Data analysis for the journal

Prepare and merge data

SSP grades

Removing the one person who is in progress

Check that all observation coinside and insert the values

Renaming questions

Tidying up

Replacing missing values with mean of column

Only 5 values

SA, GPRO final and MT grades

Load SA scores

Load MT and final scores

Load Peer Grade scores

Load Khan scores

Merge all data sources created above

SA and Midterm

Midterm and Khan and SSP

Making additional dataset with dropout status

Data cleaning

Removes all PDP students and all incomplete entries, resulting in only 77 students.

Deleting unnecesarry colums

Normalizes all scores.

This is used for the ROC and cost analysis.

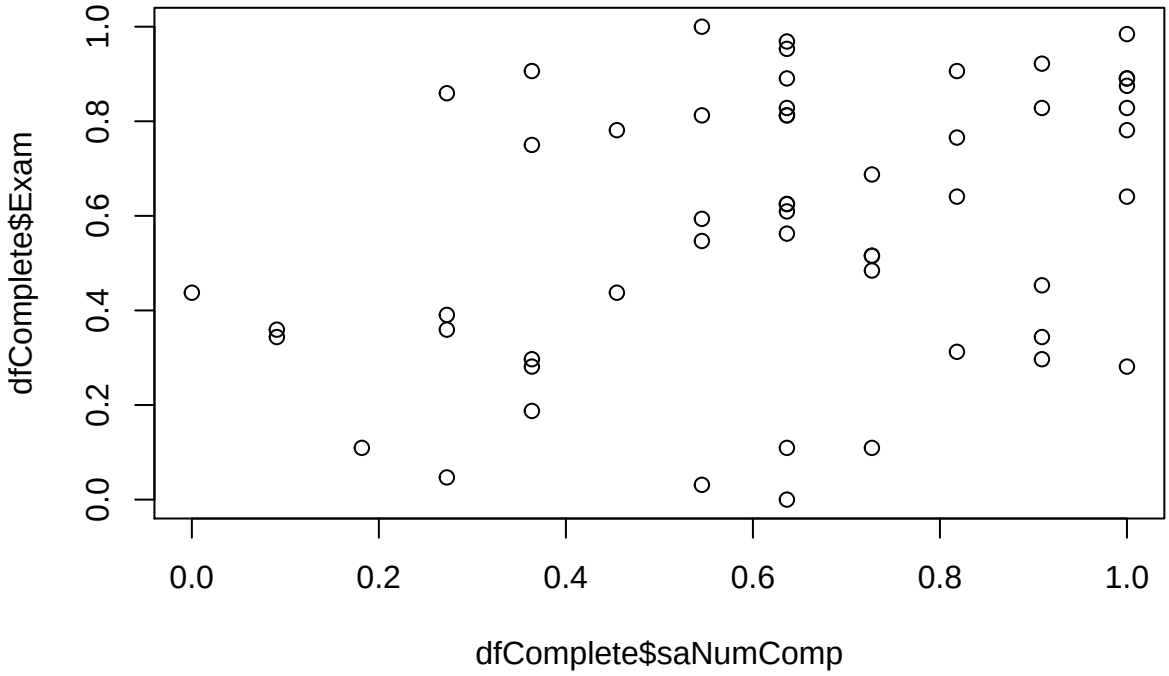
Training and test samples

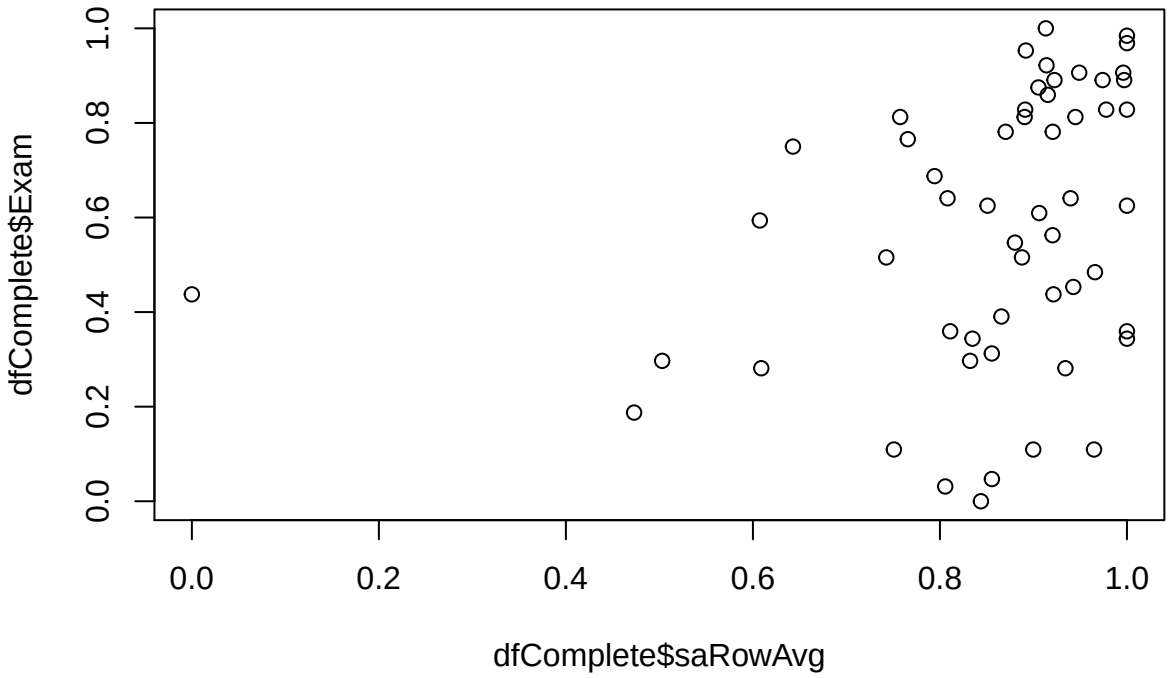
This is used for the ROC and cost analysis. 75% of the sample size of 72 students.

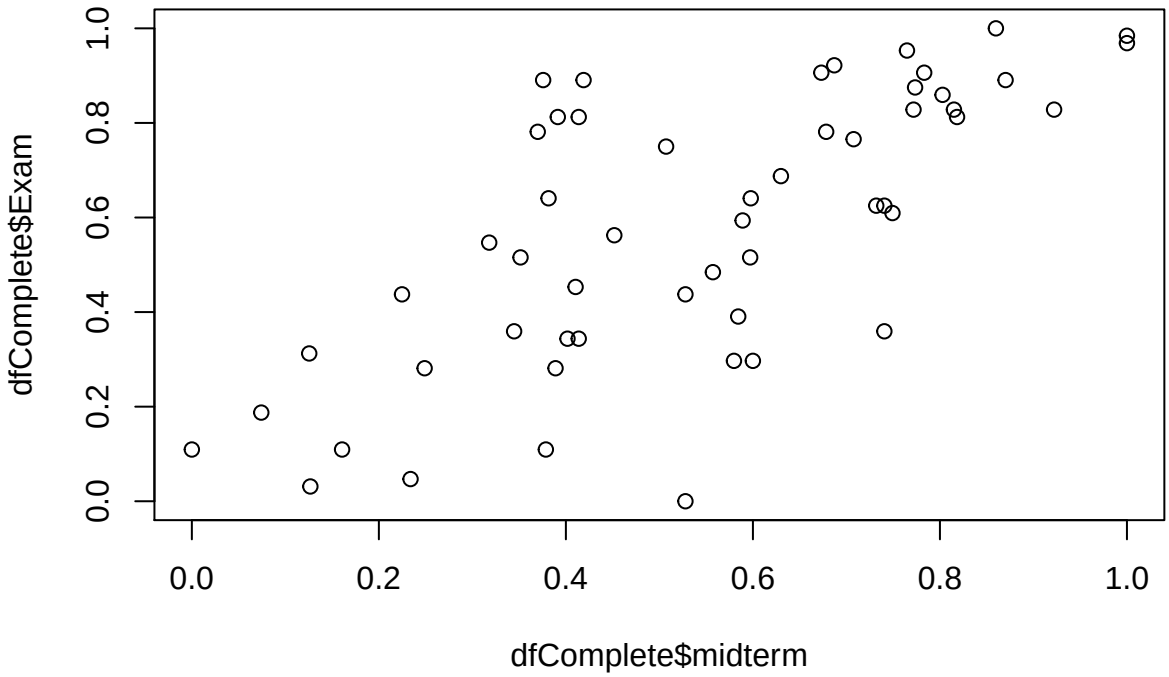
Set the seed to make your partition reproducible

