Statistics | HW2

Problem 3: Linear regression analysis

Task 1

```
data = read.csv('student-mat.csv',sep =";")
data$G1 = NULL
data$G2 = NULL
```

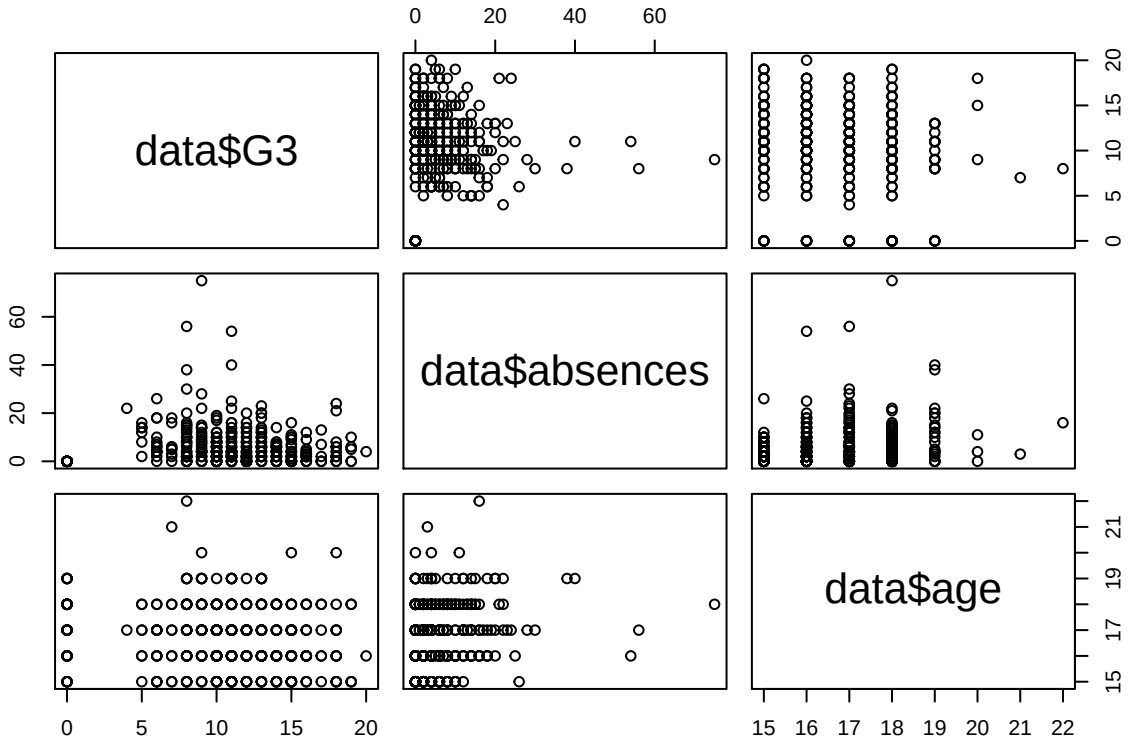
1) Mjob can have one of 5 values 'teacher', 'health', 'services', 'at_home' or 'other'. It's clear that Mjob is a nominal scale variable. If variable is nominal scaled, then we can only conclude about equality or inequality. For example, that Mjob='teacher' does not equal Mjob='health'. We can't order values. We can't include nominal scale variables into model, so we should include Mjob with dummies.
   A dummy variable takes the value one for some observations to indicate the presence of an effect or membership in a group and zero for the remaining observations. So, in case of Mjob, we can have 4 dummies: Mjob_teacher, Mjob_health, Mjob_services, Mjob_at_home. We should'n include dummy variables for all possible cases, beacause of colinearity of columns of matrix X. If all listed above dummies are equal to 0, that means that Mjob is 'other'.
2) Goout variable can have numeric values from 1 to 5(1 - very low to 5 - very high). We can see that goout is an ordinal variable, because we can naturally order the values. Since we can't provide a natural meaning of differences between two values, goout is not an interval scaled variable. For example, goout=1 is less than goout=2. We can include ordinal scale variables into model without dummies.

All other nominal scale variables, such as Fjob, address, etc. should be included into model with dummies.

Task 2

The only interval scaled variables that we have are absenses and age.
As we can see from scatter plot, absenses seem to have nonlinear impact on G3.
So, let's run RESET test to decide if we should include powers(2,3) of absenses and age into model.

```
pairs(data$G3 ~ data$absences+data$age)
```



We want to tests whether non-linear combinations of the fitted values help explain the response variable. As we can see, adding absences of power 2 and 3 will have significant influence. On the other hand, adding powers of age will not.

```
library("lmtest")
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
resettest(lm(G3~absences, data=data), power=2:3, type="fitted")
```

```
##
##  RESET test
##
## data:  lm(G3 ~ absences, data = data)
## RESET = 5.1655, df1 = 2, df2 = 391, p-value = 0.006106
```

```
resettest(lm(G3~age, data=data), power=2:3, type="fitted")
```

```
##
##  RESET test
##
## data:  lm(G3 ~ age, data = data)
```

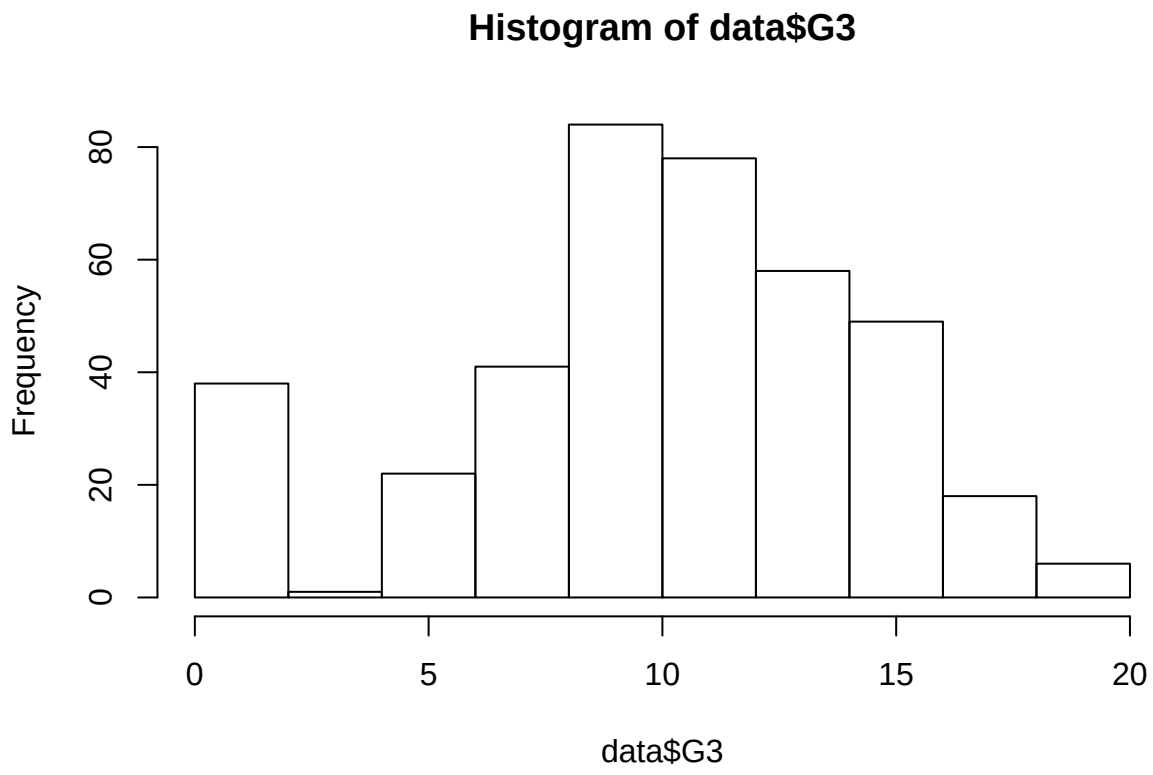
```
## RESET = 0.2666, df1 = 2, df2 = 391, p-value = 0.7661
```

So, let's add absences2 and absences3 to our data.

```
data$absences2 <- data$absences * data$absences
data$absences3 <- data$absences * data$absences * data$absences
```

Let's look at y variable(G3). As we can see from histogram and measure of skewness, G3 is almost normally distributed with some left skeweness. There is no need to transform G3 variable.

```
library(moments)
hist(data$G3)
```




```
cat("Measure of skeweness: ", skewness(data$G3))
```

```
## Measure of skeweness:  -0.7298871
```

Task 3

From this summary of linear regression we can see that some variables such as failures have extremely low p-value and therefore are very significant.
Coefficients that have p-value larger than 0.5 we consider as insignificant. We can also see estimated values for coefficients, std errors and t values. Our model has R-squared: 0.2965 and Adjusted R-squared: 0.2148. In futher model selection we would look at adjusted R-squared, because R-squared increases if number of variables increases.

```
g3.model <- lm(G3~., data=data)
summary(g3.model)
```

```
##
## Call:
## lm(formula = G3 ~ ., data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.4742  -1.9967   0.1044   2.8734   8.5870
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.407e+01  4.480e+00   3.142 0.001822 **
## schoolMS          8.087e-01  7.828e-01   1.033 0.302225
## sexM              1.201e+00  4.947e-01   2.428 0.015693 *
## age              -4.181e-01  2.154e-01  -1.941 0.053032 .
## addressU          6.635e-01  5.811e-01   1.142 0.254308
## famsizeLE3        6.557e-01  4.828e-01   1.358 0.175270
## PstatusT         -1.940e-01  7.223e-01  -0.269 0.788378
## Medu              4.039e-01  3.198e-01   1.263 0.207438
## Fedu             -9.848e-03  2.760e-01  -0.036 0.971551
## Mjobhealth        1.069e+00  1.108e+00   0.964 0.335591
## Mjobother        -2.102e-01  7.080e-01  -0.297 0.766721
## Mjobservices      6.222e-01  7.886e-01   0.789 0.430674
## Mjobteacher      -1.233e+00  1.026e+00  -1.201 0.230547
## Fjobhealth        2.629e-01  1.422e+00   0.185 0.853421
## Fjobother        -5.976e-01  1.012e+00  -0.590 0.555327
## Fjobservices     -2.721e-01  1.046e+00  -0.260 0.794925
## Fjobteacher       1.278e+00  1.282e+00   0.998 0.319190
## reasonhome        9.125e-02  5.474e-01   0.167 0.867694
## reasonother       5.529e-01  8.114e-01   0.681 0.496058
## reasonreputation  5.768e-01  5.706e-01   1.011 0.312773
## guardianmother    1.188e-01  5.394e-01   0.220 0.825780
## guardianother     7.214e-01  9.905e-01   0.728 0.466895
## traveltime       -1.898e-01  3.354e-01  -0.566 0.571833
## studytime         5.441e-01  2.849e-01   1.909 0.057015 .
## failures         -1.718e+00  3.313e-01  -5.185 3.65e-07 ***
## schoolsupyes     -1.394e+00  6.593e-01  -2.114 0.035244 *
## famsupyes        -8.814e-01  4.739e-01  -1.860 0.063732 .
## paidyes           3.585e-01  4.722e-01   0.759 0.448217
## activitiesyes    -3.923e-01  4.402e-01  -0.891 0.373450
## nurseryyes       -2.115e-01  5.489e-01  -0.385 0.700241
## higheryes         1.014e+00  1.102e+00   0.920 0.358070
## internetyes       4.695e-01  6.125e-01   0.767 0.443865
## romanticyes      -1.015e+00  4.650e-01  -2.183 0.029675 *
## famrel            3.108e-01  2.444e-01   1.272 0.204279
## freetime          2.866e-01  2.348e-01   1.221 0.223050
## goout            -6.469e-01  2.226e-01  -2.906 0.003886 **
## Dalc             -3.113e-01  3.273e-01  -0.951 0.342104
## Walc              2.206e-01  2.461e-01   0.897 0.370594
## health           -1.529e-01  1.596e-01  -0.958 0.338545
## absences          3.269e-01  9.049e-02   3.612 0.000347 ***
## absences2        -1.329e-02  4.910e-03  -2.707 0.007119 **
## absences3         1.265e-04  5.668e-05   2.231 0.026302 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 4.06 on 353 degrees of freedom
## Multiple R-squared:  0.2965, Adjusted R-squared:  0.2148
## F-statistic: 3.629 on 41 and 353 DF,  p-value: 2.661e-11
```

