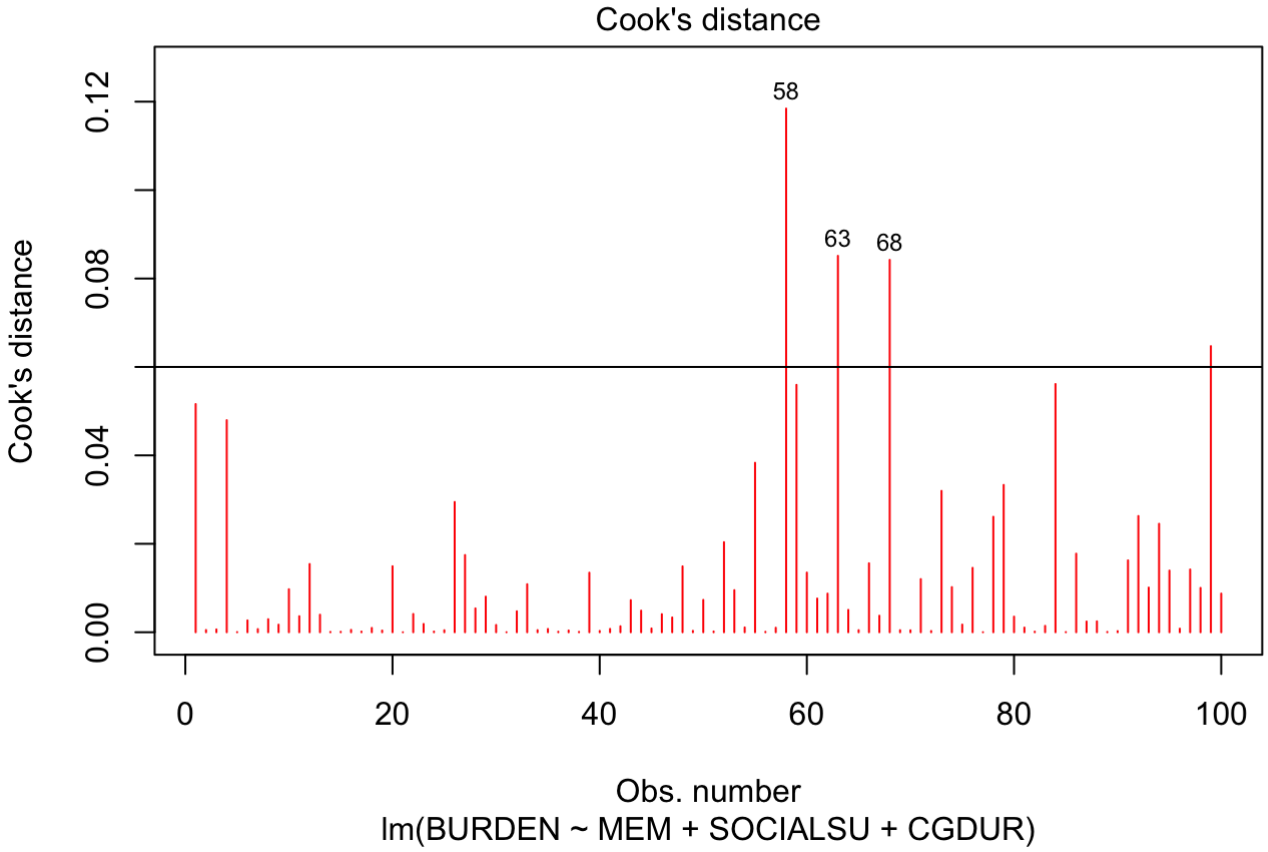


Based on the residual plot above, we can see there is no general pattern. The linearity assumption holds.

g

Do you detect any outliers by using leverage values greater than $\frac{2p}{n}$? Remove those outliers and fit again the model with variables selected in question a

```
n_KBI <- 100
p_KBI <- 3
plot(reduced_model_KBI, pch = 18, col = "red", which = 4)
abline(h = 2 * p_KBI / n_KBI, lty = 1)
```