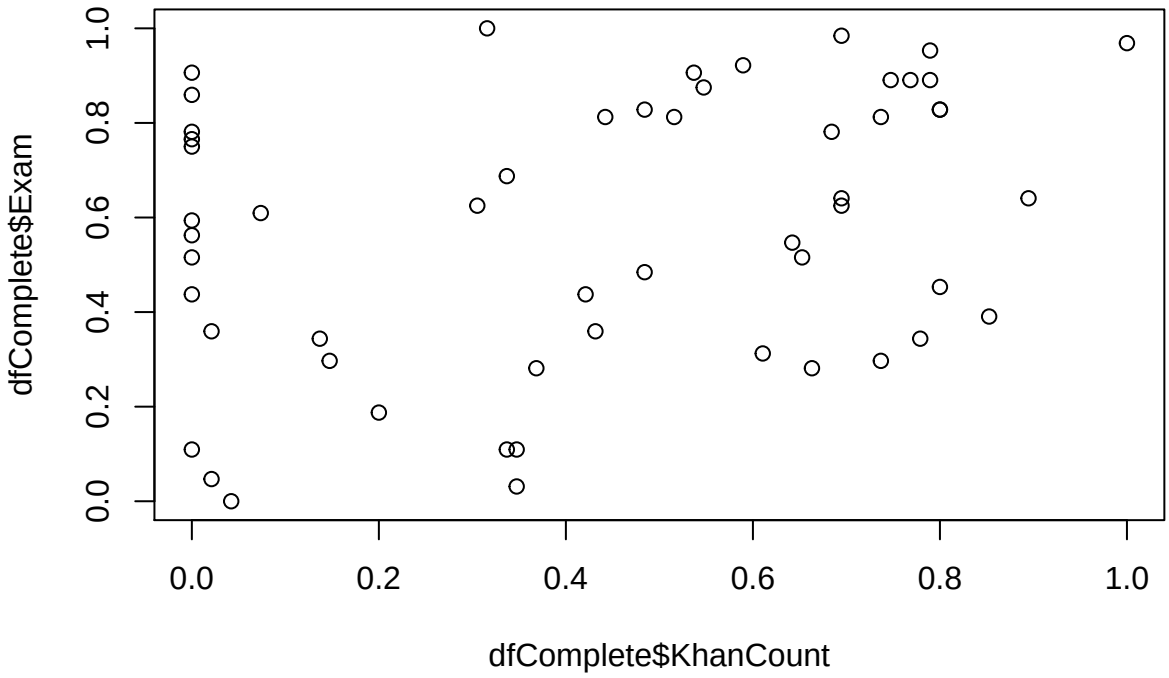
Data Analysis

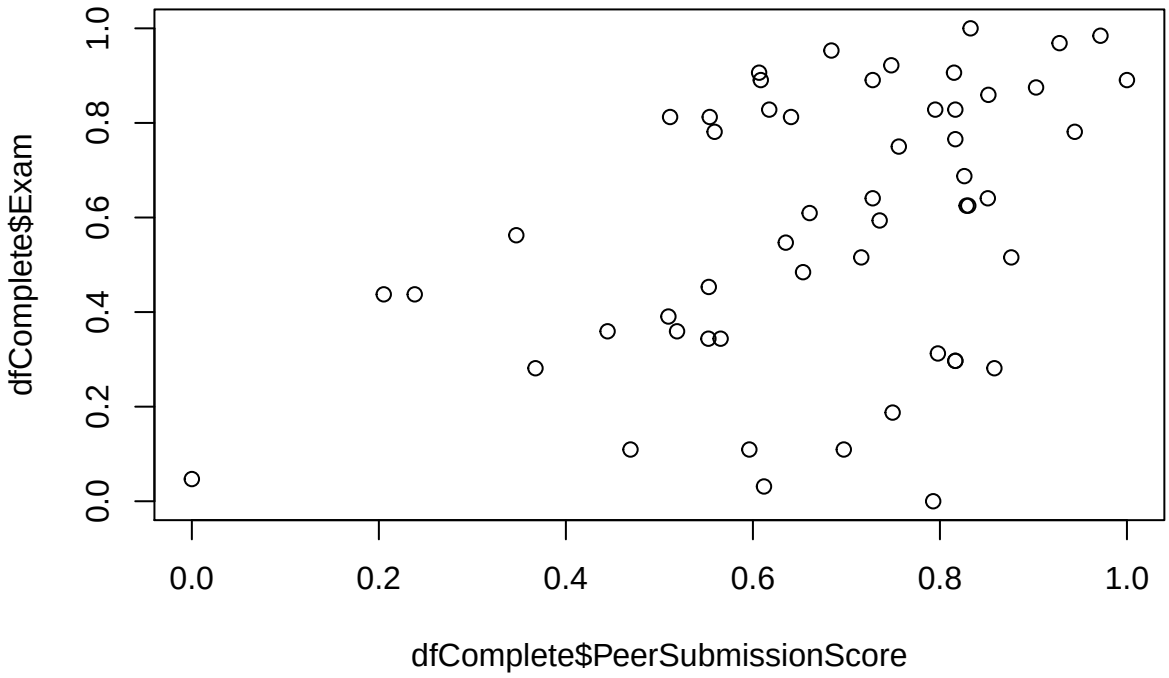
```
## [1] 0.4074074
```

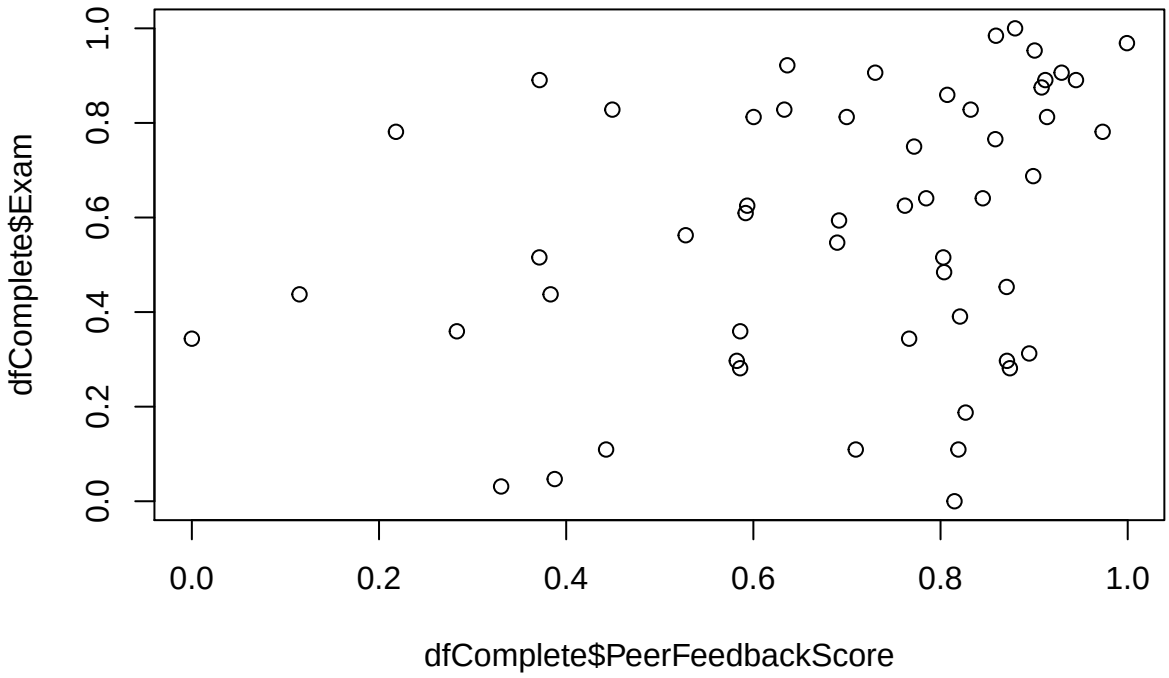


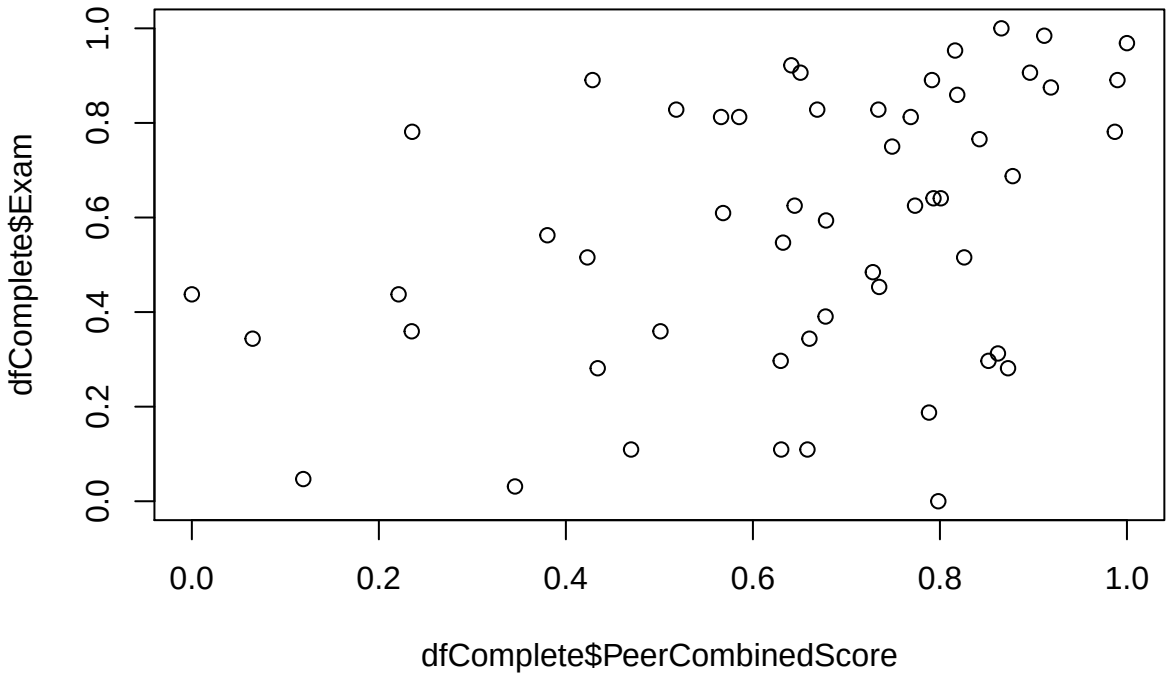

dfComplete$Exam
0.0
0.2
0.4
0.6
0.8
1.0
dfComplete$saRowAvg
0.0
0.2
0.4
0.6
0.8
1.0

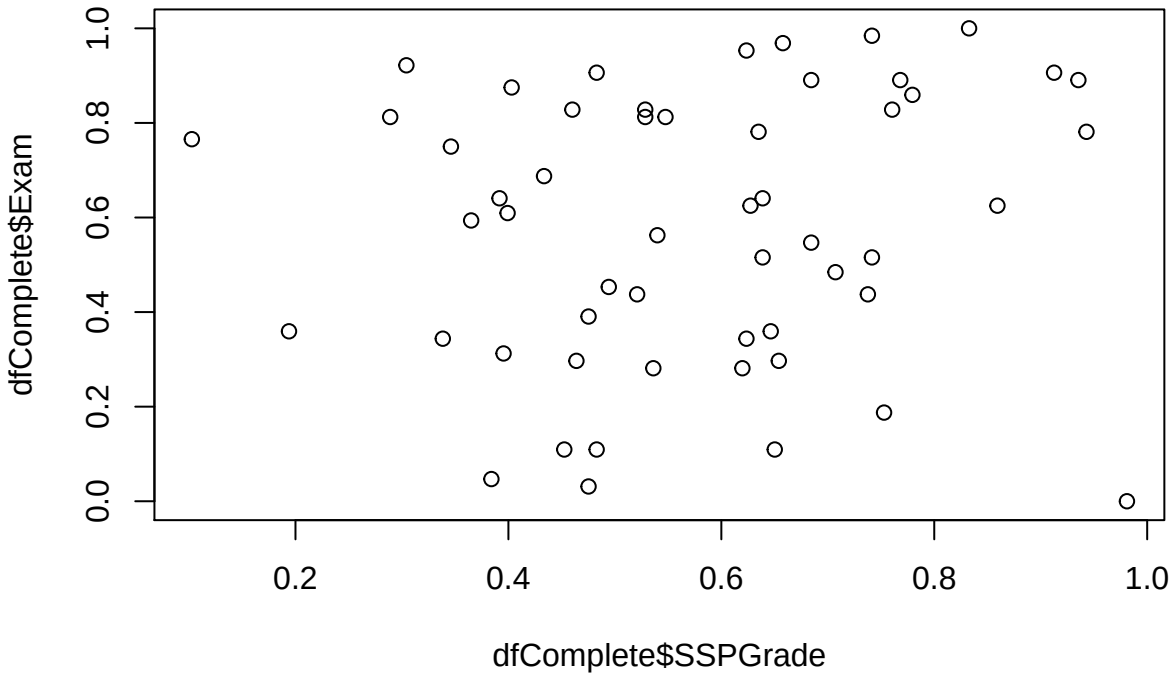

dfComplete$Exam
0.0
0.2
0.4
0.6
0.8
1.0
dfComplete$midterm
0.0
0.2
0.4
0.6
0.8
1.0

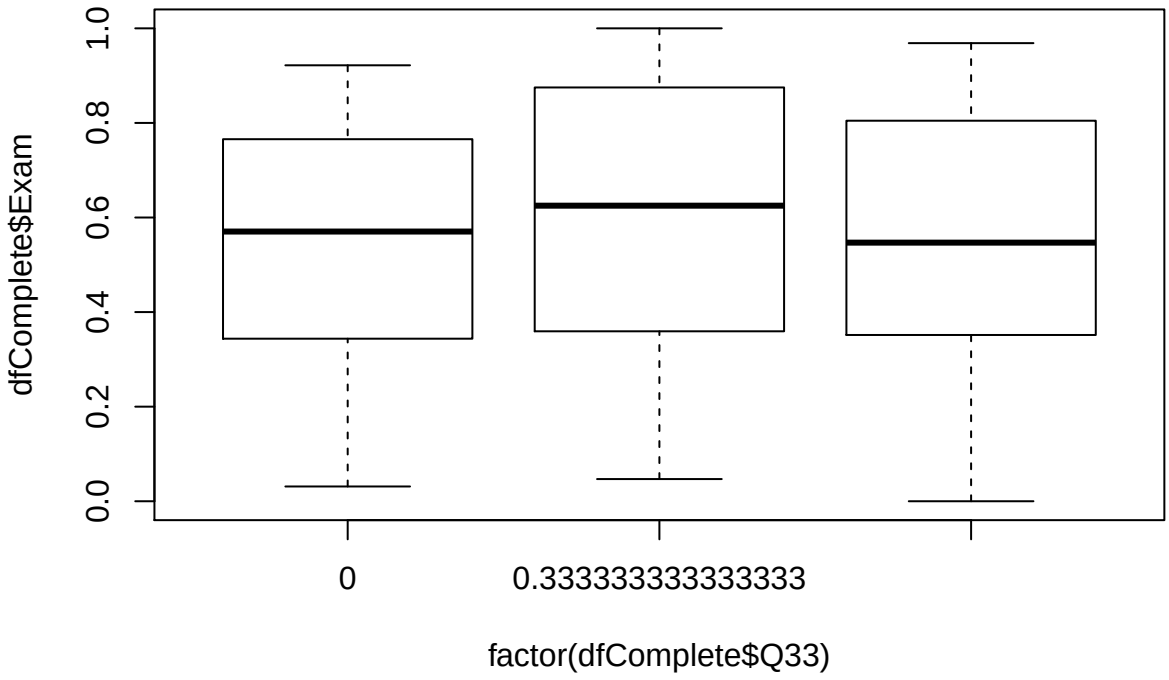

dfComplete$Exam
0.0
0.2
0.4
0.6
0.8
1.0
dfComplete$KhanCount
0.0
0.2
0.4
0.6
0.8
1.0

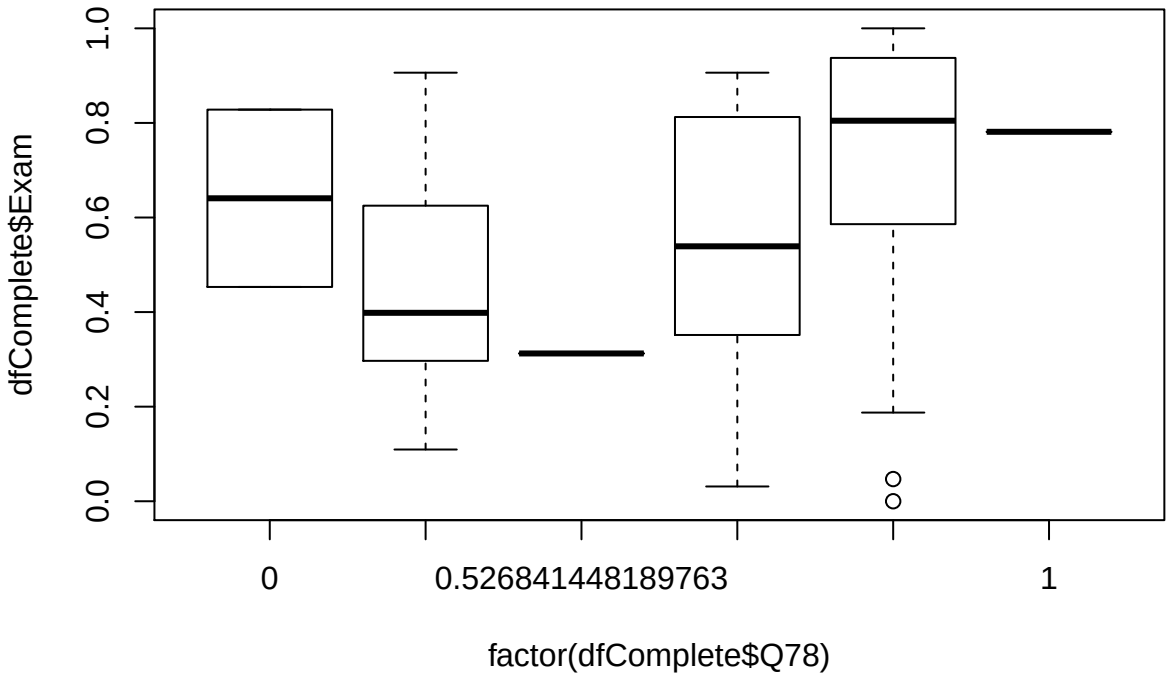

dfComplete$Exam
0.0
0.2
0.4
0.6
0.8
1.0
dfComplete$PeerSubmissionScore
0.0
0.2
0.4
0.6
0.8
1.0