As we can see from summary of lm, Fjob seems to be insignificant. So we need to check the simultaneous insignificance of all dummies.
For that we check linear hypothesis. H0 is that Fjobhealth, Fjobother, Fjobservices, Fjobteacher are simultaneously equal to zero.
As we can see the p-value is > 0.05, so do not reject the H0 hypothesis and can conclude that we all our dummies for Fjob are simultaneously insignificant.

```
library("car")
linearHypothesis(g3.model, c("Fjobother", "Fjobhealth", "Fjobservices", "Fjobteacher"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## Fjobother = 0
## Fjobhealth = 0
## Fjobservices = 0
## Fjobteacher = 0
##
## Model 1: restricted model
## Model 2: G3 ~ school + sex + age + address + famsize + Pstatus + Medu +
##     Fedu + Mjob + Fjob + reason + guardian + traveltime + studytime +
##     failures + schoolsup + famsup + paid + activities + nursery +
##     higher + internet + romantic + famrel + freetime + goout +
##     Dalc + Walc + health + absences + absences2 + absences3
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    357 5889.3
## 2    353 5817.6  4    71.644 1.0868 0.3628
```

Task 4

age coefficient: 0.-4181. That means for fixed other variables, if you are 1 year older than your G3 decreases on 0.4181.
Fjob is represented with dummy variables: Fjobother: 0.-5976
Fjobhealth: 0.2629.
Fjobservices: -2.721e-01 Fjobteacher: 1.278e+0
This means, for example, for fixed other variables, If student's Fjob is health, he will have G3 by 0.2629 larger than comparing to having another Fjob.
goout coefficient: 0.-6469. That means for fixed other variables, if you have goout increased by 1, you will get G3 decreased by 0.-6469, which is very logical. The more you going out, the works marks you will have.

Task 5

Confidence intervals for famsupyes and absences.
We can see that with probability of 95% coefficient for absences will be from 0.1489176 to 0.50483572. It means that increase in absences will increase G3(if other variables are fixed) which is kind of a strange.
We can see that with probability of 95% coefficient for famsupyes will be from -1.8134225 to 0.05061711. If confint containes 0, that means that variable if insignificant, because there is rather high probability for

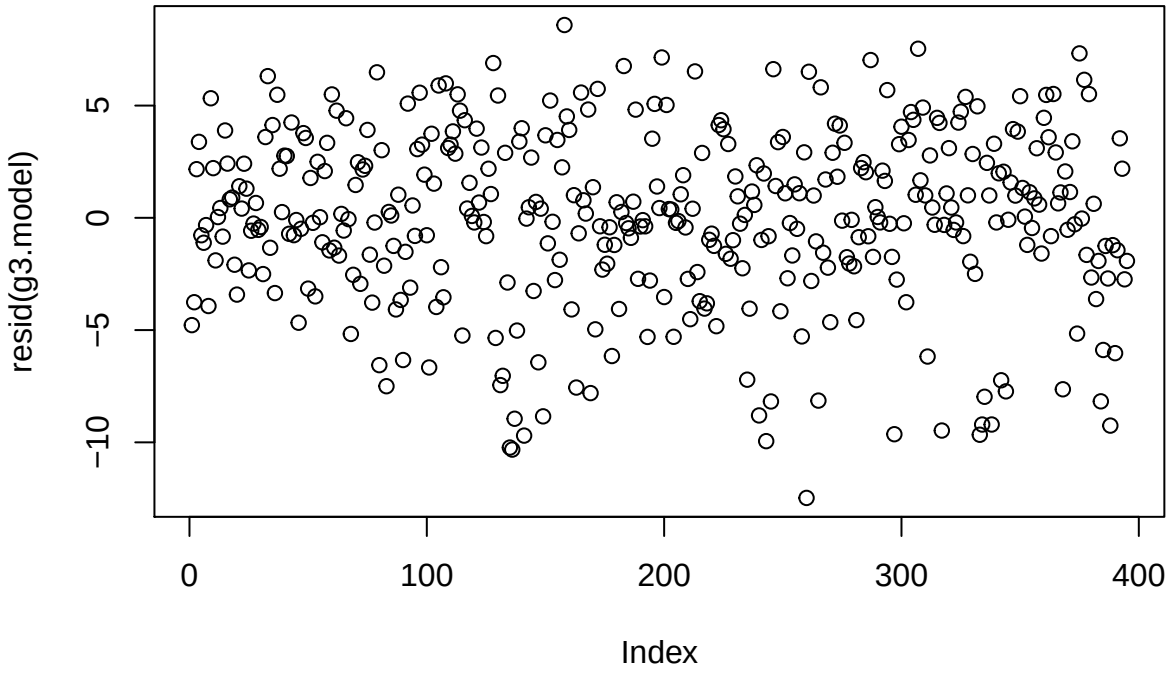
coefficient to be 0. If 0 is not in confint, that means, that variable is significant.(if we choose p-value threshold as 0.05 and take into account 95% confints)

```
confint(g3.model, c("famsupyes", "absences"))
```

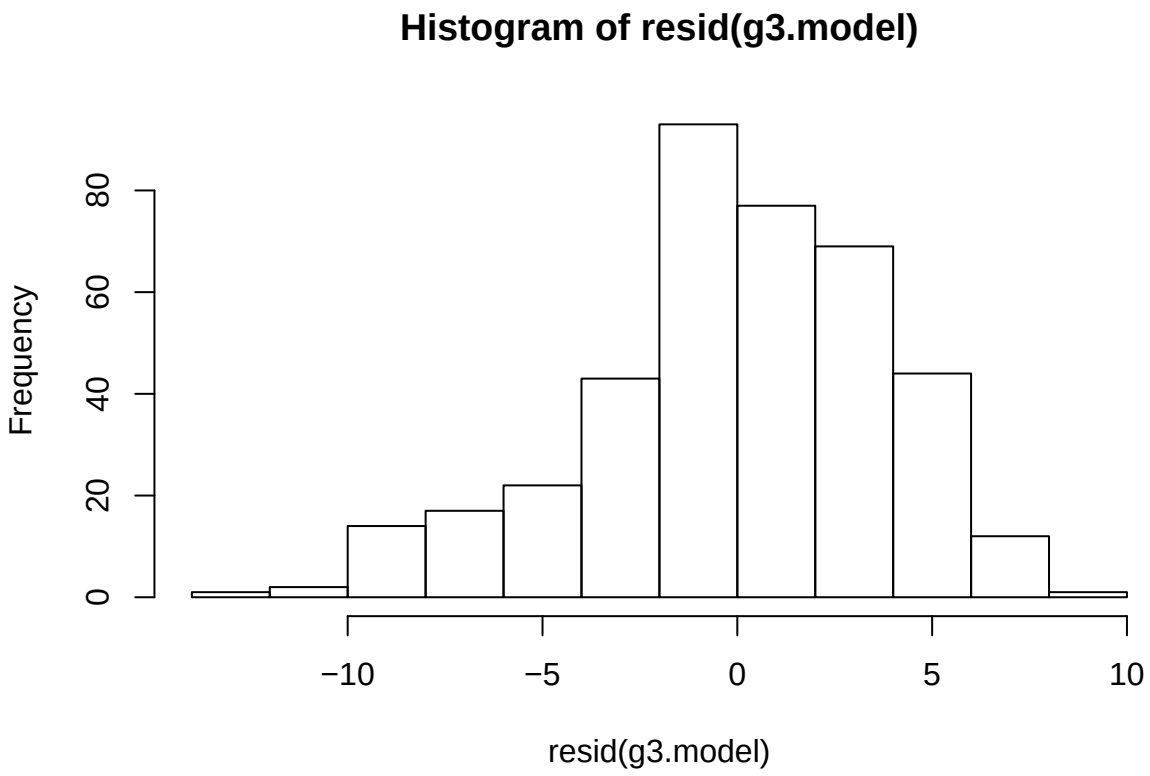
```
##                2.5 %     97.5 %
## famsupyes -1.8134225 0.05061711
## absences   0.1489176 0.50483572
```

Here we can see plotted residuals, we can see that they are no patterns. Also we can see from hist that residuals seems to follow normal distribution.
In addition we run a Kolmogorov-Smirnov test and see that our residuals follow normal distribution.

```
plot(resid(g3.model))
```




```
hist(resid(g3.model))
```




```
ks.test(resid(g3.model), mean(resid(g3.model)), sd(resid(g3.model)))
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  resid(g3.model) and mean(resid(g3.model))
## D = 0.51392, p-value = 0.9747
## alternative hypothesis: two-sided
```