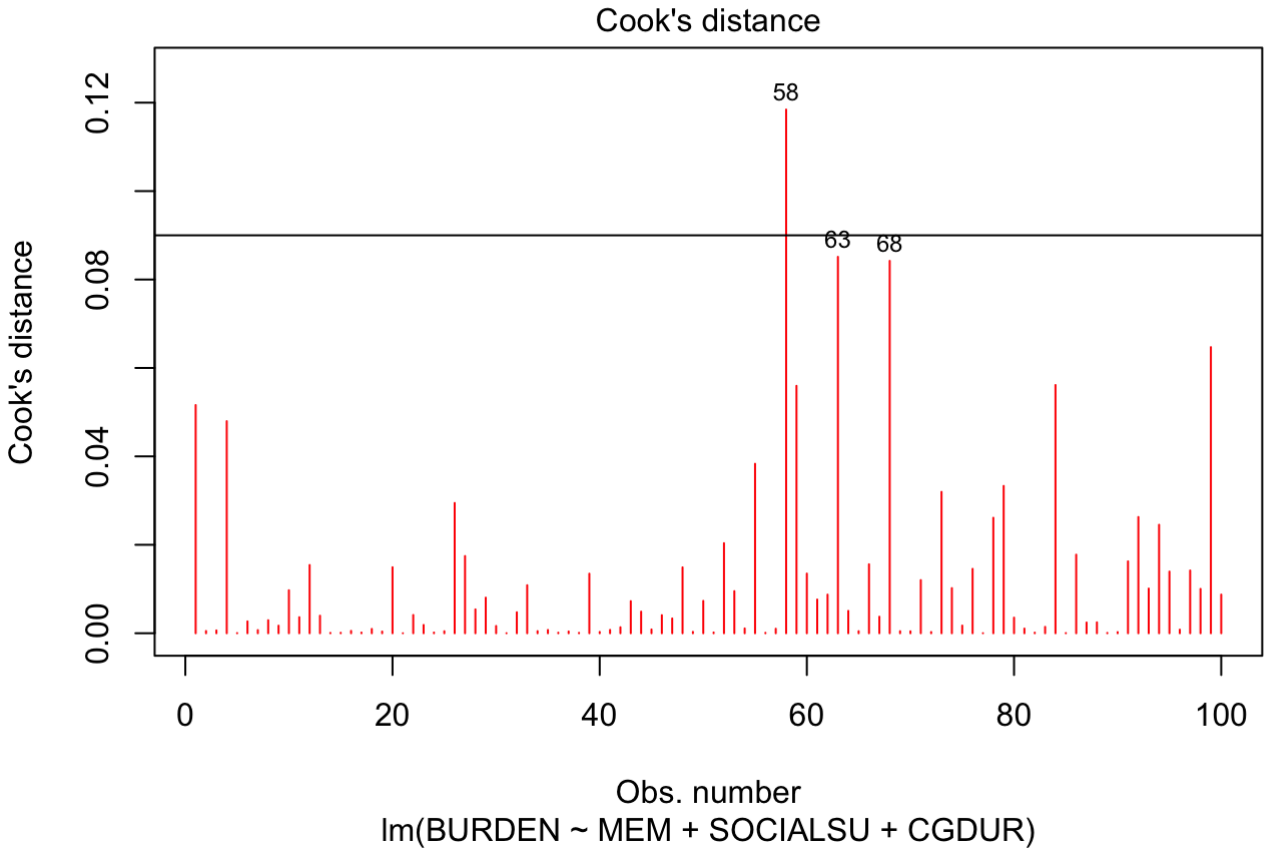
Observations 58, 62 and 68 are above the $\frac{2p}{n}$ cutoff. These observation will be dropped.

```
KBI_reduced1 <- KBI[-c(58, 62, 68), ]
reduced_model_KBI1 <- lm(data = KBI_reduced1, BURDEN ~ MEM + SOCIALSU + CGDUR)
```

h

Do you detect any outliers by using leverage values greater than $\frac{3p}{n}$? Remove those outliers and fit again the model with variables selected in question a

```
plot(reduced_model_KBI, pch = 18, col = "red", which = 4)
abline(h = 3 * p_KBI / n_KBI, lty = 1)
```



Observation 58 is above the $\frac{3p}{n}$ cutoff. This observation will be dropped.

```
KBI_reduced2 <- KBI[-c(58), ]
reduced_model_KBI2 <- lm(data = KBI_reduced2, BURDEN ~ MEM + SOCIALSU + CGDUR)
```

i

Do you notice any difference in the results with the model from part a and the best fit model between part g and h? Comment.

```
summary(reduced_model_KBI1)$adj.r.squared
```

```
## [1] 0.4653651
```

```
summary(reduced_model_KBI2)$adj.r.squared
```

```
## [1] 0.4357068
```

```
summary(reduced_model_KBI)$adj.r.squared
```

```
## [1] 0.4222207
```

In terms of R^2_{adj}, the model improves as you take out the observation outliers. These outliers have a significant enough impact on the model that they should be removed from the best fitted model.