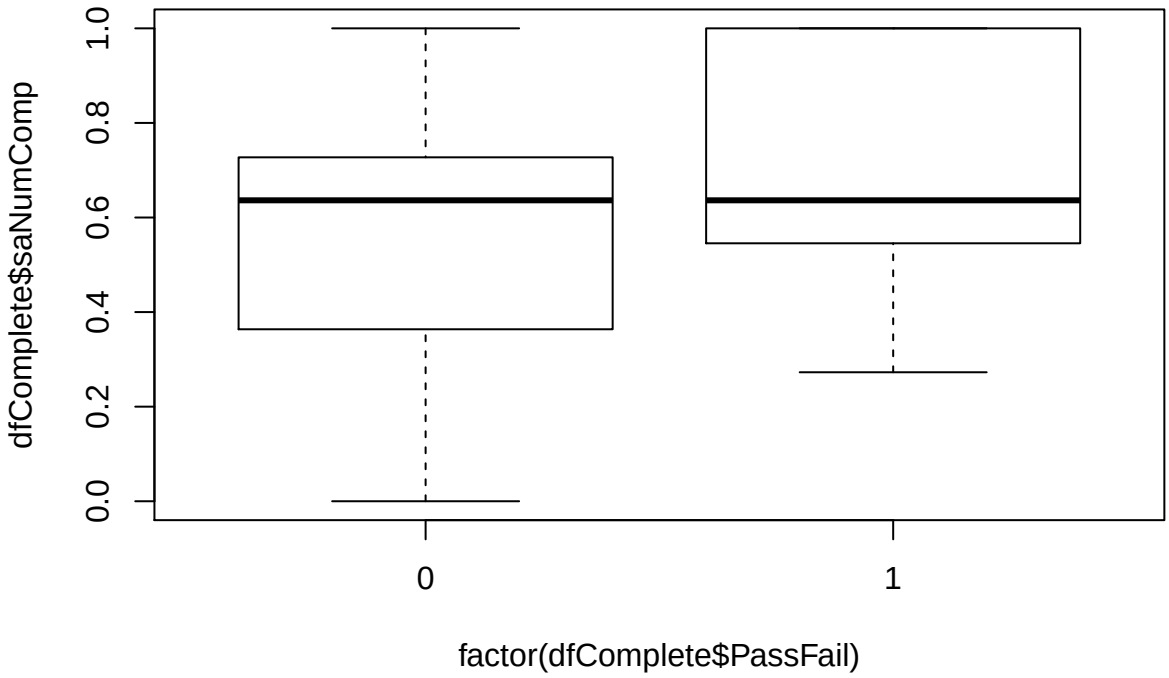

dfComplete$Exam
0.0
0.2
0.4
0.6
0.8
1.0
dfComplete$PeerFeedbackScore
0.0
0.2
0.4
0.6
0.8
1.0

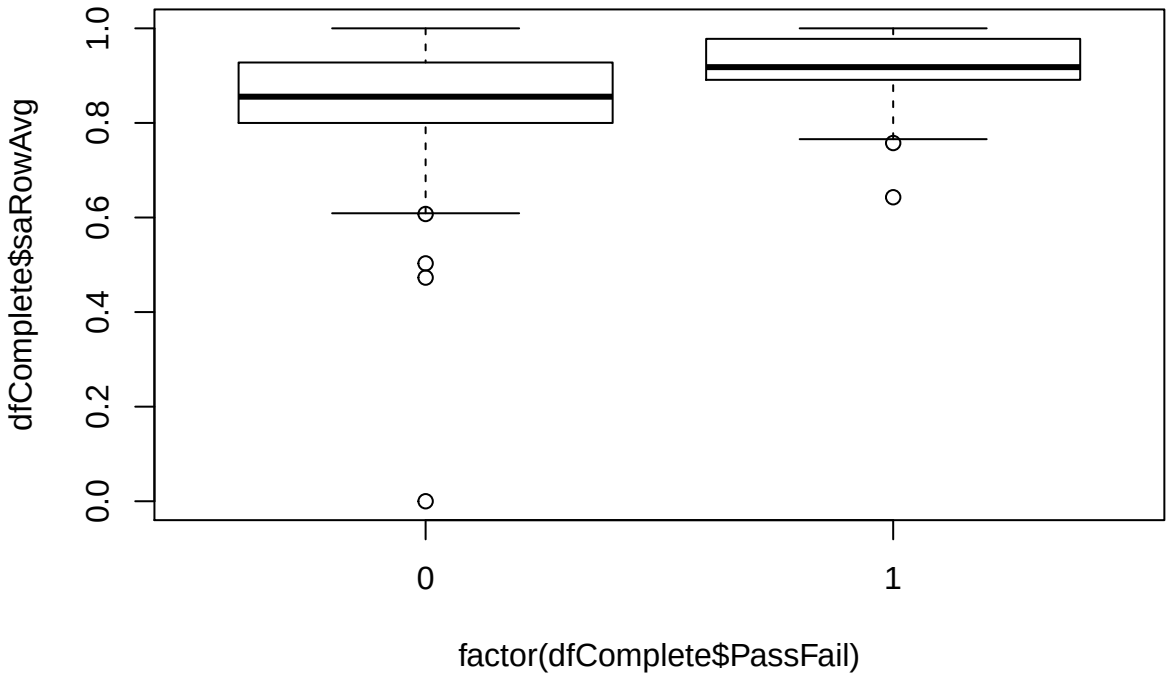

dfComplete$Exam
0.0
0.2
0.4
0.6
0.8
1.0
dfComplete$PeerCombinedScore

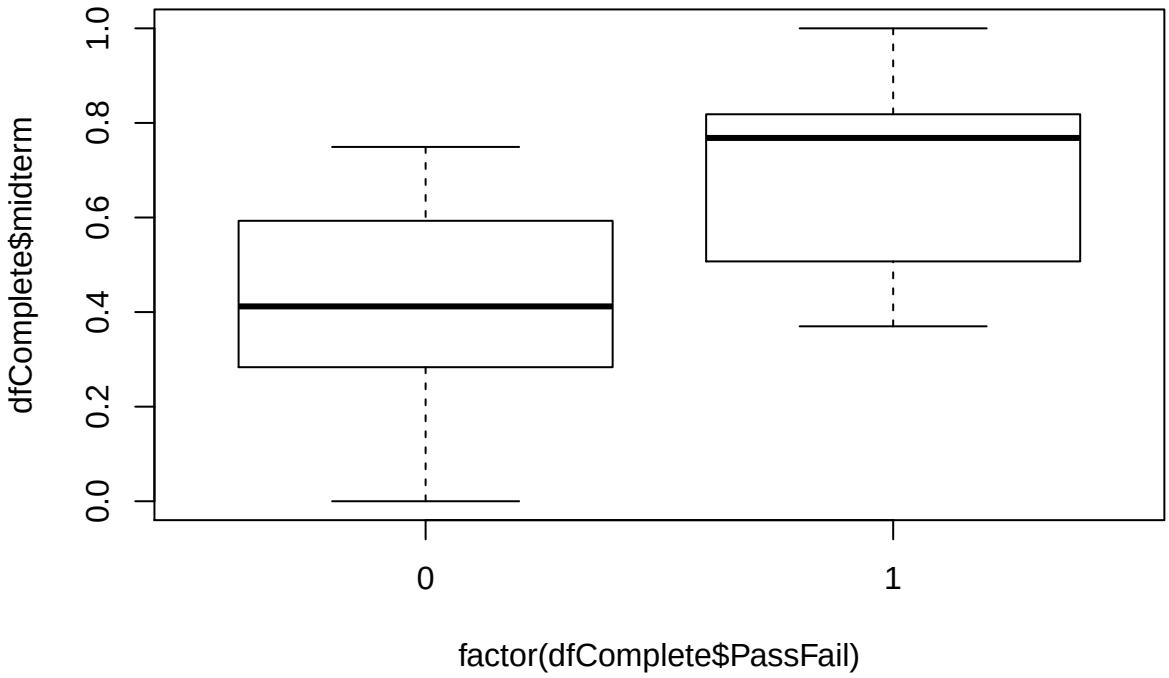

dfComplete$Exam
0.0
0.2
0.4
0.6
0.8
1.0
0.2
0.4
0.6
0.8
1.0
dfComplete$SSPGrade

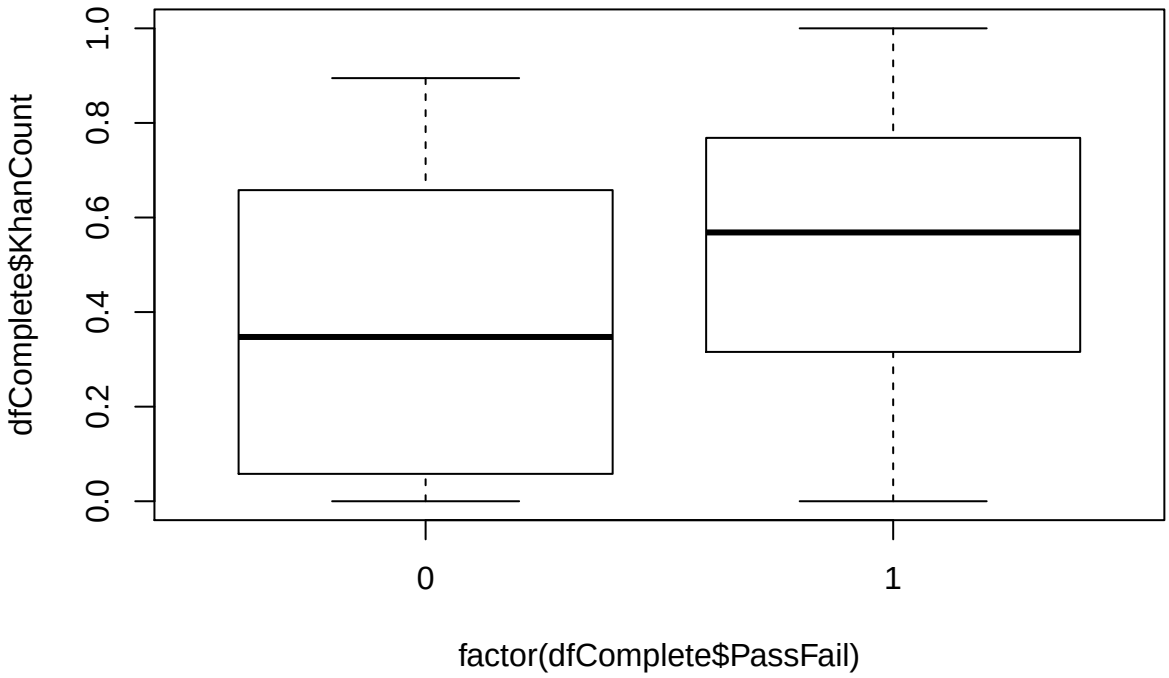

dfComplete$Exam
0.0
0.2
0.4
0.6
0.8
1.0
0
0.333333333333333
factor(dfComplete$Q33)

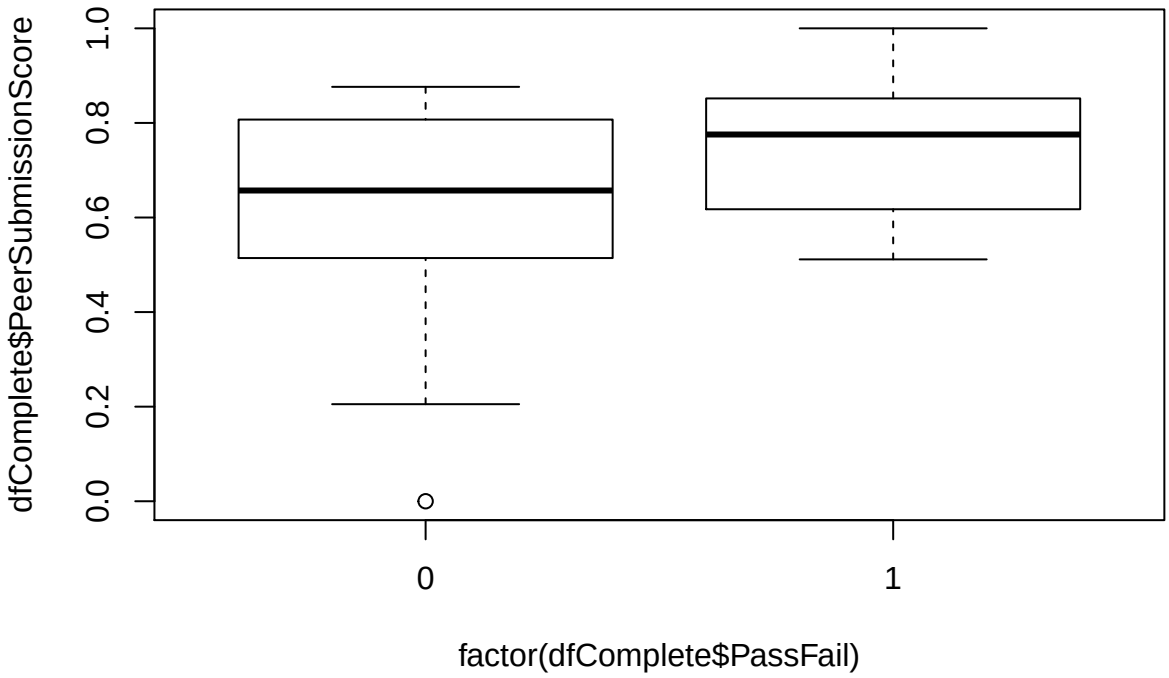

dfComplete$Exam
0.0
0.2
0.4
0.6
0.8
1.0
0
0.526841448189763
1
factor(dfComplete$Q78)

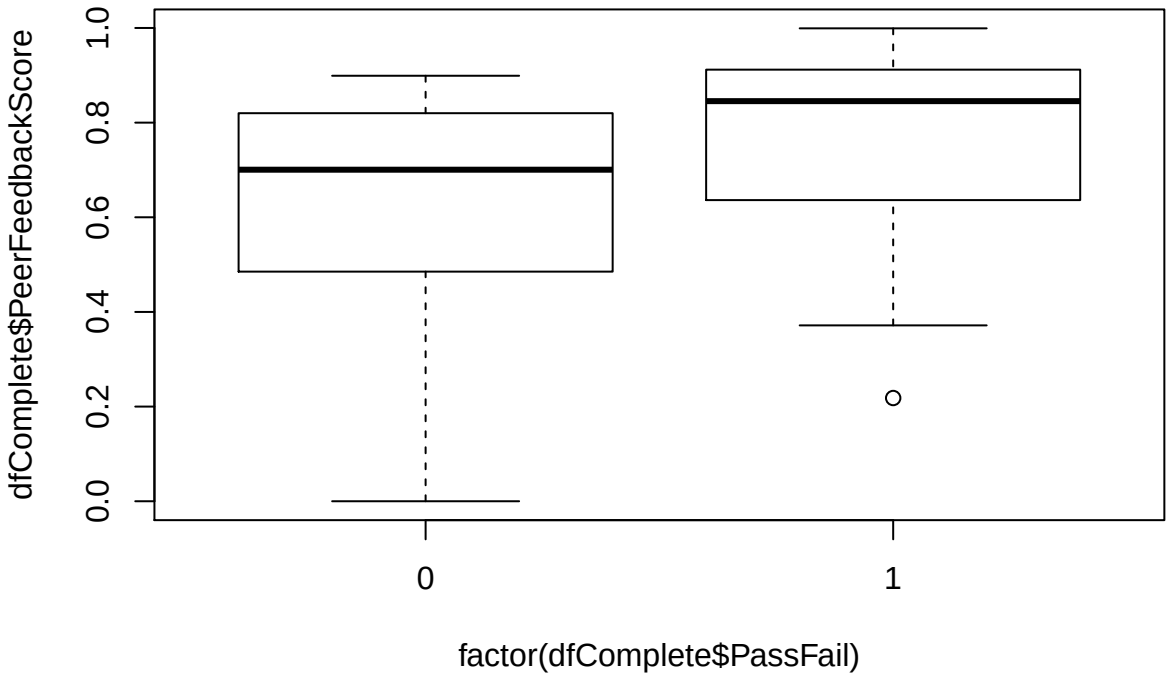

dfComplete$saNumComp
0.0
0.2
0.4
0.6
0.8
1.0
0
1
factor(dfComplete$PassFail)