Task 6

Here we use stepwise model selection based on AIC.
AIC can be thought of goodness of fit minus complexity.
AIC estimates the quality of each model, relative to each of the other models.
On each step we calculate AIC for model, and AIC for pissible models as if we drop one variable.
Then we choose to drop coefficient that will lead to the smallest AIC. If we can't find smallest AIC - stop.

```
selected_model <- step(g3.model)
```

```
## Start:  AIC=1146.45
## G3 ~ school + sex + age + address + famsize + Pstatus + Medu +
##     Fedu + Mjob + Fjob + reason + guardian + traveltime + studytime +
##     failures + schoolsup + famsup + paid + activities + nursery +
##     higher + internet + romantic + famrel + freetime + goout +
##     Dalc + Walc + health + absences + absences2 + absences3
##
```

```
##              Df Sum of Sq    RSS    AIC
## - reason      3     22.32 5839.9 1142.0
## - guardian    2      8.86 5826.5 1143.1
## - Fjob        4     71.64 5889.3 1143.3
## - Fedu        1      0.02 5817.6 1144.5
## - Pstatus     1      1.19 5818.8 1144.5
## - nursery     1      2.45 5820.1 1144.6
## - traveltime  1      5.28 5822.9 1144.8
## - paid        1      9.50 5827.1 1145.1
## - internet    1      9.68 5827.3 1145.1
## - activities  1     13.09 5830.7 1145.3
## - Walc        1     13.25 5830.9 1145.3
## - higher      1     13.96 5831.6 1145.4
## - Dalc        1     14.91 5832.5 1145.5
## - health      1     15.14 5832.7 1145.5
## - school      1     17.59 5835.2 1145.7
## - address     1     21.49 5839.1 1145.9
## - freetime    1     24.55 5842.2 1146.1
## - Medu        1     26.29 5843.9 1146.2
## - famrel      1     26.66 5844.3 1146.3
## <none>                    5817.6 1146.5
## - famsize     1     30.40 5848.0 1146.5
## - famsup      1     57.01 5874.6 1148.3
## - studytime   1     60.09 5877.7 1148.5
## - Mjob        4    151.31 5968.9 1148.6
## - age         1     62.10 5879.7 1148.7
## - schoolsup   1     73.63 5891.2 1149.4
## - romantic    1     78.56 5896.2 1149.8
## - absences3   1     82.04 5899.6 1150.0
## - sex         1     97.13 5914.7 1151.0
## - absences2   1    120.77 5938.4 1152.6
## - goout       1    139.22 5956.8 1153.8
## - absences    1    215.07 6032.7 1158.8
## - failures    1    443.03 6260.6 1173.5
##
## Step:  AIC=1141.97
## G3 ~ school + sex + age + address + famsize + Pstatus + Medu +
##     Fedu + Mjob + Fjob + guardian + traveltime + studytime +
##     failures + schoolsup + famsup + paid + activities + nursery +
##     higher + internet + romantic + famrel + freetime + goout +
##     Dalc + Walc + health + absences + absences2 + absences3
##
##              Df Sum of Sq    RSS    AIC
## - guardian    2      9.26 5849.2 1138.6
## - Fjob        4     79.55 5919.5 1139.3
## - Fedu        1      0.15 5840.1 1140.0
## - Pstatus     1      1.02 5840.9 1140.0
## - nursery     1      2.21 5842.1 1140.1
## - traveltime  1      6.07 5846.0 1140.4
## - internet    1      9.49 5849.4 1140.6
## - activities  1      9.54 5849.5 1140.6
## - higher      1     11.69 5851.6 1140.8
## - paid        1     12.06 5852.0 1140.8
## - Walc        1     14.21 5854.1 1140.9
```

```
## - Dalc        1     14.87 5854.8 1141.0
## - school      1     16.77 5856.7 1141.1
## - address     1     16.83 5856.8 1141.1
## - health      1     21.66 5861.6 1141.4
## - freetime    1     25.58 5865.5 1141.7
## - famrel      1     27.48 5867.4 1141.8
## - Medu        1     29.29 5869.2 1141.9
## <none>                    5839.9 1142.0
## - famsize     1     30.39 5870.3 1142.0
## - famsup      1     57.18 5897.1 1143.8
## - age         1     62.70 5902.6 1144.2
## - studytime   1     64.65 5904.6 1144.3
## - schoolsup   1     75.52 5915.4 1145.0
## - romantic    1     79.24 5919.2 1145.3
## - absences3   1     81.53 5921.5 1145.4
## - Mjob        4    174.01 6013.9 1145.6
## - sex         1     92.03 5932.0 1146.1
## - absences2   1    122.44 5962.4 1148.2
## - goout       1    148.52 5988.4 1149.9
## - absences    1    224.87 6064.8 1154.9
## - failures    1    460.15 6300.1 1169.9
##
## Step:  AIC=1138.59
## G3 ~ school + sex + age + address + famsize + Pstatus + Medu +
##     Fedu + Mjob + Fjob + traveltime + studytime + failures +
##     schoolsup + famsup + paid + activities + nursery + higher +
##     internet + romantic + famrel + freetime + goout + Dalc +
##     Walc + health + absences + absences2 + absences3
##
##              Df Sum of Sq    RSS    AIC
## - Fjob        4     75.33 5924.5 1135.7
## - Fedu        1      0.07 5849.3 1136.6
## - Pstatus     1      1.81 5851.0 1136.7
## - nursery     1      4.06 5853.3 1136.9
## - traveltime  1      5.21 5854.4 1136.9
## - internet    1      8.63 5857.8 1137.2
## - activities  1      9.70 5858.9 1137.2
## - Walc        1     11.55 5860.7 1137.4
## - paid        1     13.25 5862.4 1137.5
## - Dalc        1     13.83 5863.0 1137.5
## - higher      1     14.05 5863.2 1137.5
## - school      1     14.67 5863.9 1137.6
## - address     1     19.14 5868.3 1137.9
## - health      1     23.02 5872.2 1138.1
## - Medu        1     27.84 5877.0 1138.5
## - famrel      1     28.29 5877.5 1138.5
## - freetime    1     28.68 5877.9 1138.5
## - famsize     1     29.14 5878.3 1138.6
## <none>                    5849.2 1138.6
## - age         1     54.04 5903.2 1140.2
## - famsup      1     56.63 5905.8 1140.4
## - studytime   1     67.31 5916.5 1141.1
## - schoolsup   1     74.79 5924.0 1141.6
## - romantic    1     76.18 5925.4 1141.7
```

```
## - absences3   1     79.11 5928.3 1141.9
## - Mjob        4    174.21 6023.4 1142.2
## - sex         1     92.10 5941.3 1142.8
## - absences2   1    120.45 5969.6 1144.6
## - goout       1    155.55 6004.7 1147.0
## - absences    1    227.61 6076.8 1151.7
## - failures    1    456.90 6306.1 1166.3
##
## Step:  AIC=1135.65
## G3 ~ school + sex + age + address + famsize + Pstatus + Medu +
##     Fedu + Mjob + traveltime + studytime + failures + schoolsup +
##     famsup + paid + activities + nursery + higher + internet +
##     romantic + famrel + freetime + goout + Dalc + Walc + health +
##     absences + absences2 + absences3
##
##              Df Sum of Sq    RSS    AIC
## - Pstatus     1      2.32 5926.8 1133.8
## - nursery     1      3.76 5928.3 1133.9
## - Walc        1      4.04 5928.6 1133.9
## - traveltime  1      4.48 5929.0 1134.0
## - internet    1      5.29 5929.8 1134.0
## - Fedu        1      7.15 5931.7 1134.1
## - Dalc        1      7.35 5931.9 1134.1
## - paid        1      8.96 5933.5 1134.2
## - higher      1     11.74 5936.3 1134.4
## - activities  1     13.40 5937.9 1134.5
## - school      1     15.14 5939.7 1134.7
## - health      1     21.34 5945.9 1135.1
## - address     1     21.86 5946.4 1135.1
## - famrel      1     22.59 5947.1 1135.2
## - Medu        1     23.78 5948.3 1135.2
## - famsize     1     24.32 5948.8 1135.3
## - freetime    1     26.86 5951.4 1135.4
## <none>                    5924.5 1135.7
## - famsup      1     56.50 5981.0 1137.4
## - age         1     56.57 5981.1 1137.4
## - schoolsup   1     62.79 5987.3 1137.8
## - romantic    1     66.37 5990.9 1138.0
## - studytime   1     70.88 5995.4 1138.3
## - Mjob        4    165.96 6090.5 1138.6
## - absences3   1     88.03 6012.5 1139.5
## - sex         1     92.66 6017.2 1139.8
## - absences2   1    131.87 6056.4 1142.3
## - goout       1    152.03 6076.5 1143.7
## - absences    1    240.18 6164.7 1149.3
## - failures    1    449.97 6374.5 1162.6
##
## Step:  AIC=1133.8
## G3 ~ school + sex + age + address + famsize + Medu + Fedu + Mjob +
##     traveltime + studytime + failures + schoolsup + famsup +
##     paid + activities + nursery + higher + internet + romantic +
##     famrel + freetime + goout + Dalc + Walc + health + absences +
##     absences2 + absences3
##
```

```
##              Df Sum of Sq    RSS    AIC
## - nursery     1      3.43 5930.3 1132.0
## - Walc        1      3.94 5930.8 1132.1
## - traveltime  1      4.52 5931.4 1132.1
## - internet    1      4.65 5931.5 1132.1
## - Dalc        1      7.18 5934.0 1132.3
## - Fedu        1      7.49 5934.3 1132.3
## - paid        1      8.57 5935.4 1132.4
## - higher      1     12.44 5939.3 1132.6
## - school      1     14.77 5941.6 1132.8
## - activities  1     14.87 5941.7 1132.8
## - health      1     21.69 5948.5 1133.2
## - address     1     22.26 5949.1 1133.3
## - famrel      1     22.51 5949.4 1133.3
## - Medu        1     25.56 5952.4 1133.5
## - freetime    1     26.69 5953.5 1133.6
## - famsize     1     27.35 5954.2 1133.6
## <none>                    5926.8 1133.8
## - famsup      1     56.93 5983.8 1135.6
## - age         1     57.26 5984.1 1135.6
## - schoolsup   1     62.38 5989.2 1135.9
## - romantic    1     65.54 5992.4 1136.2
## - studytime   1     70.49 5997.3 1136.5
## - Mjob        4    167.48 6094.3 1136.8
## - sex         1     92.30 6019.1 1137.9
## - absences3   1     93.31 6020.1 1138.0
## - absences2   1    137.88 6064.7 1140.9
## - goout       1    152.05 6078.9 1141.8
## - absences    1    249.18 6176.0 1148.1
## - failures    1    447.91 6374.8 1160.6
##
## Step:  AIC=1132.03
## G3 ~ school + sex + age + address + famsize + Medu + Fedu + Mjob +
##     traveltime + studytime + failures + schoolsup + famsup +
##     paid + activities + higher + internet + romantic + famrel +
##     freetime + goout + Dalc + Walc + health + absences + absences2 +
##     absences3
##
##              Df Sum of Sq    RSS    AIC
## - Walc        1      4.57 5934.8 1130.3
## - traveltime  1      4.81 5935.1 1130.3
## - internet    1      5.20 5935.5 1130.4
## - Dalc        1      6.84 5937.1 1130.5
## - Fedu        1      6.91 5937.2 1130.5
## - paid        1      7.75 5938.0 1130.5
## - higher      1     13.04 5943.3 1130.9
## - activities  1     14.21 5944.5 1131.0
## - school      1     15.48 5945.7 1131.1
## - health      1     21.80 5952.1 1131.5
## - address     1     22.39 5952.7 1131.5
## - famrel      1     22.71 5953.0 1131.5
## - Medu        1     24.58 5954.8 1131.7
## - famsize     1     25.47 5955.7 1131.7
## - freetime    1     26.89 5957.2 1131.8
```

```
## <none>                   5930.3 1132.0
## - famsup      1     56.23 5986.5 1133.8
## - age         1     56.26 5986.5 1133.8
## - schoolsup   1     63.67 5993.9 1134.2
## - romantic    1     66.50 5996.8 1134.4
## - studytime   1     68.86 5999.1 1134.6
## - Mjob        4    166.52 6096.8 1135.0
## - sex         1     91.18 6021.5 1136.1
## - absences3   1     96.70 6027.0 1136.4
## - absences2   1    140.67 6070.9 1139.3
## - goout       1    154.58 6084.9 1140.2
## - absences    1    249.47 6179.7 1146.3
## - failures    1    445.64 6375.9 1158.7
##
## Step:  AIC=1130.33
## G3 ~ school + sex + age + address + famsize + Medu + Fedu + Mjob +
##     traveltime + studytime + failures + schoolsup + famsup +
##     paid + activities + higher + internet + romantic + famrel +
##     freetime + goout + Dalc + health + absences + absences2 +
##     absences3
##
##              Df Sum of Sq    RSS    AIC
## - Dalc        1      2.97 5937.8 1128.5
## - traveltime  1      4.55 5939.4 1128.6
## - internet    1      5.10 5939.9 1128.7
## - Fedu        1      7.50 5942.3 1128.8
## - paid        1      9.25 5944.1 1129.0
## - higher      1     12.46 5947.3 1129.2
## - activities  1     14.41 5949.2 1129.3
## - school      1     14.85 5949.7 1129.3
## - health      1     20.60 5955.4 1129.7
## - famrel      1     20.78 5955.6 1129.7
## - address     1     20.85 5955.7 1129.7
## - Medu        1     22.53 5957.4 1129.8
## - freetime    1     25.37 5960.2 1130.0
## - famsize     1     26.11 5960.9 1130.1
## <none>                    5934.8 1130.3
## - age         1     57.21 5992.0 1132.1
## - famsup      1     59.28 5994.1 1132.3
## - studytime   1     65.70 6000.5 1132.7
## - schoolsup   1     66.49 6001.3 1132.7
## - romantic    1     67.71 6002.5 1132.8
## - Mjob        4    169.50 6104.3 1133.5
## - absences3   1     95.90 6030.7 1134.7
## - sex         1     97.64 6032.5 1134.8
## - absences2   1    140.53 6075.4 1137.6
## - goout       1    156.39 6091.2 1138.6
## - absences    1    254.06 6188.9 1144.9
## - failures    1    445.87 6380.7 1157.0
##
## Step:  AIC=1128.53
## G3 ~ school + sex + age + address + famsize + Medu + Fedu + Mjob +
##     traveltime + studytime + failures + schoolsup + famsup +
##     paid + activities + higher + internet + romantic + famrel +
```

```
##     freetime + goout + health + absences + absences2 + absences3
##
##              Df Sum of Sq    RSS    AIC
## - internet    1      4.78 5942.6 1126.8
## - traveltime  1      5.20 5943.0 1126.9
## - Fedu        1      7.35 5945.2 1127.0
## - paid        1      8.00 5945.8 1127.1
## - higher      1     12.23 5950.0 1127.3
## - activities  1     13.27 5951.1 1127.4
## - school      1     14.29 5952.1 1127.5
## - Medu        1     21.87 5959.7 1128.0
## - address     1     22.01 5959.8 1128.0
## - health      1     22.08 5959.9 1128.0
## - famrel      1     23.12 5960.9 1128.1
## - freetime    1     23.68 5961.5 1128.1
## - famsize     1     24.95 5962.8 1128.2
## <none>                    5937.8 1128.5
## - famsup      1     58.92 5996.7 1130.4
## - age         1     59.13 5996.9 1130.5
## - romantic    1     67.72 6005.5 1131.0
## - schoolsup   1     68.27 6006.1 1131.0
## - studytime   1     68.59 6006.4 1131.1
## - Mjob        4    171.61 6109.4 1131.8
## - absences3   1     94.29 6032.1 1132.8
## - sex         1     94.82 6032.6 1132.8
## - absences2   1    138.42 6076.2 1135.6
## - goout       1    172.36 6110.2 1137.8
## - absences    1    251.11 6188.9 1142.9
## - failures    1    452.69 6390.5 1155.5
##
## Step:  AIC=1126.85
## G3 ~ school + sex + age + address + famsize + Medu + Fedu + Mjob +
##     traveltime + studytime + failures + schoolsup + famsup +
##     paid + activities + higher + romantic + famrel + freetime +
##     goout + health + absences + absences2 + absences3
##
##              Df Sum of Sq    RSS    AIC
## - traveltime  1      5.22 5947.8 1125.2
## - Fedu        1      7.85 5950.4 1125.4
## - paid        1      9.34 5951.9 1125.5
## - higher      1     11.46 5954.0 1125.6
## - activities  1     13.06 5955.6 1125.7
## - school      1     14.08 5956.7 1125.8
## - Medu        1     21.07 5963.7 1126.2
## - freetime    1     24.04 5966.6 1126.4
## - famrel      1     24.15 5966.7 1126.5
## - famsize     1     24.32 5966.9 1126.5
## - health      1     24.84 5967.4 1126.5
## - address     1     25.68 5968.3 1126.5
## <none>                    5942.6 1126.8
## - famsup      1     57.89 6000.5 1128.7
## - age         1     63.16 6005.7 1129.0
## - romantic    1     64.75 6007.3 1129.1
## - schoolsup   1     69.08 6011.7 1129.4
```

```
## - studytime   1     71.59 6014.2 1129.6
## - Mjob        4    173.99 6116.6 1130.2
## - absences3   1     93.77 6036.3 1131.0
## - sex         1     97.57 6040.1 1131.3
## - absences2   1    137.89 6080.5 1133.9
## - goout       1    169.83 6112.4 1136.0
## - absences    1    253.32 6195.9 1141.3
## - failures    1    453.29 6395.9 1153.9
##
## Step:  AIC=1125.2
## G3 ~ school + sex + age + address + famsize + Medu + Fedu + Mjob +
##     studytime + failures + schoolsup + famsup + paid + activities +
##     higher + romantic + famrel + freetime + goout + health +
##     absences + absences2 + absences3
##
##              Df Sum of Sq    RSS    AIC
## - Fedu        1      9.42 5957.2 1123.8
## - paid        1      9.83 5957.6 1123.8
## - school      1     11.65 5959.5 1124.0
## - higher      1     11.75 5959.5 1124.0
## - activities  1     13.41 5961.2 1124.1
## - Medu        1     20.91 5968.7 1124.6
## - famsize     1     22.75 5970.6 1124.7
## - famrel      1     24.03 5971.8 1124.8
## - health      1     24.77 5972.6 1124.8
## - freetime    1     25.51 5973.3 1124.9
## <none>                    5947.8 1125.2
## - address     1     34.54 5982.3 1125.5
## - age         1     61.03 6008.8 1127.2
## - famsup      1     62.00 6009.8 1127.3
## - romantic    1     65.85 6013.6 1127.5
## - schoolsup   1     69.35 6017.1 1127.8
## - studytime   1     75.40 6023.2 1128.2
## - Mjob        4    178.02 6125.8 1128.8
## - absences3   1     94.78 6042.6 1129.4
## - sex         1     95.43 6043.2 1129.5
## - absences2   1    139.45 6087.2 1132.3
## - goout       1    175.80 6123.6 1134.7
## - absences    1    255.99 6203.8 1139.8
## - failures    1    454.75 6402.5 1152.3
##
## Step:  AIC=1123.82
## G3 ~ school + sex + age + address + famsize + Medu + Mjob + studytime +
##     failures + schoolsup + famsup + paid + activities + higher +
##     romantic + famrel + freetime + goout + health + absences +
##     absences2 + absences3
##
##              Df Sum of Sq    RSS    AIC
## - paid        1      8.85 5966.1 1122.4
## - school      1     12.21 5969.4 1122.6
## - activities  1     12.23 5969.5 1122.6
## - higher      1     13.38 5970.6 1122.7
## - famsize     1     21.91 5979.1 1123.3
## - health      1     22.94 5980.2 1123.3
```

```