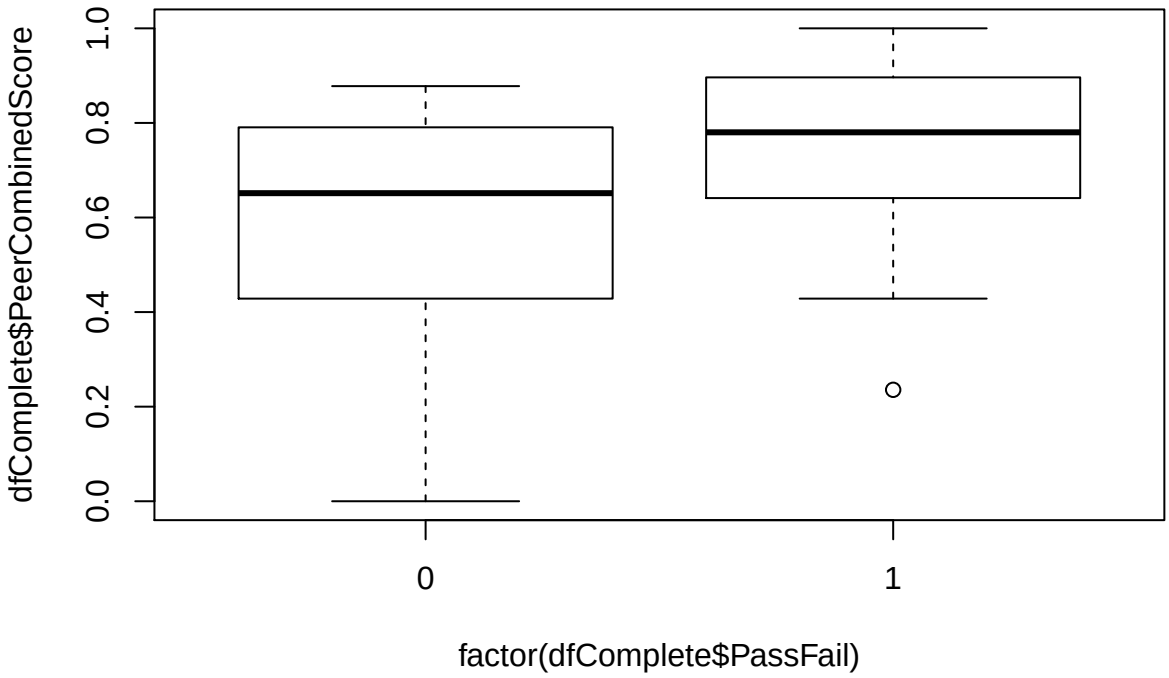

dfComplete$saRowAvg
0.0
0.2
0.4
0.6
0.8
1.0
0
1
factor(dfComplete$PassFail)

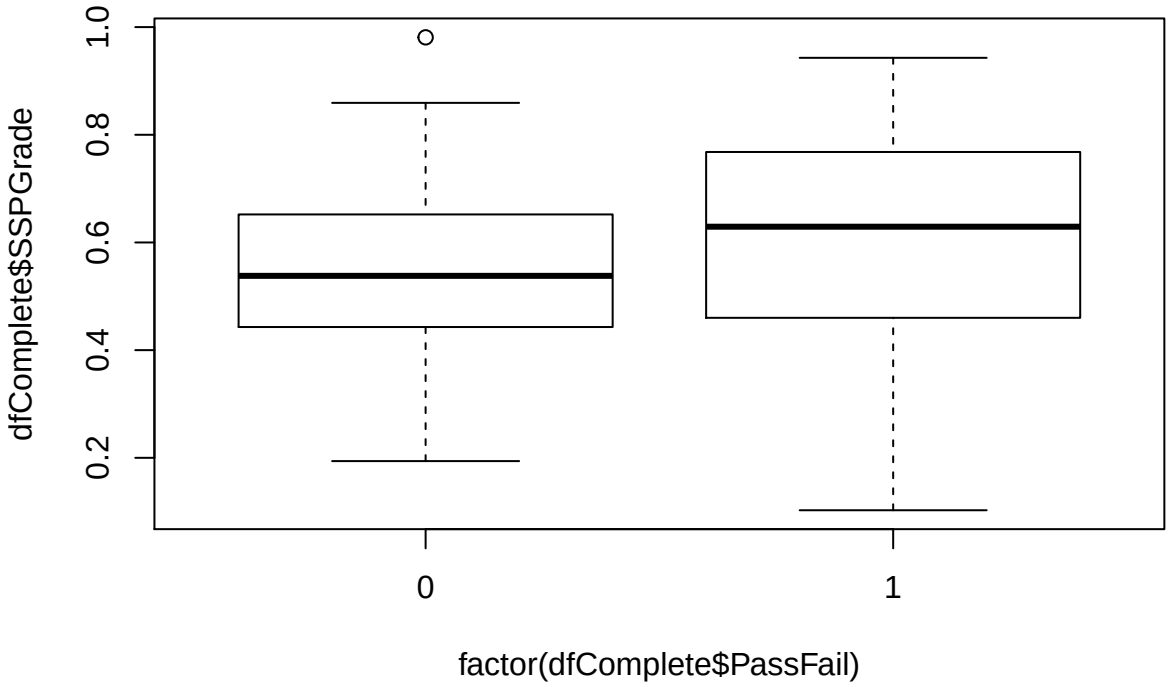

dfComplete$midterm
0.0
0.2
0.4
0.6
0.8
1.0
0
1
factor(dfComplete$PassFail)

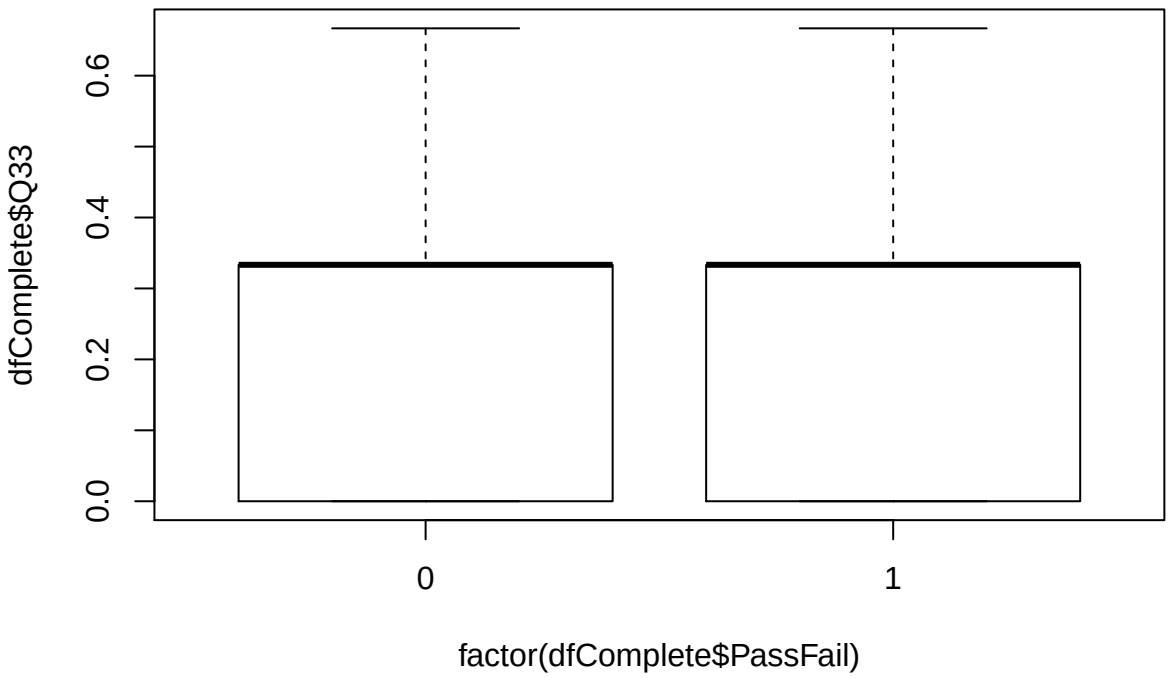

dfComplete$KhanCount
0.0
0.2
0.4
0.6
0.8
1.0
0
1
factor(dfComplete$PassFail)

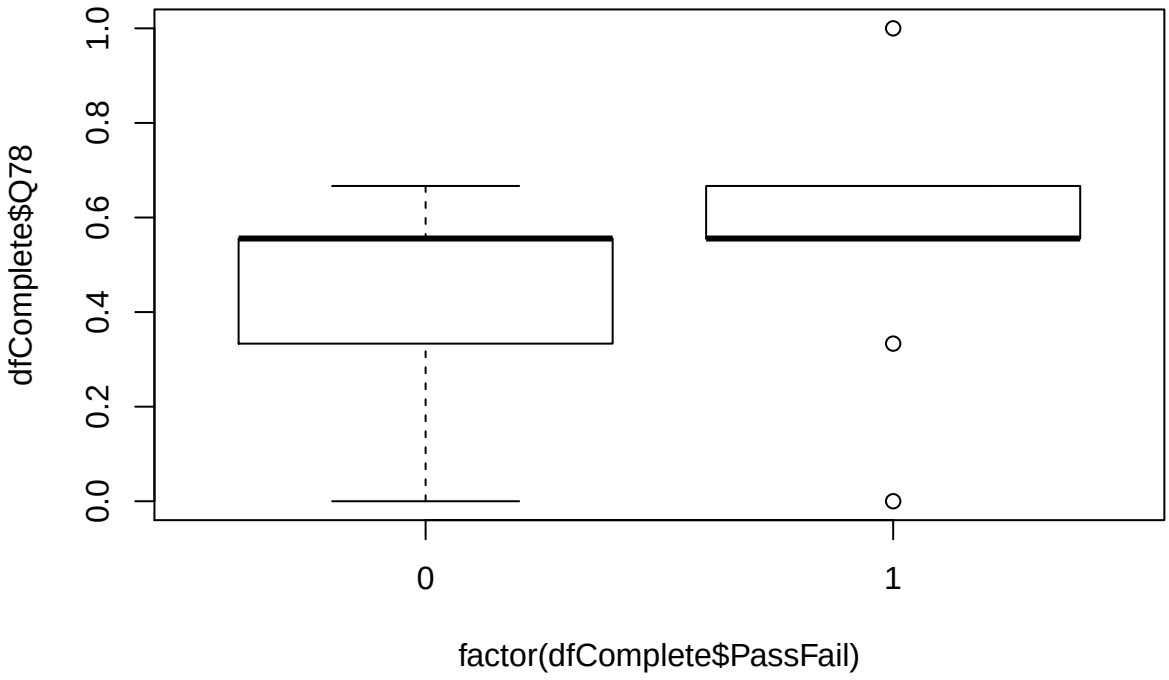

dfComplete$PeerSubmissionScore
0.0
0.2
0.4
0.6
0.8
1.0
0
1
factor(dfComplete$PassFail)


dfComplete$PeerFeedbackScore
0.0
0.2
0.4
0.6
0.8
1.0
0
1
factor(dfComplete$PassFail)


dfComplete$PeerCombinedScore
0.0
0.2
0.4
0.6
0.8
1.0
0
1
factor(dfComplete$PassFail)


dfComplete$SSPGrade
0.2
0.4
0.6
0.8
1.0
0
1
factor(dfComplete$PassFail)

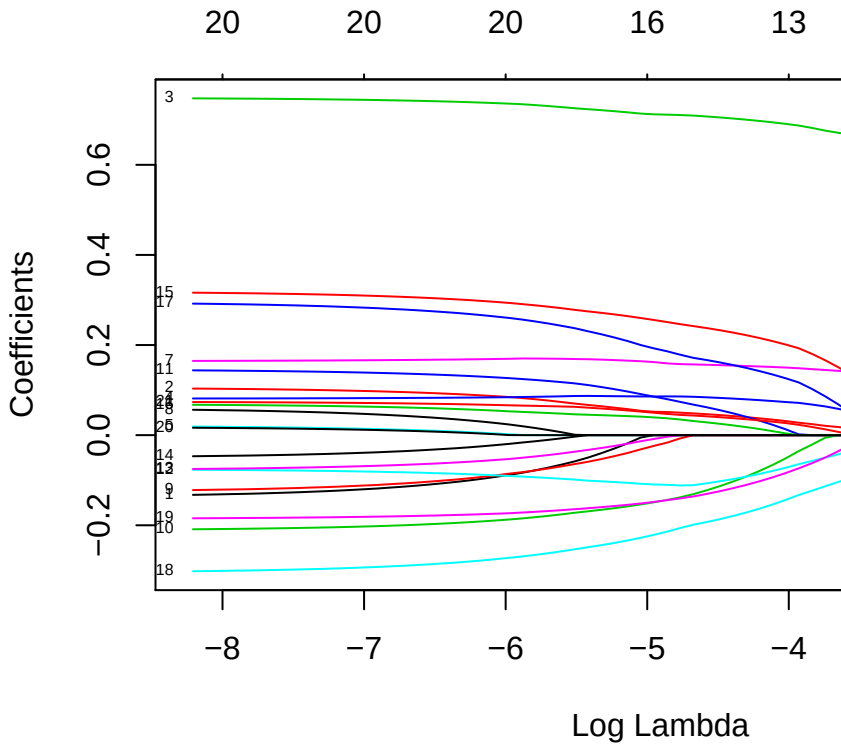

dfComplete$Q33
0.0
0.2
0.4
0.6
0
1
factor(dfComplete$PassFail)



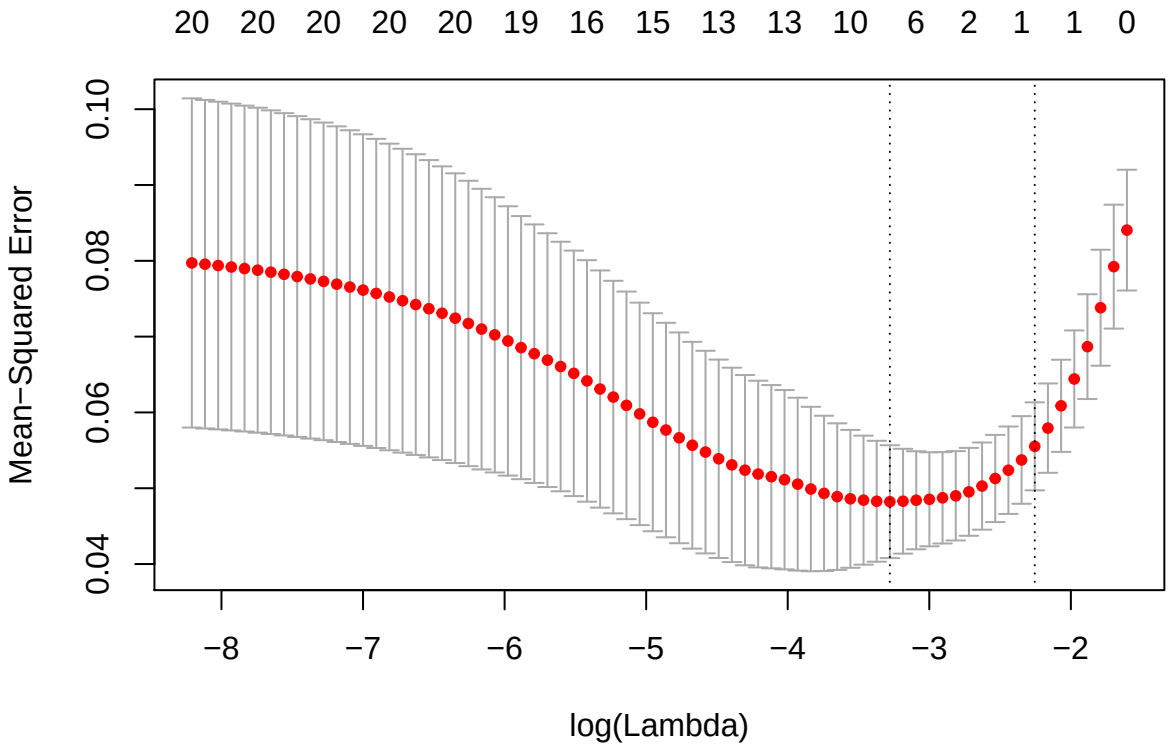
Analysis with linear models

Choosing with lasso (single questions and other predictors)

```
## Analysis of Variance Table
##
## Model 1: Exam ~ midterm + PeerCombinedScore + Q107 + Q82 + Q46
## Model 2: Exam ~ midterm + Q82
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     48 1.3363
## 2     51 1.9369 -3  -0.60065 7.1919 0.0004406 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Choosing with lasso (categories and other predictors)




```
## [1] 0.03768883
```

```
## [1] 0.1048714
```

Using best subset selection

Firstly for single questions in SSP

```
## [1] "midterm"
```

```
## [1] "midterm" "Q82"
```

```
## [1] "midterm" "Q46"     "Q82"
```

```
## [1] "midterm" "Q46"     "Q61"     "Q82"
```

```
## [1] "midterm" "Q46"     "Q61"     "Q82"     "Q107"
```

Fits a linear regression based on the single questions from above:

Secondly for the SSP categories:

```
## [1] "midterm"
```

```
## [1] "midterm"           "PeerCombinedScore"
```

```
## [1] "midterm"           "PeerCombinedScore" "HighSchoolTrust"
```

```
## [1] "midterm"           "PeerCombinedScore" "HighSchoolTrust"
## [4] "Selfcontrol"
```

```
## [1] "midterm"                 "PeerCombinedScore"
## [3] "HighSchoolTrust"         "PersonalTraitComparison"
## [5] "Selfcontrol"
```

Fits a linear regression based on the categories from above:

Forwards and backwards selection

Questions - which are even significant

```
## Single term additions
##
## Model:
## Exam ~ 1
##                      Df Sum of Sq    RSS     AIC  Pr(>Chi)
## saRowAvg              1   0.40551 3.9953 -136.61  0.022327 *
## saNumComp             1   0.55154 3.8493 -138.62  0.007166 **
## midterm               1   2.18456 2.2162 -168.43 1.156e-09 ***
## KhanCount             1   0.43402 3.9668 -137.00  0.017889 *
## PeerSubmissionScore   1   0.79019 3.6106 -142.08  0.001079 **
## PeerFeedbackScore     1   0.43818 3.9626 -137.05  0.017320 *
## PeerCombinedScore     1   0.67523 3.7256 -140.38  0.002708 **
## Q2                    1   0.56035 3.8405 -138.74  0.006689 **
## Q9                    1   0.32292 4.0779 -135.50  0.042496 *
## Q25                   1   0.35073 4.0501 -135.87  0.034197 *
## Q36                   1   0.31477 4.0860 -135.40  0.045300 *
## Q43                   1   0.52056 3.8802 -138.19  0.009126 **
## Q75                   1   0.42485 3.9760 -136.87  0.019211 *
## Q82                   1   1.10691 3.2939 -147.03 7.642e-05 ***
## Q99                   1   0.47949 3.9213 -137.62  0.012564 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Single term additions
##
## Model:
## Exam ~ 1
##                     Df Sum of Sq    RSS     AIC  Pr(>Chi)
## saNumComp            1   0.55154 3.8493 -138.62  0.007166 **
## midterm              1   2.18456 2.2162 -168.43 1.156e-09 ***
## PeerSubmissionScore  1   0.79019 3.6106 -142.08  0.001079 **
## PeerCombinedScore    1   0.67523 3.7256 -140.38  0.002708 **
## Q2                   1   0.56035 3.8405 -138.74  0.006689 **
## Q43                  1   0.52056 3.8802 -138.19  0.009126 **
## Q82                  1   1.10691 3.2939 -147.03 7.642e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Sig.level 0.05

Sig.level 0.01

Categories - which are even significant

```
## Single term additions
##
## Model:
```

```
## Exam ~ 1
##                       Df Sum of Sq    RSS     AIC  Pr(>Chi)
## saRowAvg               1   0.40551 3.9953 -136.61  0.022327 *
## saNumComp              1   0.55154 3.8493 -138.62  0.007166 **
## midterm                1   2.18456 2.2162 -168.43 1.156e-09 ***
## KhanCount              1   0.43402 3.9668 -137.00  0.017889 *
## PeerSubmissionScore    1   0.79019 3.6106 -142.08  0.001079 **
## PeerFeedbackScore      1   0.43818 3.9626 -137.05  0.017320 *
## PeerCombinedScore      1   0.67523 3.7256 -140.38  0.002708 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Single term additions
##
## Model:
## Exam ~ 1
##                       Df Sum of Sq    RSS     AIC  Pr(>Chi)
## saNumComp              1   0.55154 3.8493 -138.62  0.007166 **
## midterm                1   2.18456 2.2162 -168.43 1.156e-09 ***
## PeerSubmissionScore    1   0.79019 3.6106 -142.08  0.001079 **
## PeerCombinedScore      1   0.67523 3.7256 -140.38  0.002708 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

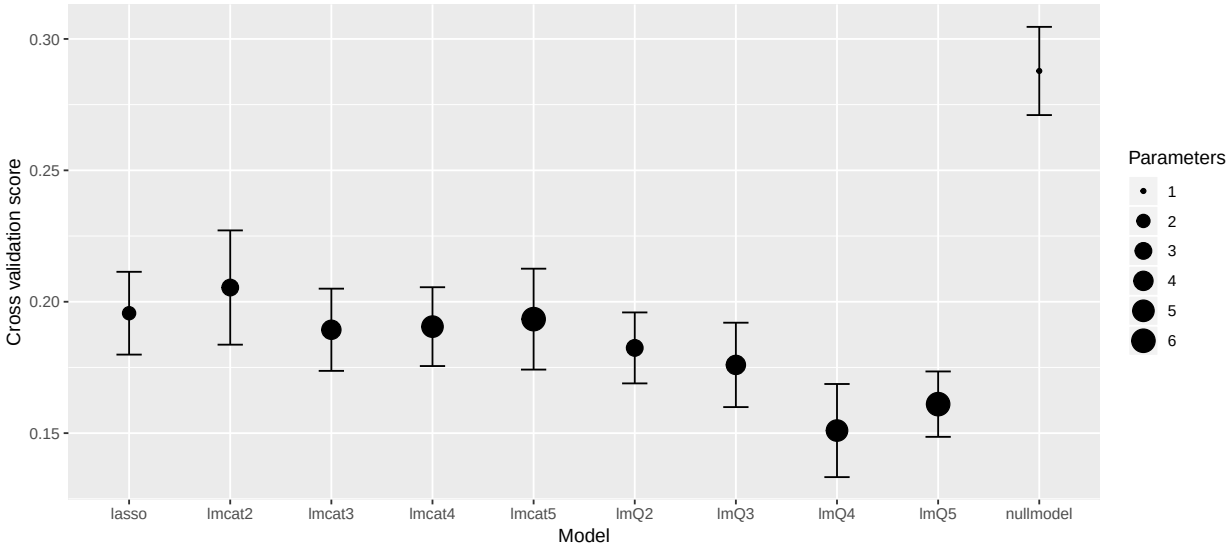
Sig.level 0.05

Sig.level 0.01

Doing crossvalidation for all models

post.lasso.1selmfit1 and lmcatfit1 are the same model so only one

lmQ5 is stable in the bottom:




```
## % latex table generated in R 3.5.1 by xtable 1.8-3 package
## % Tue Dec 11 22:57:44 2018
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrr}
```

```
##   \hline
##  & df & AIC \\
##   \hline
## null.model & 2.00 & 21.86 \\
##   post.lasso.1se & 4.00 & -18.46 \\
##   lmfit2 & 4.00 & -18.46 \\
##   lmfit3 & 5.00 & -25.53 \\
##   lmfit4 & 6.00 & -32.44 \\
##   lmfit5 & 7.00 & -38.86 \\
##   lmcatfit2 & 4.00 & -17.32 \\
##   lmcatfit3 & 5.00 & -20.52 \\
##   lmcatfit4 & 6.00 & -21.67 \\
##   lmcatfit5 & 7.00 & -23.17 \\
##    \hline
## \end{tabular}
## \end{table}
```

```
## % latex table generated in R 3.5.1 by xtable 1.8-3 package
## % Tue Dec 11 22:57:44 2018
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrr}
##   \hline
##  & df & AIC \\
##   \hline
## null.model & 2.00 & 25.83 \\
##   post.lasso.1se & 4.00 & -10.50 \\
##   lmfit2 & 4.00 & -10.50 \\
##   lmfit3 & 5.00 & -15.59 \\
##   lmfit4 & 6.00 & -20.51 \\
##   lmfit5 & 7.00 & -24.93 \\
##   lmcatfit2 & 4.00 & -9.36 \\
##   lmcatfit3 & 5.00 & -10.57 \\
##   lmcatfit4 & 6.00 & -9.74 \\
##   lmcatfit5 & 7.00 & -9.25 \\
##    \hline
## \end{tabular}
## \end{table}
```

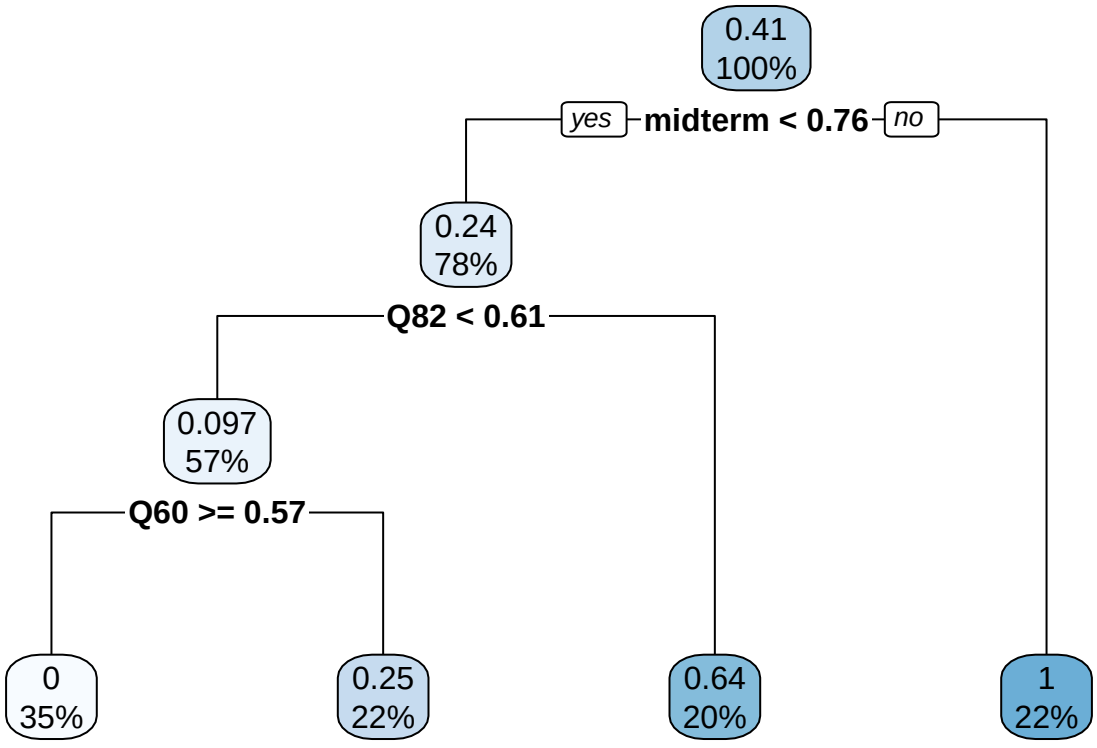
```
## % latex table generated in R 3.5.1 by xtable 1.8-3 package
## % Tue Dec 11 22:57:44 2018
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrrrr}
##   \hline
##  & Estimate & Std. Error & t value & Pr($>$$|$t$|$) \\
##   \hline
## (Intercept) & -0.40 & 0.12 & -3.40 & 0.00 \\
##   midterm & 0.68 & 0.10 & 7.03 & 0.00 \\
##   Q46 & 0.26 & 0.09 & 3.04 & 0.00 \\
##   Q61 & -0.20 & 0.06 & -3.27 & 0.00 \\
##   Q82 & 0.67 & 0.15 & 4.51 & 0.00 \\
##   Q107 & 0.19 & 0.07 & 2.85 & 0.01 \\
##    \hline
## \end{tabular}
```