## - famrel      1     23.67 5980.9 1123.4
## - freetime    1     24.01 5981.2 1123.4
## <none>                    5957.2 1123.8
## - address     1     34.17 5991.4 1124.1
## - Medu        1     52.49 6009.7 1125.3
## - famsup      1     57.86 6015.1 1125.6
## - age         1     61.52 6018.7 1125.9
## - romantic    1     64.97 6022.2 1126.1
## - schoolsup   1     66.58 6023.8 1126.2
## - studytime   1     70.28 6027.5 1126.5
## - Mjob        4    175.66 6132.9 1127.3
## - absences3   1     90.57 6047.8 1127.8
## - sex         1     95.37 6052.6 1128.1
## - absences2   1    134.00 6091.2 1130.6
## - goout       1    171.77 6129.0 1133.0
## - absences    1    248.43 6205.7 1138.0
## - failures    1    480.08 6437.3 1152.4
##
## Step:  AIC=1122.41
## G3 ~ school + sex + age + address + famsize + Medu + Mjob + studytime +
##     failures + schoolsup + famsup + activities + higher + romantic +
##     famrel + freetime + goout + health + absences + absences2 +
##     absences3
##
##              Df Sum of Sq    RSS    AIC
## - school      1     13.00 5979.1 1121.3
## - activities  1     13.31 5979.4 1121.3
## - higher      1     15.68 5981.8 1121.5
## - famsize     1     21.88 5988.0 1121.8
## - freetime    1     22.56 5988.6 1121.9
## - famrel      1     24.24 5990.3 1122.0
## - health      1     25.07 5991.1 1122.1
## <none>                    5966.1 1122.4
## - address     1     35.10 6001.2 1122.7
## - famsup      1     50.17 6016.2 1123.7
## - Medu        1     51.54 6017.6 1123.8
## - age         1     60.71 6026.8 1124.4
## - romantic    1     64.15 6030.2 1124.6
## - schoolsup   1     68.54 6034.6 1124.9
## - studytime   1     74.87 6040.9 1125.3
## - Mjob        4    175.51 6141.6 1125.9
## - absences3   1     90.59 6056.7 1126.4
## - sex         1     93.01 6059.1 1126.5
## - absences2   1    134.36 6100.4 1129.2
## - goout       1    168.43 6134.5 1131.4
## - absences    1    249.93 6216.0 1136.6
## - failures    1    502.57 6468.6 1152.4
##
## Step:  AIC=1121.27
## G3 ~ sex + age + address + famsize + Medu + Mjob + studytime +
##     failures + schoolsup + famsup + activities + higher + romantic +
##     famrel + freetime + goout + health + absences + absences2 +
##     absences3
##
```

```
##              Df Sum of Sq    RSS    AIC
## - activities  1     16.23 5995.3 1120.3
## - higher      1     18.70 5997.8 1120.5
## - famrel      1     21.53 6000.6 1120.7
## - famsize     1     24.22 6003.3 1120.9
## - freetime    1     25.10 6004.2 1120.9
## - health      1     26.31 6005.4 1121.0
## - address     1     26.59 6005.7 1121.0
## <none>                    5979.1 1121.3
## - age         1     48.58 6027.7 1122.5
## - Medu        1     50.78 6029.9 1122.6
## - famsup      1     56.17 6035.2 1123.0
## - romantic    1     61.85 6040.9 1123.3
## - studytime   1     68.60 6047.7 1123.8
## - schoolsup   1     70.20 6049.3 1123.9
## - Mjob        4    173.58 6152.7 1124.6
## - sex         1     89.55 6068.6 1125.1
## - absences3   1     90.39 6069.5 1125.2
## - absences2   1    132.88 6112.0 1128.0
## - goout       1    171.24 6150.3 1130.4
## - absences    1    241.48 6220.6 1134.9
## - failures    1    510.59 6489.7 1151.6
##
## Step:  AIC=1120.34
## G3 ~ sex + age + address + famsize + Medu + Mjob + studytime +
##     failures + schoolsup + famsup + higher + romantic + famrel +
##     freetime + goout + health + absences + absences2 + absences3
##
##             Df Sum of Sq    RSS    AIC
## - higher     1     16.40 6011.7 1119.4
## - famrel     1     20.59 6015.9 1119.7
## - freetime   1     22.63 6017.9 1119.8
## - famsize    1     24.67 6020.0 1120.0
## - health     1     26.29 6021.6 1120.1
## - address    1     30.10 6025.4 1120.3
## <none>                   5995.3 1120.3
## - age        1     44.33 6039.6 1121.2
## - Medu       1     50.07 6045.4 1121.6
## - famsup     1     53.59 6048.9 1121.8
## - studytime  1     62.10 6057.4 1122.4
## - romantic   1     66.20 6061.5 1122.7
## - schoolsup  1     74.11 6069.4 1123.2
## - Mjob       4    172.43 6167.7 1123.5
## - sex        1     82.58 6077.9 1123.7
## - absences3  1     88.84 6084.1 1124.2
## - absences2  1    130.94 6126.2 1126.9
## - goout      1    176.11 6171.4 1129.8
## - absences   1    238.58 6233.9 1133.8
## - failures   1    507.16 6502.5 1150.4
##
## Step:  AIC=1119.42
## G3 ~ sex + age + address + famsize + Medu + Mjob + studytime +
##     failures + schoolsup + famsup + romantic + famrel + freetime +
##     goout + health + absences + absences2 + absences3
```

```
##
##             Df Sum of Sq    RSS    AIC
## - freetime   1     21.51 6033.2 1118.8
## - famrel     1     22.13 6033.8 1118.9
## - health     1     25.09 6036.8 1119.1
## - famsize    1     25.35 6037.1 1119.1
## - address    1     28.13 6039.8 1119.3
## <none>                   6011.7 1119.4
## - Medu       1     52.47 6064.2 1120.8
## - famsup     1     53.21 6064.9 1120.9
## - age        1     53.22 6064.9 1120.9
## - studytime  1     69.46 6081.2 1122.0
## - romantic   1     72.85 6084.6 1122.2
## - sex        1     73.42 6085.1 1122.2
## - schoolsup  1     74.64 6086.3 1122.3
## - Mjob       4    174.62 6186.3 1122.7
## - absences3  1     81.59 6093.3 1122.7
## - absences2  1    126.29 6138.0 1125.6
## - goout      1    177.09 6188.8 1128.9
## - absences   1    241.61 6253.3 1133.0
## - failures   1    583.01 6594.7 1154.0
##
## Step:  AIC=1118.83
## G3 ~ sex + age + address + famsize + Medu + Mjob + studytime +
##     failures + schoolsup + famsup + romantic + famrel + goout +
##     health + absences + absences2 + absences3
##
##             Df Sum of Sq    RSS    AIC
## - health     1     23.94 6057.2 1118.4
## - famsize    1     24.79 6058.0 1118.5
## - address    1     28.80 6062.0 1118.7
## - famrel     1     29.28 6062.5 1118.7
## <none>                   6033.2 1118.8
## - famsup     1     49.75 6083.0 1120.1
## - Medu       1     49.75 6083.0 1120.1
## - age        1     54.60 6087.8 1120.4
## - studytime  1     64.11 6097.3 1121.0
## - romantic   1     70.37 6103.6 1121.4
## - schoolsup  1     75.44 6108.6 1121.7
## - Mjob       4    169.10 6202.3 1121.8
## - absences3  1     81.35 6114.6 1122.1
## - sex        1     89.26 6122.5 1122.6
## - absences2  1    127.53 6160.7 1125.1
## - goout      1    157.22 6190.4 1127.0
## - absences   1    243.25 6276.5 1132.4
## - failures   1    572.95 6606.2 1152.7
##
## Step:  AIC=1118.39
## G3 ~ sex + age + address + famsize + Medu + Mjob + studytime +
##     failures + schoolsup + famsup + romantic + famrel + goout +
##     absences + absences2 + absences3
##
##             Df Sum of Sq    RSS    AIC
## - famrel     1     24.61 6081.8 1118.0
```

```
## - famsize    1     27.20 6084.4 1118.2
## <none>                   6057.2 1118.4
## - address    1     31.63 6088.8 1118.5
## - age        1     48.94 6106.1 1119.6
## - famsup     1     53.63 6110.8 1119.9
## - Medu       1     59.37 6116.5 1120.2
## - Mjob       4    158.98 6216.1 1120.6
## - studytime  1     66.79 6123.9 1120.7
## - schoolsup  1     72.84 6130.0 1121.1
## - romantic   1     76.64 6133.8 1121.4
## - sex        1     79.66 6136.8 1121.5
## - absences3  1     80.13 6137.3 1121.6
## - absences2  1    127.50 6184.7 1124.6
## - goout      1    156.32 6213.5 1126.5
## - absences   1    246.28 6303.4 1132.1
## - failures   1    585.36 6642.5 1152.8
##
## Step:  AIC=1118
## G3 ~ sex + age + address + famsize + Medu + Mjob + studytime +
##     failures + schoolsup + famsup + romantic + goout + absences +
##     absences2 + absences3
##
##             Df Sum of Sq    RSS    AIC
## - famsize    1     26.18 6108.0 1117.7
## <none>                   6081.8 1118.0
## - address    1     32.58 6114.3 1118.1
## - age        1     43.56 6125.3 1118.8
## - famsup     1     53.85 6135.6 1119.5
## - Medu       1     60.45 6142.2 1119.9
## - Mjob       4    161.53 6243.3 1120.3
## - studytime  1     71.12 6152.9 1120.6
## - schoolsup  1     71.16 6152.9 1120.6
## - absences3  1     74.71 6156.5 1120.8
## - romantic   1     82.45 6164.2 1121.3
## - sex        1     86.74 6168.5 1121.6
## - absences2  1    118.85 6200.6 1123.6
## - goout      1    147.76 6229.5 1125.5
## - absences   1    231.44 6313.2 1130.8
## - failures   1    602.52 6684.3 1153.3
##
## Step:  AIC=1117.69
## G3 ~ sex + age + address + Medu + Mjob + studytime + failures +
##     schoolsup + famsup + romantic + goout + absences + absences2 +
##     absences3
##
##             Df Sum of Sq    RSS    AIC
## <none>                   6108.0 1117.7
## - address    1     37.25 6145.2 1118.1
## - age        1     42.23 6150.2 1118.4
## - Medu       1     51.79 6159.7 1119.0
## - famsup     1     61.89 6169.8 1119.7
## - studytime  1     68.47 6176.4 1120.1
## - Mjob       4    163.53 6271.5 1120.1
## - schoolsup  1     70.36 6178.3 1120.2
```

```
## - absences3  1     76.77 6184.7 1120.6
## - romantic   1     77.51 6185.5 1120.7
## - sex        1     92.61 6200.6 1121.6
## - absences2  1    122.79 6230.7 1123.5
## - goout      1    146.49 6254.4 1125.0
## - absences   1    241.71 6349.7 1131.0
## - failures   1    619.34 6727.3 1153.8
```