```
## \end{table}
```

```
## % latex table generated in R 3.5.1 by xtable 1.8-3 package
## % Tue Dec 11 22:57:44 2018
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrr}
##   \hline
##  & 2.5 \% & 97.5 \% \\
##   \hline
## (Intercept) & -0.63 & -0.16 \\
##   midterm & 0.49 & 0.88 \\
##   Q46 & 0.09 & 0.43 \\
##   Q61 & -0.33 & -0.08 \\
##   Q82 & 0.37 & 0.97 \\
##   Q107 & 0.06 & 0.32 \\
##    \hline
## \end{tabular}
## \end{table}
```

Classification tree

Trying to predict passed failed with classification tree




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.2000  0.4667  0.5062  0.8167  1.0000
```

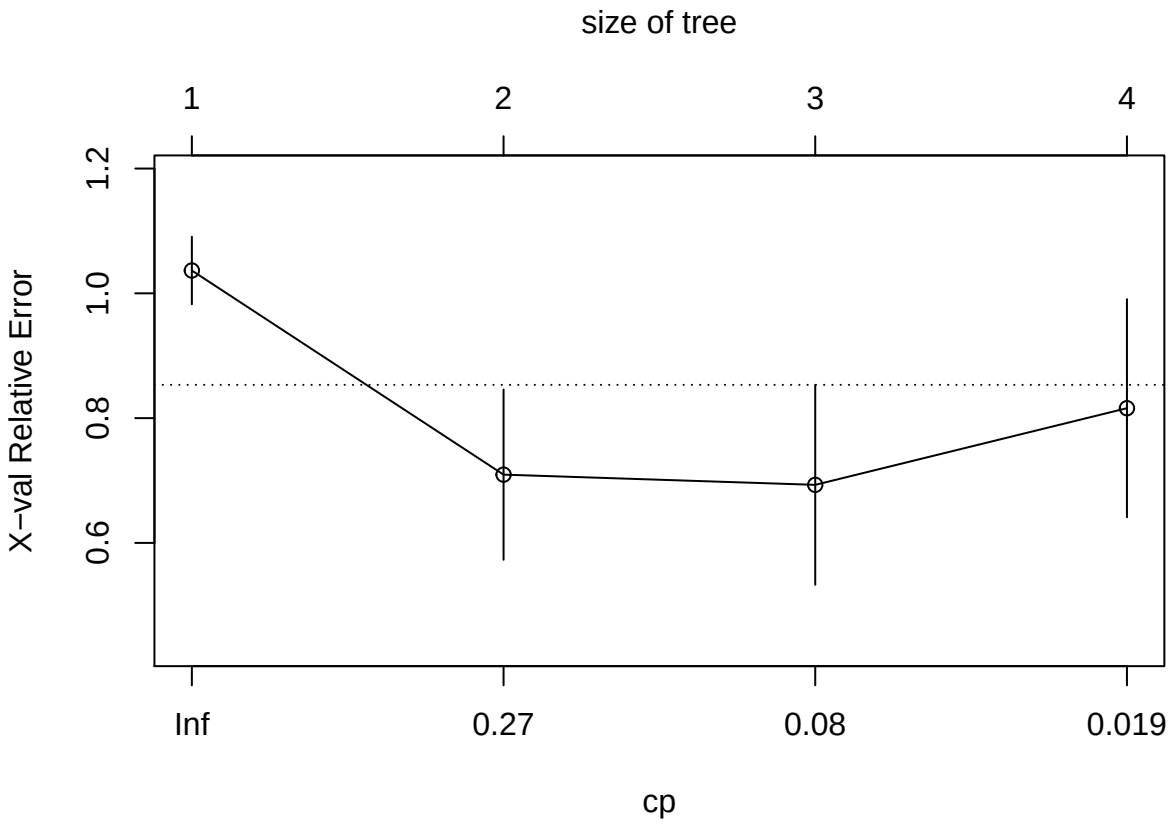
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##
```

```
## numeric(0)
```

```
##             midterm                  Q82                  Q55
##           5.8477923            3.2669139            1.0128315
##           KhanCount                   Q3                  Q35
##           0.9029982            0.9029982            0.9029982
##                Q81   PeerCombinedScore   PeerFeedbackScore
##           0.9029982            0.6447043            0.6447043
##                Q60 PeerSubmissionScore                 Q106
##           0.4596774            0.4298028            0.2298387
##                Q56                  Q61                  Q38
##           0.2298387            0.1915323            0.1532258
```

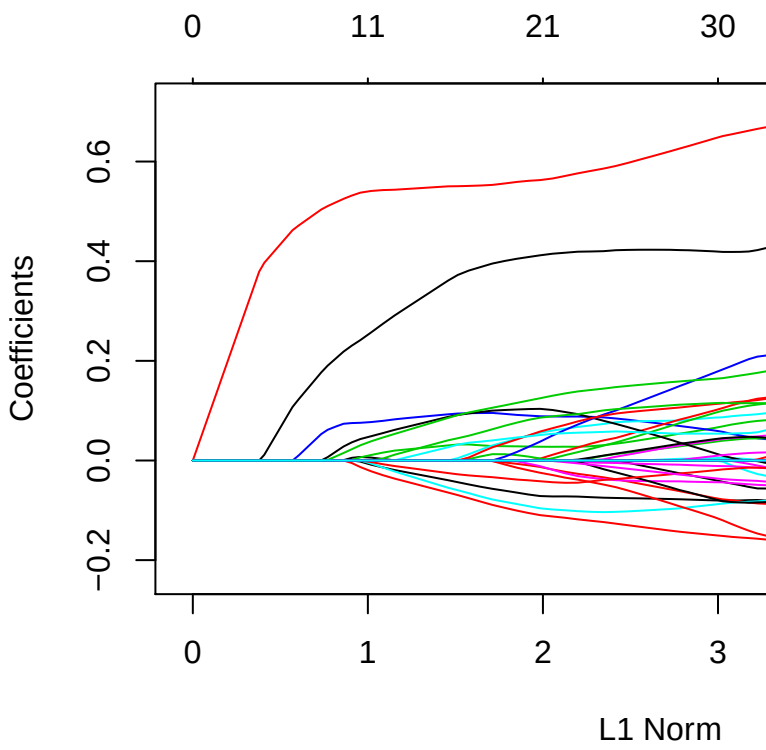
```
## Call:
## rpart(formula = PassFail ~ ., data = Qdatapf)
##   n= 54
##
##           CP nsplit rel error    xerror       xstd
## 1 0.41558442      0 1.0000000 1.0365806 0.05412692
## 2 0.18132308      1 0.5844156 0.7094032 0.13645118
## 3 0.03525935      2 0.4030925 0.6930333 0.16027513
## 4 0.01000000      3 0.3678332 0.8158651 0.17479445
##
## Variable importance
##             midterm                  Q82                  Q55
##                  35                   20                    6
##           KhanCount                   Q3                  Q35
##                   5                    5                    5
##                Q81   PeerCombinedScore   PeerFeedbackScore
##                   5                    4                    4
##                Q60 PeerSubmissionScore                 Q106
##                   3                    3                    1
##                Q56                  Q61                  Q38
##                   1                    1                    1
##
## Node number 1: 54 observations,    complexity param=0.4155844
##   mean=0.4074074, MSE=0.2414266
##   left son=2 (42 obs) right son=3 (12 obs)
##   Primary splits:
##       midterm           < 0.7569994 to the left,  improve=0.4155844, (0 missing)
##       PeerFeedbackScore < 0.8997808 to the left,  improve=0.2529644, (0 missing)
##       Q82               < 0.6111111 to the left,  improve=0.2046950, (0 missing)
##       saRowAvg          < 0.8891536 to the left,  improve=0.1920455, (0 missing)
##       Q43               < 0.7777778 to the left,  improve=0.1695422, (0 missing)
##   Surrogate splits:
##       KhanCount < 0.7842105 to the left,  agree=0.815, adj=0.167, (0 split)
##       Q3        < 0.8571429 to the left,  agree=0.815, adj=0.167, (0 split)
##       Q35       < 0.2777778 to the right, agree=0.815, adj=0.167, (0 split)
##       Q81       < 0.8333333 to the left,  agree=0.815, adj=0.167, (0 split)
##       Q82       < 0.8333333 to the left,  agree=0.815, adj=0.167, (0 split)
##
## Node number 2: 42 observations,    complexity param=0.1813231
```

```
##   mean=0.2380952, MSE=0.1814059
##   left son=4 (31 obs) right son=5 (11 obs)
##   Primary splits:
##       Q82               < 0.6111111 to the left,  improve=0.3102639, (0 missing)
##       PeerFeedbackScore < 0.8844334 to the left,  improve=0.2500000, (0 missing)
##       Q48               < 0.6666667 to the left,  improve=0.2343750, (0 missing)
##       Q22               < 0.3571429 to the right, improve=0.1941636, (0 missing)
##       Q43               < 0.7777778 to the left,  improve=0.1847874, (0 missing)
##   Surrogate splits:
##       Q55                 < 0.8571429 to the left,  agree=0.833, adj=0.364, (0 split)
##       PeerFeedbackScore   < 0.9215901 to the left,  agree=0.810, adj=0.273, (0 split)
##       PeerCombinedScore   < 0.8870538 to the left,  agree=0.810, adj=0.273, (0 split)
##       midterm             < 0.651552  to the left,  agree=0.786, adj=0.182, (0 split)
##       PeerSubmissionScore < 0.910241  to the left,  agree=0.786, adj=0.182, (0 split)
##
## Node number 3: 12 observations
##   mean=1, MSE=0
##
## Node number 4: 31 observations,    complexity param=0.03525935
##   mean=0.09677419, MSE=0.08740895
##   left son=8 (19 obs) right son=9 (12 obs)
##   Primary splits:
##       Q60               < 0.5714286 to the right, improve=0.1696429, (0 missing)
##       PeerCombinedScore < 0.7419641 to the left,  improve=0.1696429, (0 missing)
##       Q84               < 0.6111111 to the left,  improve=0.1696429, (0 missing)
##       Q55               < 0.5714286 to the right, improve=0.1483516, (0 missing)
##       Q59               < 0.4       to the right, improve=0.1191185, (0 missing)
##   Surrogate splits:
##       Q56  < 0.5714286 to the right, agree=0.806, adj=0.500, (0 split)
##       Q106 < 0.25      to the right, agree=0.806, adj=0.500, (0 split)
##       Q61  < 0.25      to the right, agree=0.774, adj=0.417, (0 split)
##       Q38  < 0.2777778 to the right, agree=0.742, adj=0.333, (0 split)
##       Q55  < 0.3571429 to the right, agree=0.742, adj=0.333, (0 split)
##
## Node number 5: 11 observations
##   mean=0.6363636, MSE=0.231405
##
## Node number 8: 19 observations
##   mean=0, MSE=0
##
## Node number 9: 12 observations
##   mean=0.25, MSE=0.1875
```

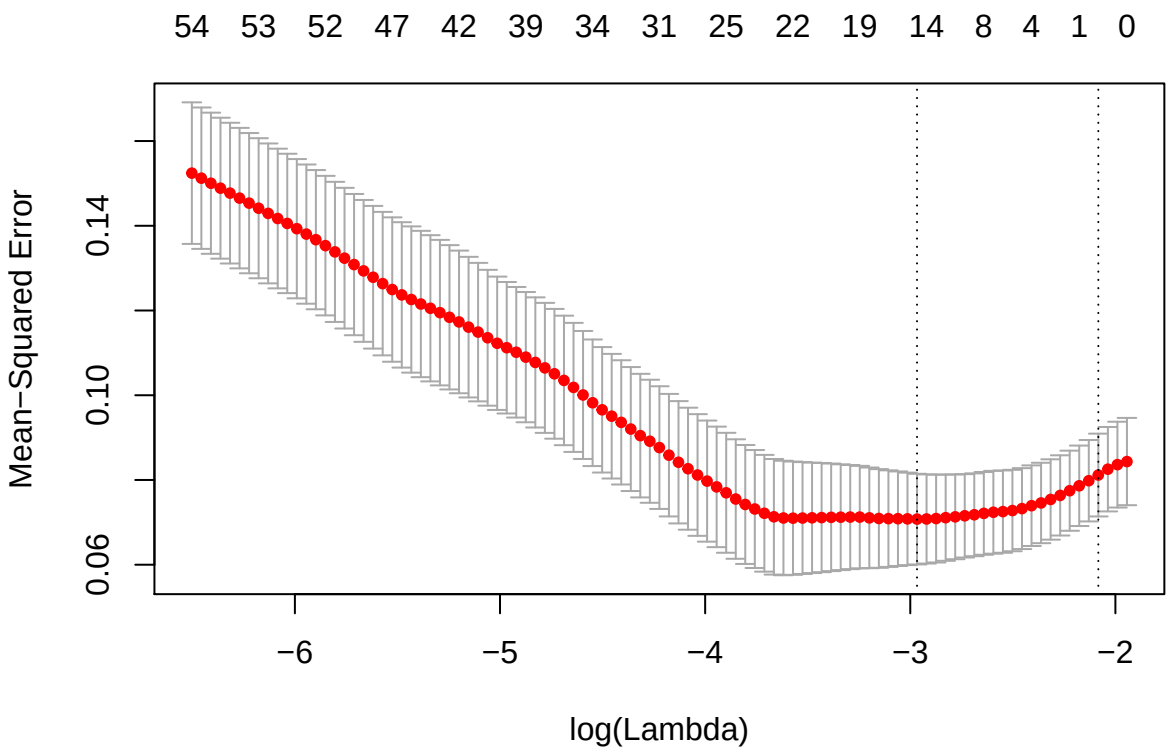


Predicting exam score excluding midterm

Trying to predict the exam score excluding midterm as predictor since it is part of the total score.



Choosing with lasso (single questions and other predictors):

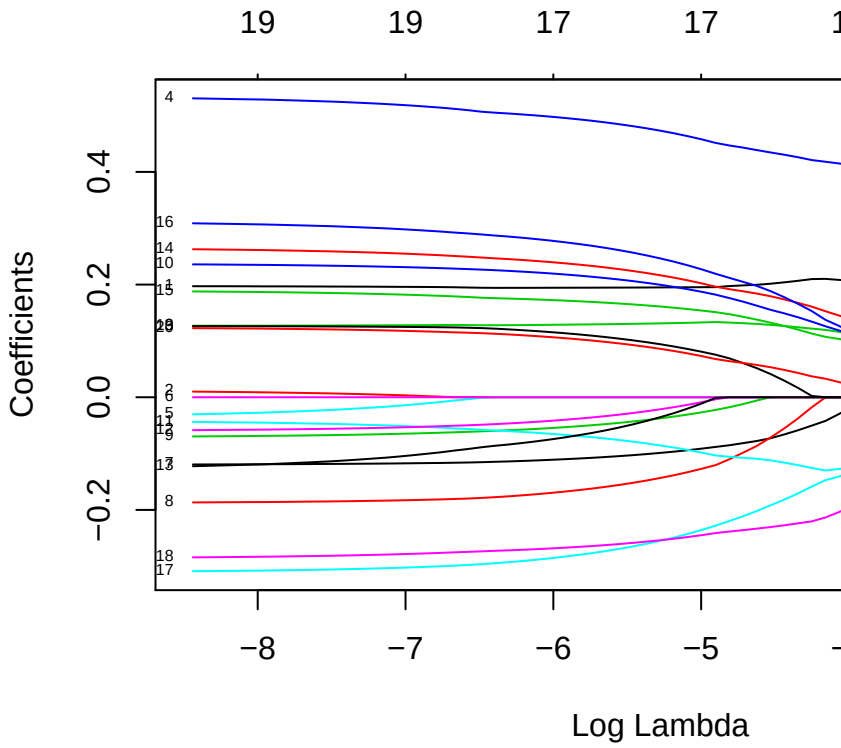



```
## [1] 0.0514536
```

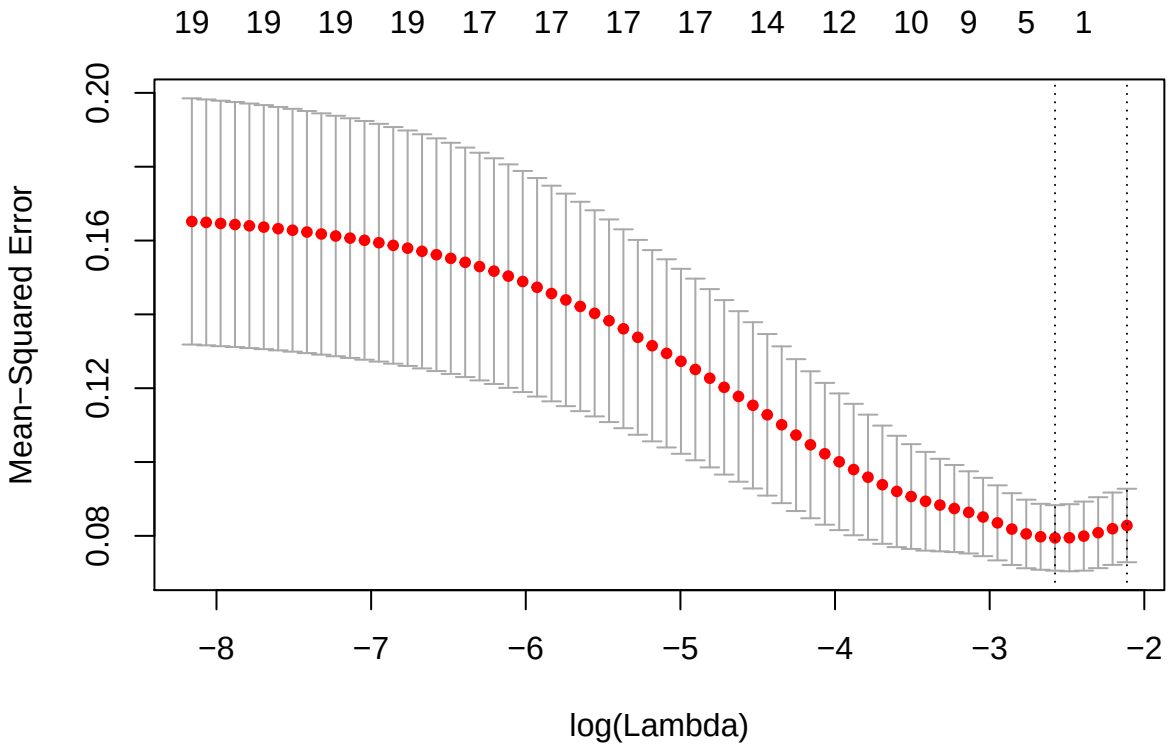
```
## [1] 0.1245243
```

Even though lasso does not choose them some seems to be significant anyway:

```
## Analysis of Variance Table
##
## Model 1: Exam ~ Q99 + Q82 + Q57 + Q56 + Q46 + Q43 + Q36 + Q25 + Q6 + Q2 +
##     PeerCombinedScore + PeerSubmissionScore + KhanCount + saRowAvg
## Model 2: Exam ~ Q82
##   Res.Df    RSS  Df Sum of Sq      F    Pr(>F)
## 1     39 1.3290
## 2     52 3.2939 -13   -1.9649 4.4354 0.0001448 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Choosing with lasso (categories and other predictors)




```
## [1] 0.07597109
```

```
## [1] 0.1209675
```

Using best subset selection

Firstly single SSP questions

```
## [1] "Q82"
```

```
## [1] "Q56" "Q82"
```

```
## [1] "PeerSubmissionScore" "Q56"                 "Q82"
```

```
## [1] "Q46" "Q56" "Q82" "Q99"
```

```
## [1] "PeerCombinedScore" "Q46"               "Q56"
## [4] "Q82"               "Q99"
```

Fits a linear regression based on the single questions from above:

```
##             df        AIC
## lmfittotal1  3   8.211992
## lmfittotal2  4   1.982298
## lmfittotal3  5  -4.424863
## lmfittotal4  6 -11.323344
## lmfittotal5  7 -19.068621
```

Secondly SSP categories

```
## [1] "PeerSubmissionScore"
```

```
## [1] "saRowAvg"            "PeerSubmissionScore"
```

```
## [1] "saRowAvg"            "PeerSubmissionScore" "Selfcontrol"
```

```
## [1] "saRowAvg"            "PeerSubmissionScore" "HighSchoolTrust"
## [4] "Selfcontrol"
```

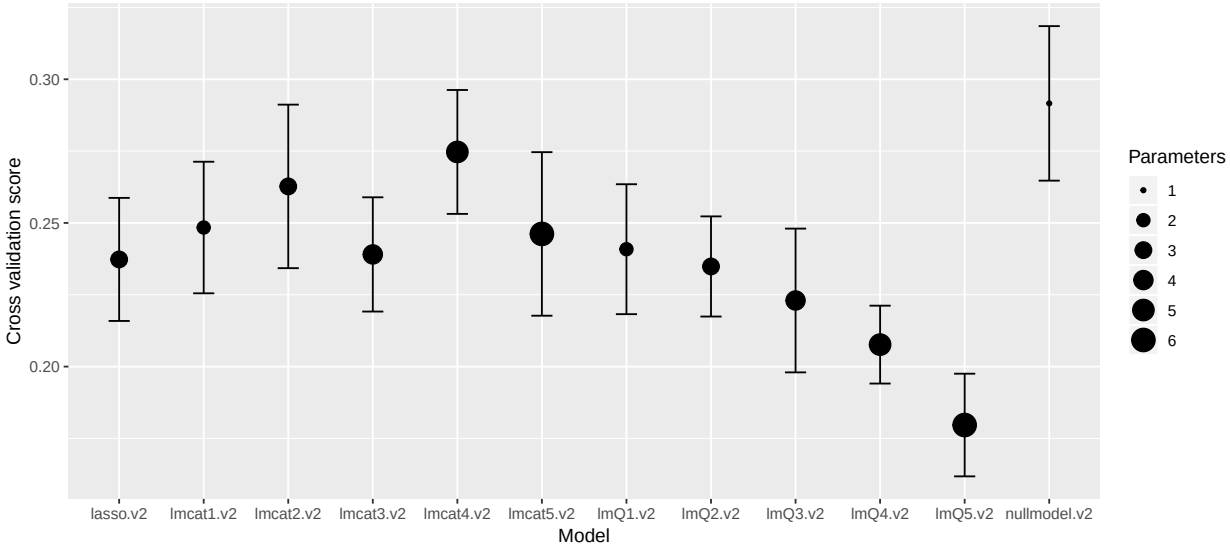
```
## [1] "saRowAvg"                   "PeerSubmissionScore"
## [3] "HighSchoolTrust"            "PerceivedAcademicAbilities"
## [5] "Selfcontrol"
```

Fits a linear regression based on the categories from above:

Table

```
##                df       AIC
## lmcatfittotal1  3 13.169684
## lmcatfittotal2  4 12.039546
## lmcatfittotal3  5 10.301624
## lmcatfittotal4  6  8.535013
## lmcatfittotal5  7  8.418489
```

Crossvalidation



none seems to be a clear improvement to the null model:

```
## % latex table generated in R 3.5.1 by xtable 1.8-3 package
## % Tue Dec 11 22:58:19 2018
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrr}
##   \hline
##  & df & AIC \\
##   \hline
## null.model & 2.00 & 21.86 \\
##   post.lasso.1se.total & 3.00 & 8.21 \\
##   lmfittotal1 & 3.00 & 8.21 \\
```

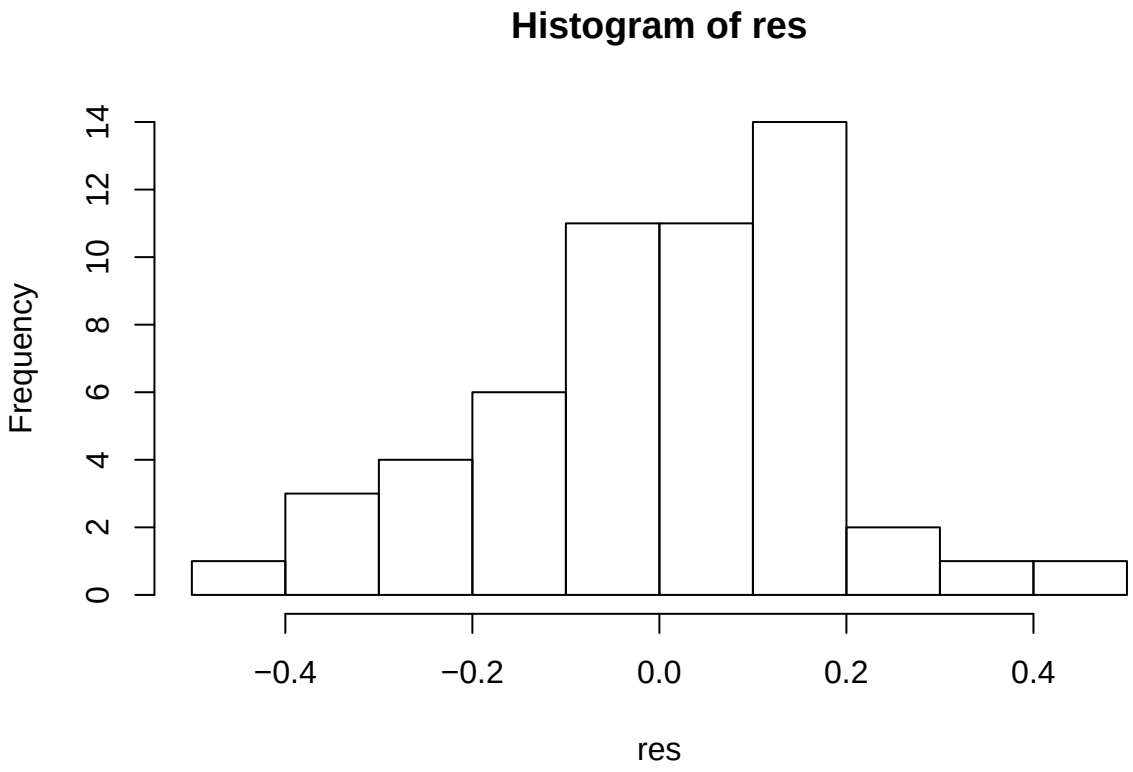
```
##   lmfittotal2 & 4.00 & 1.98 \\
##   lmfittotal3 & 5.00 & -4.42 \\
##   lmfittotal4 & 6.00 & -11.32 \\
##   lmfittotal5 & 7.00 & -19.07 \\
##   lmcatfittotal1 & 3.00 & 13.17 \\
##   lmcatfittotal2 & 4.00 & 12.04 \\
##   lmcatfittotal3 & 5.00 & 10.30 \\
##   lmcatfittotal4 & 6.00 & 8.54 \\
##   lmcatfittotal5 & 7.00 & 8.42 \\
##    \hline
## \end{tabular}
## \end{table}
```