For example, in the last step, farmsize was dropped, because It leads to model with AIC=1117.69. After that dropping variables will not lead to smaller AIC and we should stop.

Final Model:

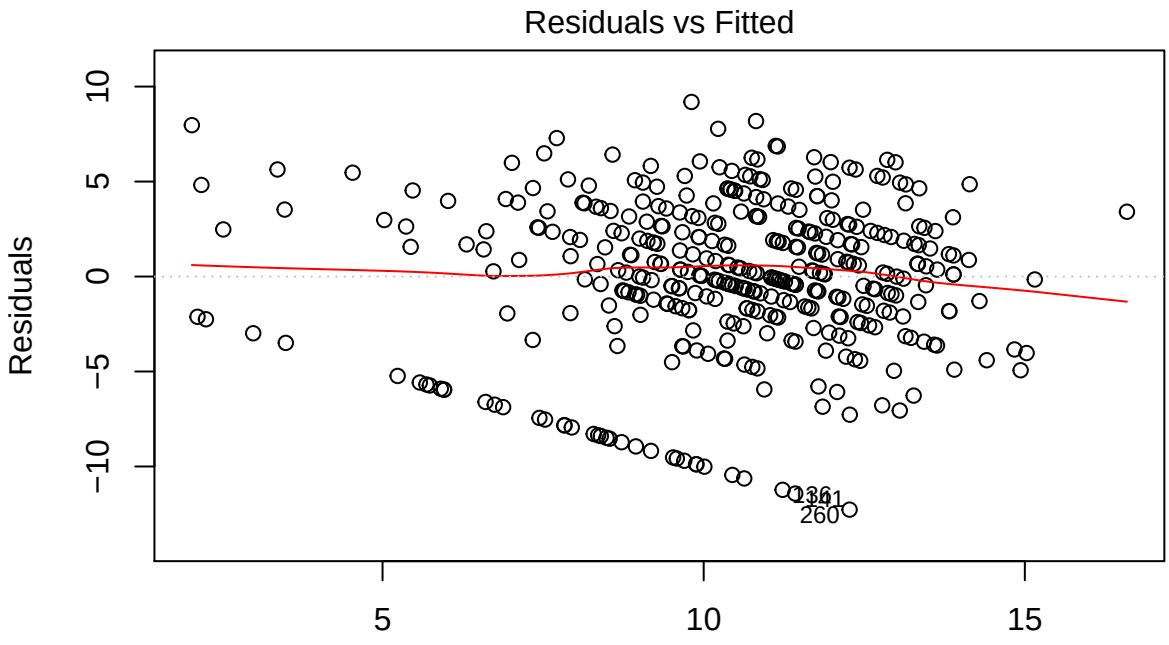
```
selected_model <- lm(formula = G3 ~ sex + age + address + Medu + Mjob + studytime +
    failures + schoolsup + famsup + romantic + goout + absences +
    absences2 + absences3, data = data)
summary(selected_model)
```

```
##
## Call:
## lm(formula = G3 ~ sex + age + address + Medu + Mjob + studytime +
##     failures + schoolsup + famsup + romantic + goout + absences +
##     absences2 + absences3, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.2713  -1.9113   0.1742   2.6427   9.1900
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.397e+01  3.233e+00   4.322 1.98e-05 ***
## sexM          1.074e+00  4.490e-01   2.391 0.017298 *
## age          -2.908e-01  1.801e-01  -1.615 0.107247
## addressU      7.690e-01  5.072e-01   1.516 0.130287
## Medu          4.546e-01  2.542e-01   1.788 0.074580 .
## Mjobhealth    1.468e+00  1.004e+00   1.463 0.144398
## Mjobother    -8.506e-02  6.569e-01  -0.129 0.897039
## Mjobservices  9.594e-01  7.276e-01   1.319 0.188111
## Mjobteacher  -6.641e-01  9.453e-01  -0.703 0.482742
## studytime     5.381e-01  2.617e-01   2.056 0.040496 *
## failures     -1.849e+00  2.990e-01  -6.183 1.64e-09 ***
## schoolsupyes -1.340e+00  6.430e-01  -2.084 0.037840 *
## famsupyes    -8.573e-01  4.386e-01  -1.955 0.051374 .
## romanticyes  -9.788e-01  4.475e-01  -2.187 0.029342 *
## goout        -5.690e-01  1.892e-01  -3.007 0.002815 **
## absences      3.319e-01  8.594e-02   3.863 0.000132 ***
## absences2    -1.305e-02  4.739e-03  -2.753 0.006191 **
## absences3     1.187e-04  5.452e-05   2.177 0.030118 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.025 on 377 degrees of freedom
## Multiple R-squared:  0.2614, Adjusted R-squared:  0.2281
## F-statistic:  7.85 on 17 and 377 DF,  p-value: < 2.2e-16
```

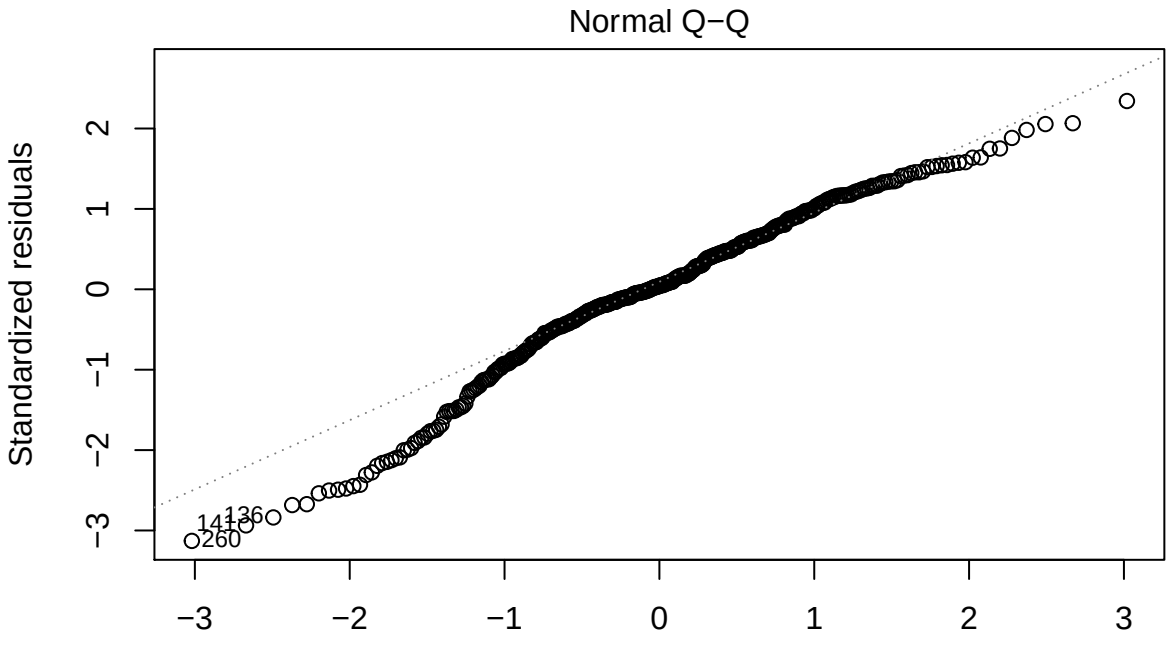
Task 7

Let's compute Cook distance to check for outliers.
As we can see from plots, there are outlier - 277 row in data.
From examing data, we can suppose that 277 row contains extremely large absences.

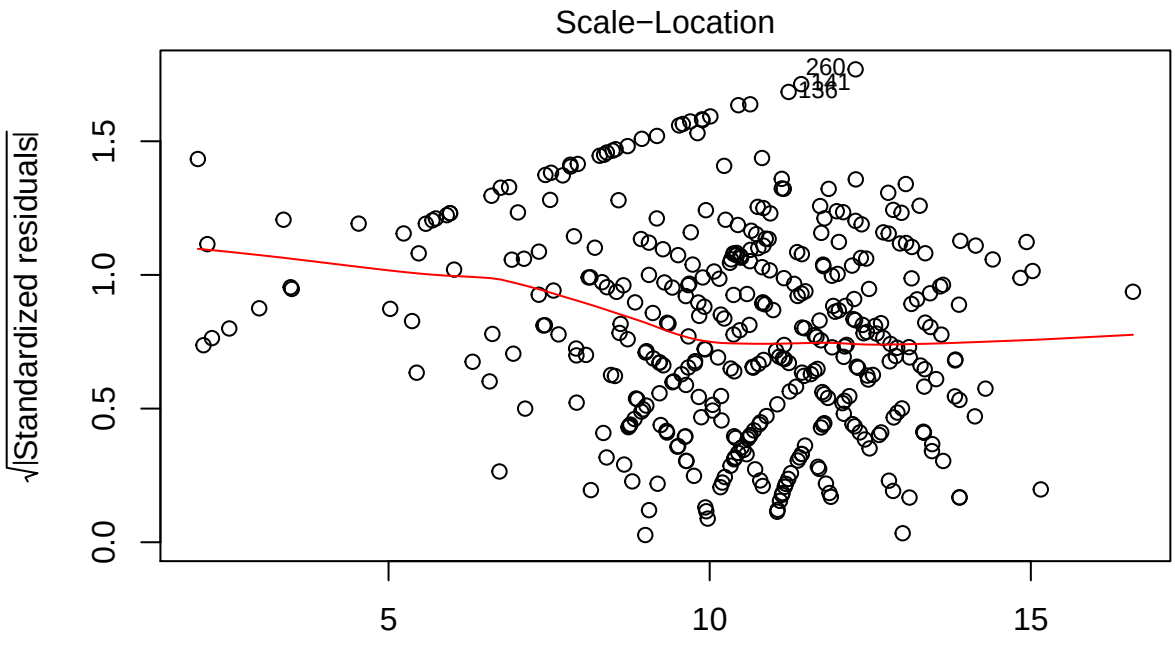
```
plot(selected_model)
```

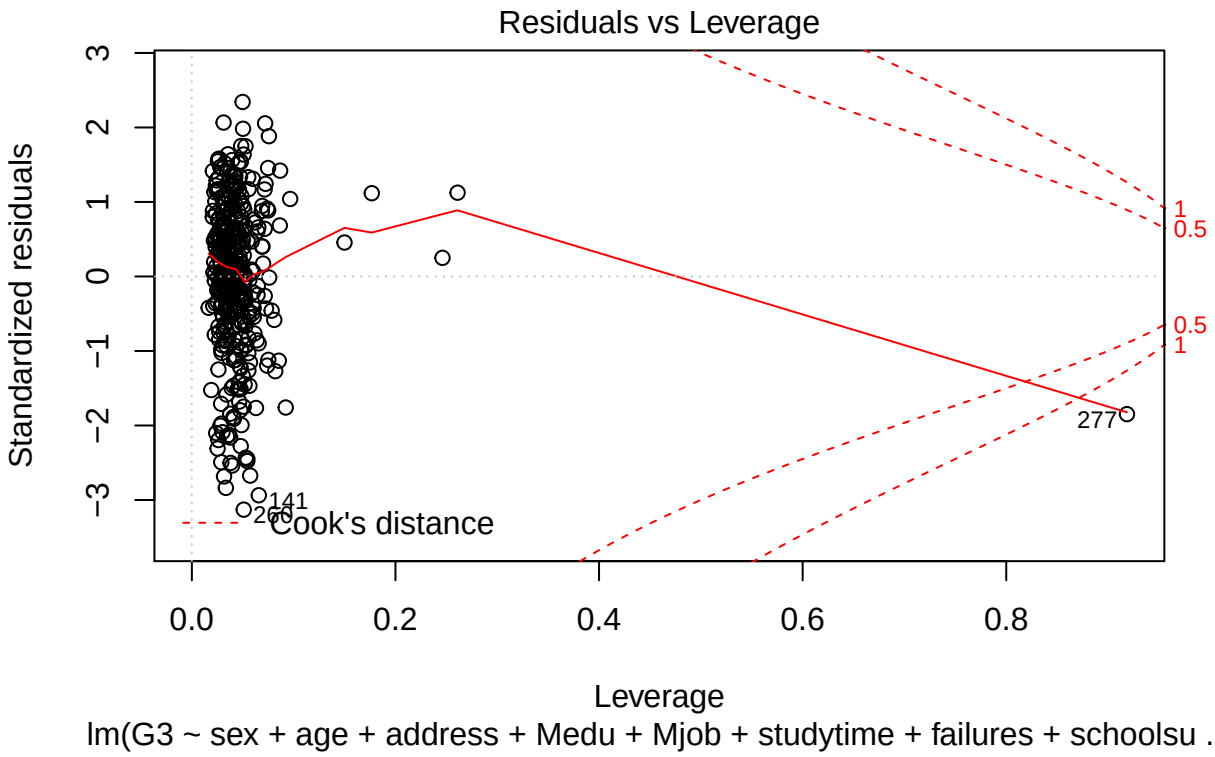


Fitted values
lm(G3 ~ sex + age + address + Medu + Mjob + studytime + failures + schoolsu ...

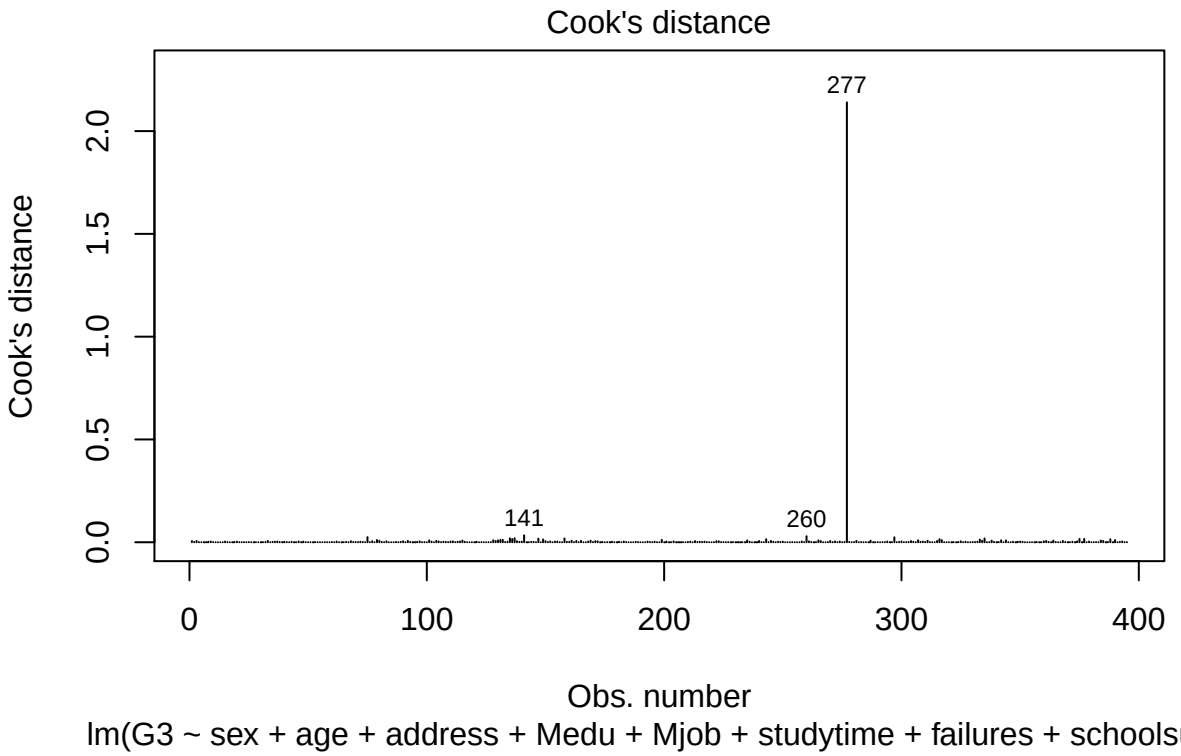


Theoretical Quantiles
lm(G3 ~ sex + age + address + Medu + Mjob + studytime + failures + schoolsu ...


Scale−Location
260
141
136
√|Standardized residuals|
0.0
0.5
1.0
1.5
5
10
15
Fitted values
lm(G3 ~ sex + age + address + Medu + Mjob + studytime + failures + schoolsu ...



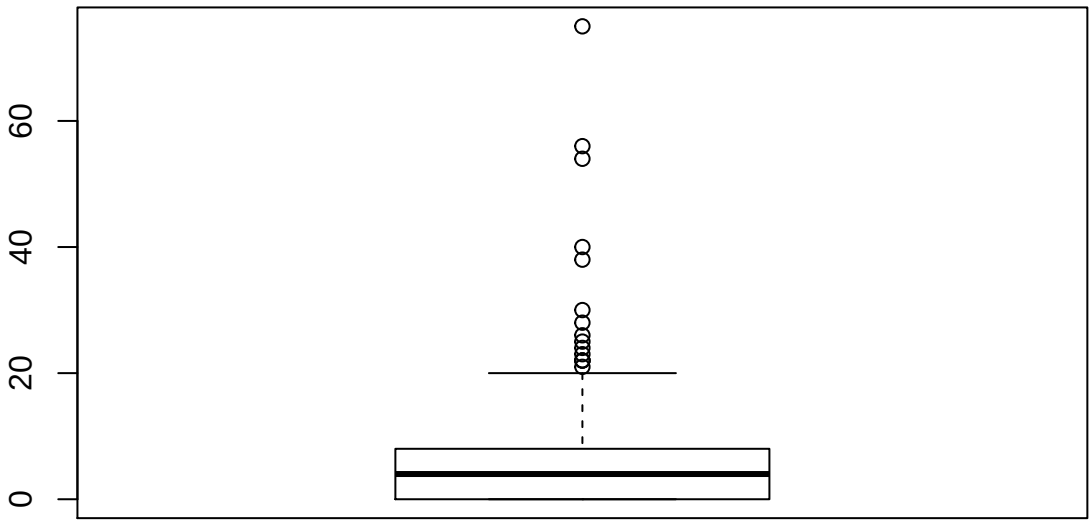

```
plot(selected_model, which=c(4))
```



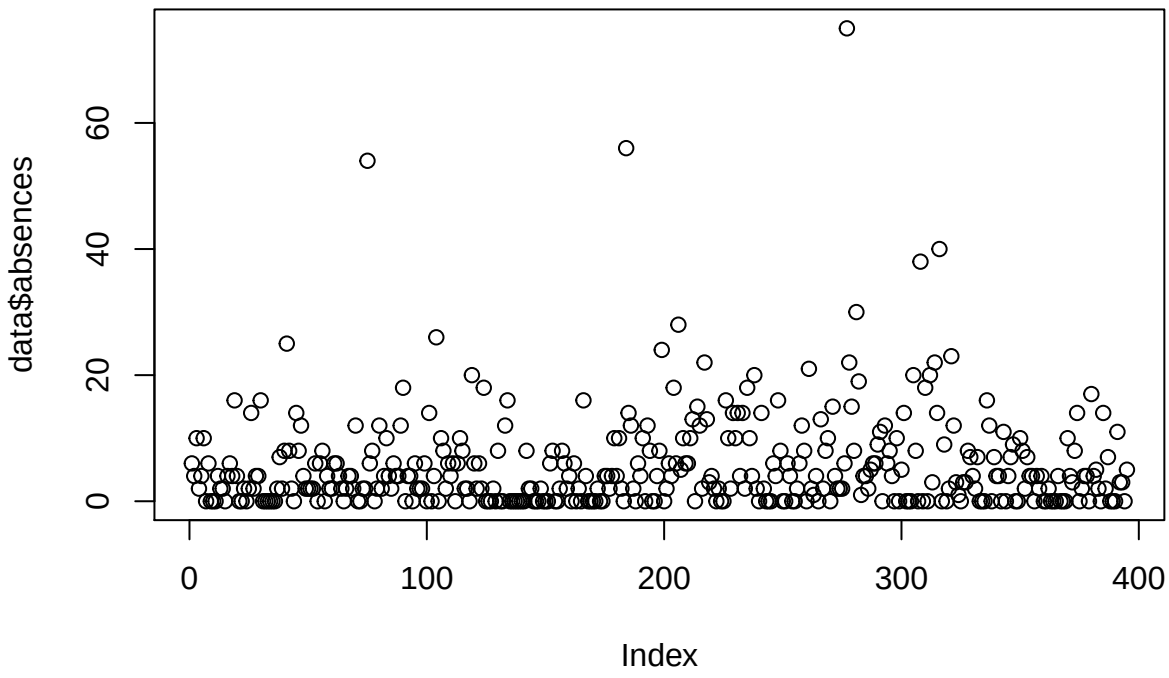

```
# cooks.distance(selected_model)
```

This can clearly be seen from boxplot and regular plot of absences.

```
boxplot(data$absences)
```




```
plot(data$absences)
```



Task 8

Randomly selecting 5 rows and delete age in them.
We can remove rows with missing data or try to simulate Next implement 3 methods of missing data imputation:
- generage age from values of age from other rows of data;
- replace missing age values with mean agea value;
- regress age on other variables.
Here you can see results:

```
set.seed(20)
data_for_testing <- data

random_5_rows <- data_for_testing[sample(nrow(data_for_testing), 5), ]

cat("original age", random_5_rows$age, "\n")
```

```
## original age 18 17 16 16 20
```

```
data_for_testing$age[as.numeric(rownames(random_5_rows))] = NA

# 1 method
generated_ages <- sample(data$age, 5)
cat("Generated from typical age:", generated_ages, "\n")
```

```
## Generated from typical age: 19 15 15 18 17
```

```r
# 2 method
mean_age <- round(mean(data$age))
cat("mean age", mean_age, "\n")
```

```
## mean age 17
```

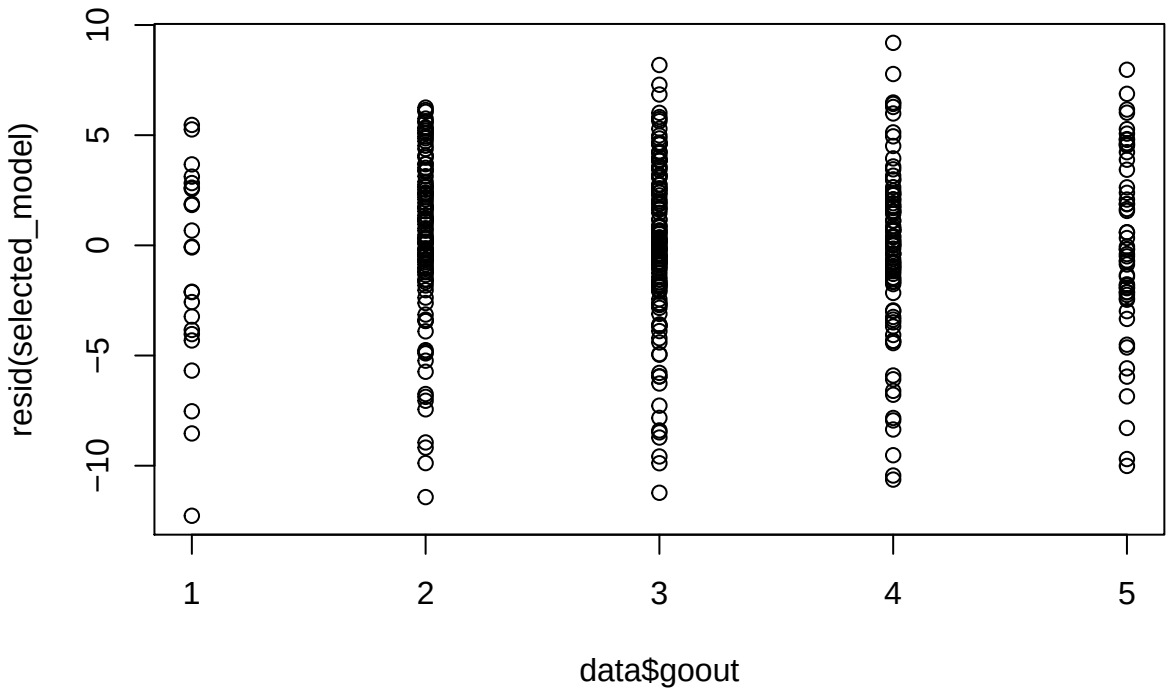
```r
data_without_age <- data_for_testing[complete.cases(data_for_testing), ]

# 3 method
age_model <- lm(age~.-G3, data=data_without_age)

random_5_rows$age <- NULL

a <- predict.lm(age_model, newdata = random_5_rows)
cat("regressed age: ", round(a))
```

```
## regressed age:  17 16 16 16 19
```

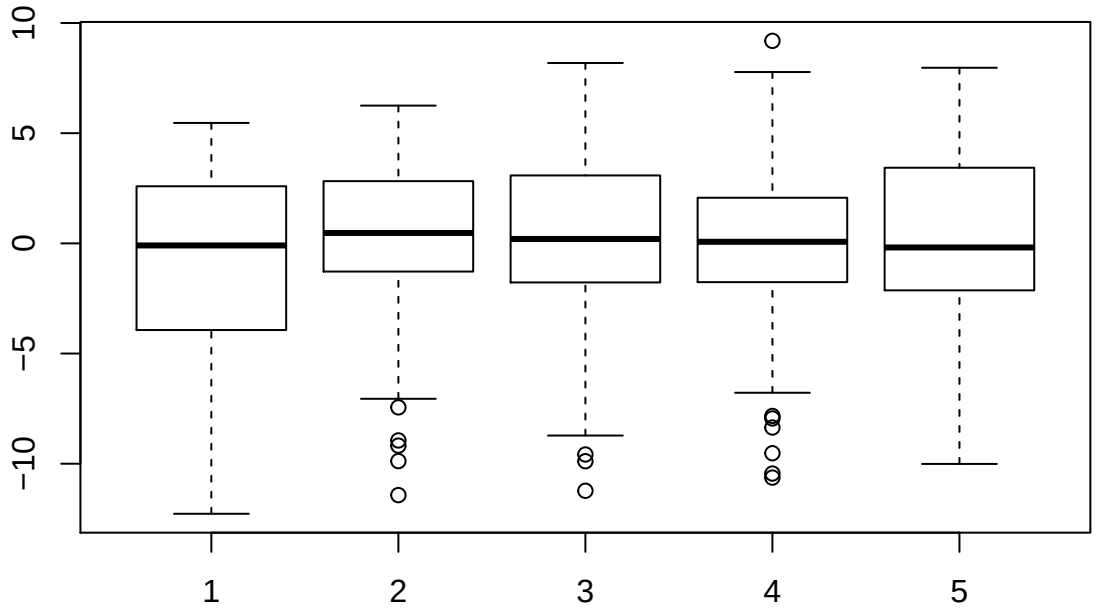
Task 9

Here we plot goout vs residuals to see if variance of the residuals is rather different for different values of goout.
Next we perform bartlett test that shows that variances are homogemeous (because p-value > 0.05), so our assumption about heteroscedasticity was wrong.

```r
plot(data$goout, resid(selected_model))
```



```r
boxplot(resid(selected_model) ~ data$goout)
```




```r
bartlett.test(resid(selected_model) ~ data$goout)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  resid(selected_model) by data$goout
## Bartlett's K-squared = 2.0929, df = 4, p-value = 0.7187
```

Task 10

Next we compute the White estimator of covariance matrix of the OLS estimators.
Here we can see that std. errors for coefficients computed with White estimator are different, because White estimator takes into account heteroskedasticity.
So there we have heteroskedasticity-consistent standard errors.
If we have had heteroskedasticity in our model, we would have seen changes in variables significance.

```r
library("sandwich")
library("lmtest")
new.model <-lm(formula = G3 ~ sex + age + address + Medu + Mjob + studytime +
    failures + schoolsup + famsup + romantic + goout + absences +
    absences2 + absences3, data = data)
covWhite = vcovHC(new.model, type="HC")
coeftest(new.model, vcov=covWhite)
```

```
##
## t test of coefficients:
##
##                 Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)   1.3973e+01  3.3673e+00  4.1497 4.118e-05 ***
## sexM          1.0735e+00  4.2043e-01  2.5534 0.0110603 *
## age          -2.9084e-01  1.9269e-01 -1.5094 0.1320408
## addressU      7.6905e-01  5.0372e-01  1.5267 0.1276623
## Medu          4.5459e-01  2.4603e-01  1.8477 0.0654330 .
## Mjobhealth    1.4678e+00  1.0145e+00  1.4469 0.1487679
## Mjobother    -8.5058e-02  7.1313e-01 -0.1193 0.9051220
## Mjobservices  9.5937e-01  7.9561e-01  1.2058 0.2286413
## Mjobteacher  -6.6414e-01  9.7845e-01 -0.6788 0.4977001
## studytime     5.3806e-01  2.7676e-01  1.9441 0.0526245 .
## failures     -1.8486e+00  3.1275e-01 -5.9110 7.619e-09 ***
## schoolsupyes -1.3399e+00  5.6523e-01 -2.3706 0.0182617 *
## famsupyes    -8.5727e-01  4.3118e-01 -1.9882 0.0475136 *
## romanticyes  -9.7882e-01  4.6158e-01 -2.1206 0.0346107 *
## goout        -5.6896e-01  1.9675e-01 -2.8918 0.0040528 **
## absences      3.3195e-01  8.7099e-02  3.8112 0.0001615 ***
## absences2    -1.3047e-02  4.2172e-03 -3.0938 0.0021237 **
## absences3     1.1868e-04  4.5519e-05  2.6073 0.0094875 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(selected_model)
```

```
##
## Call:
## lm(formula = G3 ~ sex + age + address + Medu + Mjob + studytime +
##     failures + schoolsup + famsup + romantic + goout + absences +
##     absences2 + absences3, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.2713  -1.9113   0.1742   2.6427   9.1900
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.397e+01  3.233e+00   4.322 1.98e-05 ***
## sexM          1.074e+00  4.490e-01   2.391 0.017298 *
## age          -2.908e-01  1.801e-01  -1.615 0.107247
## addressU      7.690e-01  5.072e-01   1.516 0.130287
## Medu          4.546e-01  2.542e-01   1.788 0.074580 .
## Mjobhealth    1.468e+00  1.004e+00   1.463 0.144398
## Mjobother    -8.506e-02  6.569e-01  -0.129 0.897039
## Mjobservices  9.594e-01  7.276e-01   1.319 0.188111
## Mjobteacher  -6.641e-01  9.453e-01  -0.703 0.482742
## studytime     5.381e-01  2.617e-01   2.056 0.040496 *
## failures     -1.849e+00  2.990e-01  -6.183 1.64e-09 ***
## schoolsupyes -1.340e+00  6.430e-01  -2.084 0.037840 *
## famsupyes    -8.573e-01  4.386e-01  -1.955 0.051374 .
## romanticyes  -9.788e-01  4.475e-01  -2.187 0.029342 *
## goout        -5.690e-01  1.892e-01  -3.007 0.002815 **
## absences      3.319e-01  8.594e-02   3.863 0.000132 ***
```

```
## absences2    -1.305e-02  4.739e-03  -2.753 0.006191 **
## absences3     1.187e-04  5.452e-05   2.177 0.030118 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.025 on 377 degrees of freedom
## Multiple R-squared:  0.2614, Adjusted R-squared:  0.2281
## F-statistic:  7.85 on 17 and 377 DF,  p-value: < 2.2e-16
```

Task 11

We created model that explains G3 based on different variables.
We can use It to predict G3.
The most significant variables are failures, absences, goout, studytime, romanticyes. To sum up, If you want excellent G3 you should have less failures, less go out, more studityme, do not have romantic relationships. Variable absences is really interesting, because It looks like the more you have absences the better G3 will be. It can be due to the fact that realy smart people tend to miss the lessons, to have time to learn more.

```
summary(selected_model)
```

```
##
## Call:
## lm(formula = G3 ~ sex + age + address + Medu + Mjob + studytime +
##     failures + schoolsup + famsup + romantic + goout + absences +
##     absences2 + absences3, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.2713  -1.9113   0.1742   2.6427   9.1900
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.397e+01  3.233e+00   4.322 1.98e-05 ***
## sexM          1.074e+00  4.490e-01   2.391 0.017298 *
## age          -2.908e-01  1.801e-01  -1.615 0.107247
## addressU      7.690e-01  5.072e-01   1.516 0.130287
## Medu          4.546e-01  2.542e-01   1.788 0.074580 .
## Mjobhealth    1.468e+00  1.004e+00   1.463 0.144398
## Mjobother    -8.506e-02  6.569e-01  -0.129 0.897039
## Mjobservices  9.594e-01  7.276e-01   1.319 0.188111
## Mjobteacher  -6.641e-01  9.453e-01  -0.703 0.482742
## studytime     5.381e-01  2.617e-01   2.056 0.040496 *
## failures     -1.849e+00  2.990e-01  -6.183 1.64e-09 ***
## schoolsupyes -1.340e+00  6.430e-01  -2.084 0.037840 *
## famsupyes    -8.573e-01  4.386e-01  -1.955 0.051374 .
## romanticyes  -9.788e-01  4.475e-01  -2.187 0.029342 *
## goout        -5.690e-01  1.892e-01  -3.007 0.002815 **
## absences      3.319e-01  8.594e-02   3.863 0.000132 ***
## absences2    -1.305e-02  4.739e-03  -2.753 0.006191 **
## absences3     1.187e-04  5.452e-05   2.177 0.030118 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.025 on 377 degrees of freedom
```

```
## Multiple R-squared:  0.2614, Adjusted R-squared:  0.2281
## F-statistic:  7.85 on 17 and 377 DF,  p-value: < 2.2e-16
```

Problem 4

```r
library("MASS")
d1 <- sqrt(0.4)
d2 <- sqrt(0.8)
p <- 0.2
T <- 100

sigma <- matrix(c(d1*d1, d1*d2*p, d1*d2*p, d2*d2), nrow = 2, ncol = 2)
xs <- mvrnorm(T, c(0,0), Sigma = sigma, empirical = TRUE)

x1 <- xs[,1]
x2 <- xs[,2]

# cor(x1, x2, method = "pearson")

u <- rnorm(T, 0, 1)

acf(u)
```