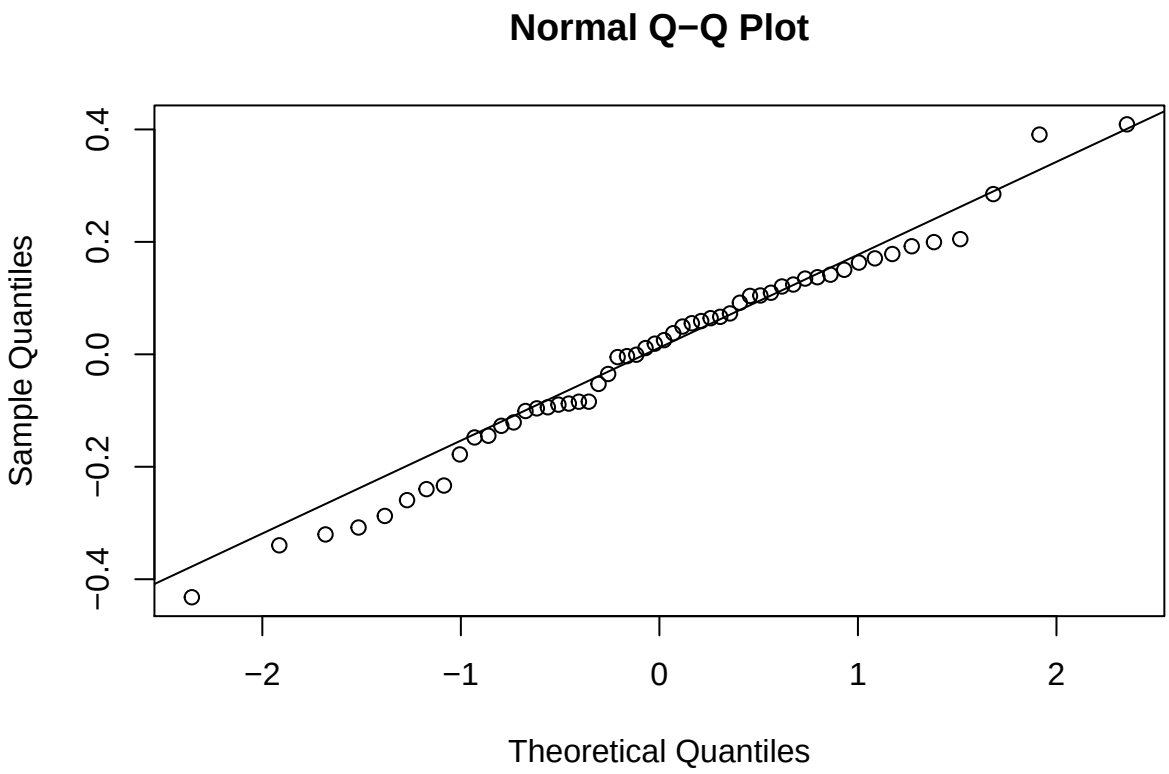
```
##                       df       AIC
## null.model             2 25.834734
## post.lasso.1se.total   3 14.178944
## lmfittotal1            3 14.178944
## lmfittotal2            4  9.938234
## lmfittotal3            5  5.520057
## lmfittotal4            6  0.610560
## lmfittotal5            7 -5.145732
## lmcatfittotal1         3 19.136636
## lmcatfittotal2         4 19.995482
## lmcatfittotal3         5 20.246544
## lmcatfittotal4         6 20.468917
## lmcatfittotal5         7 22.341377
```

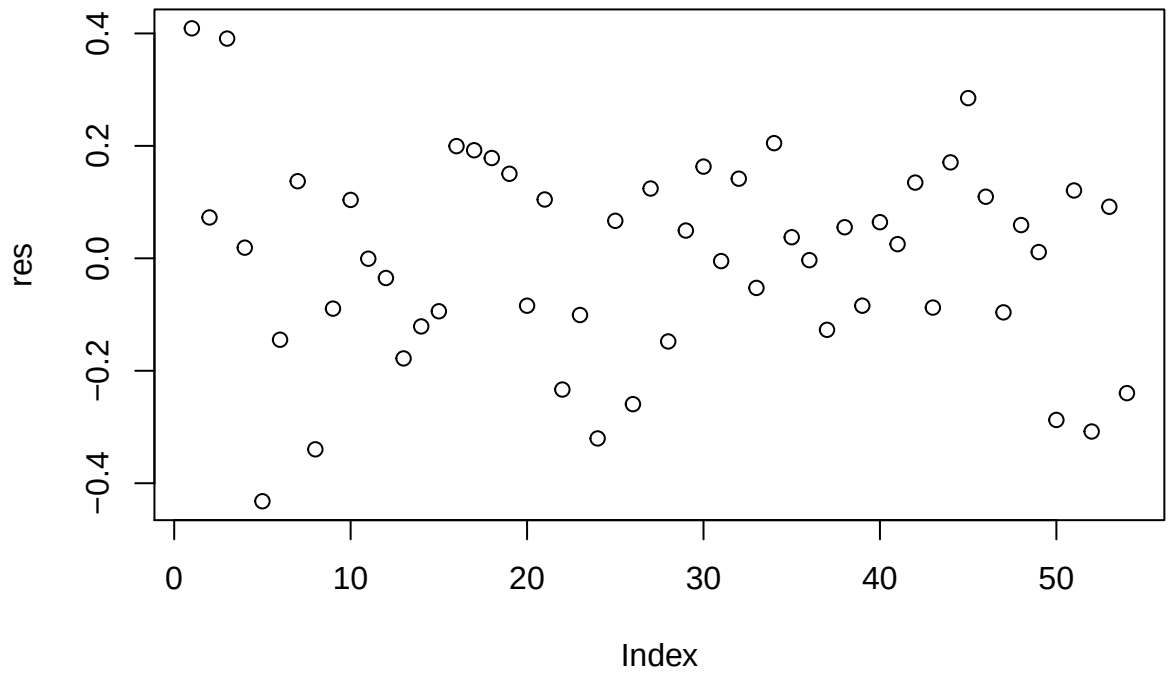
```
## % latex table generated in R 3.5.1 by xtable 1.8-3 package
## % Tue Dec 11 22:58:19 2018
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrrrr}
##   \hline
##  & Estimate & Std. Error & t value & Pr($>$$|$t$|$) \\
##   \hline
## (Intercept) & -0.61 & 0.19 & -3.18 & 0.00 \\
##   PeerCombinedScore & 0.34 & 0.11 & 3.08 & 0.00 \\
##   Q46 & 0.34 & 0.10 & 3.33 & 0.00 \\
##   Q56 & -0.32 & 0.10 & -3.31 & 0.00 \\
##   Q82 & 0.95 & 0.17 & 5.73 & 0.00 \\
##   Q99 & 0.53 & 0.16 & 3.32 & 0.00 \\
##    \hline
## \end{tabular}
## \end{table}
```

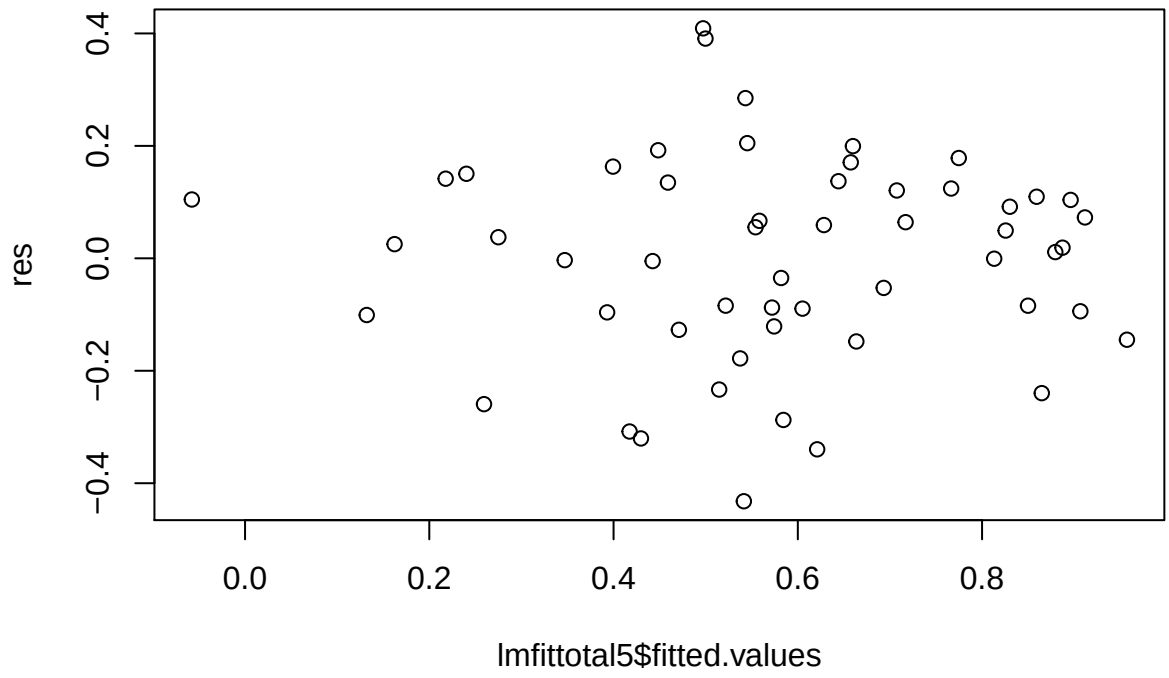
```
## % latex table generated in R 3.5.1 by xtable 1.8-3 package
## % Tue Dec 11 22:58:19 2018
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrr}
##   \hline
##  & 2.5 \% & 97.5 \% \\
##   \hline
## (Intercept) & -0.99 & -0.22 \\
##   PeerCombinedScore & 0.12 & 0.56 \\
```

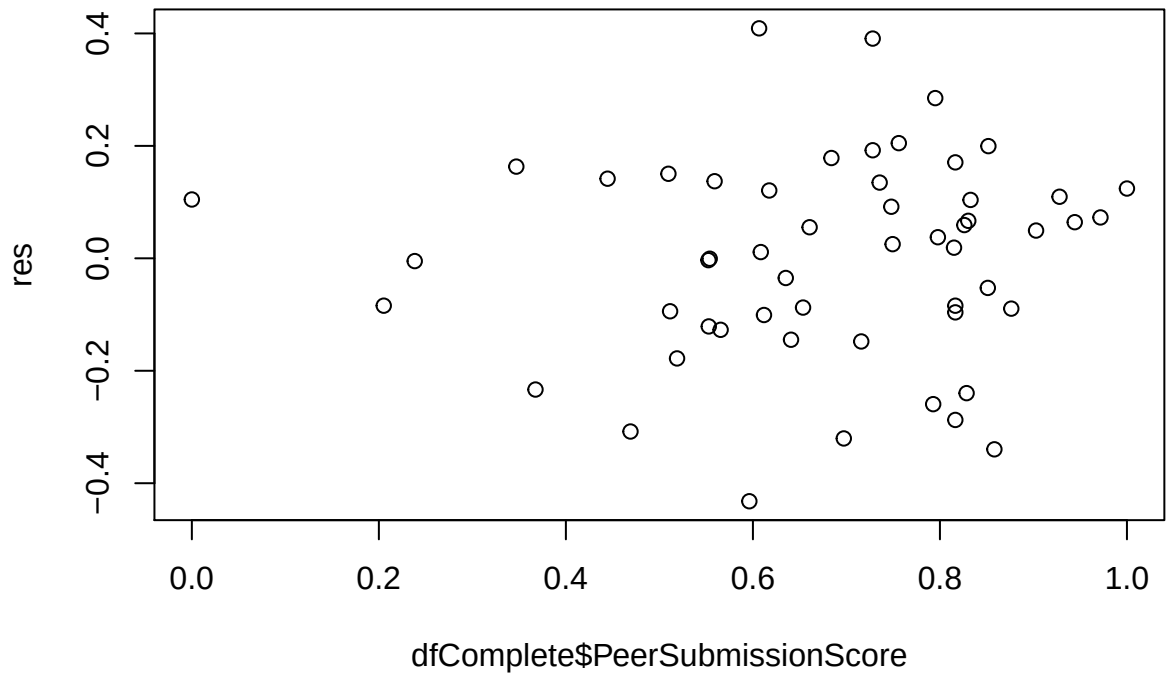
```
##   Q46 & 0.13 & 0.54 \\
##   Q56 & -0.51 & -0.12 \\
##   Q82 & 0.62 & 1.29 \\
##   Q99 & 0.21 & 0.86 \\
##    \hline
## \end{tabular}
## \end{table}
```

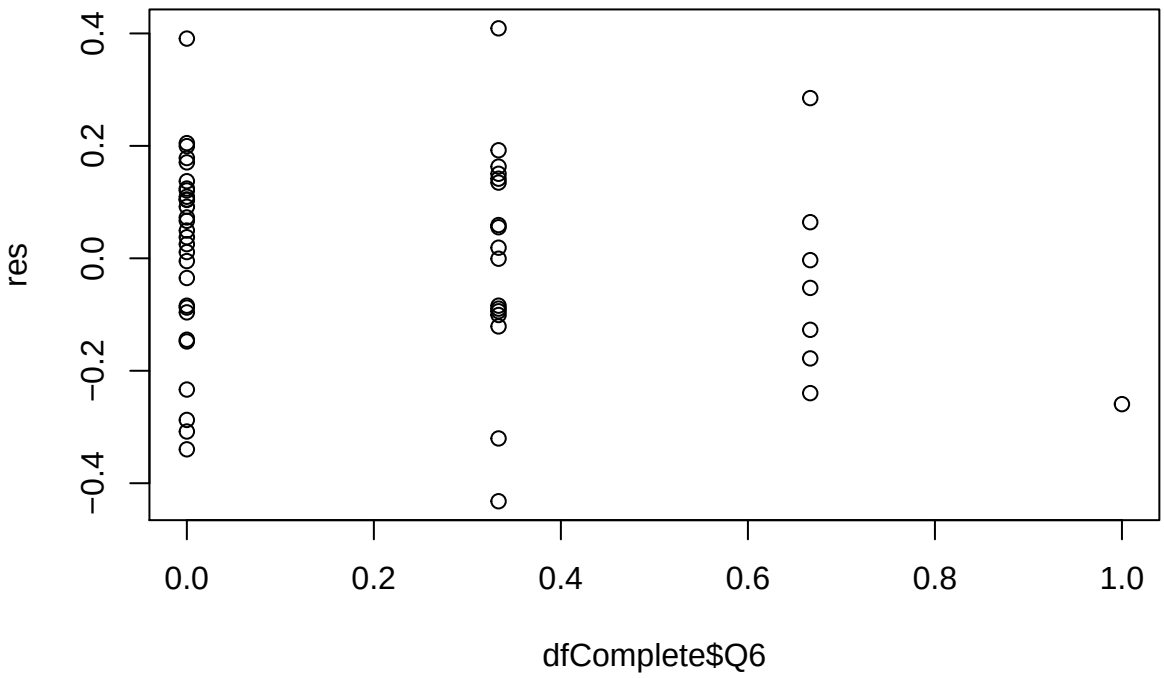


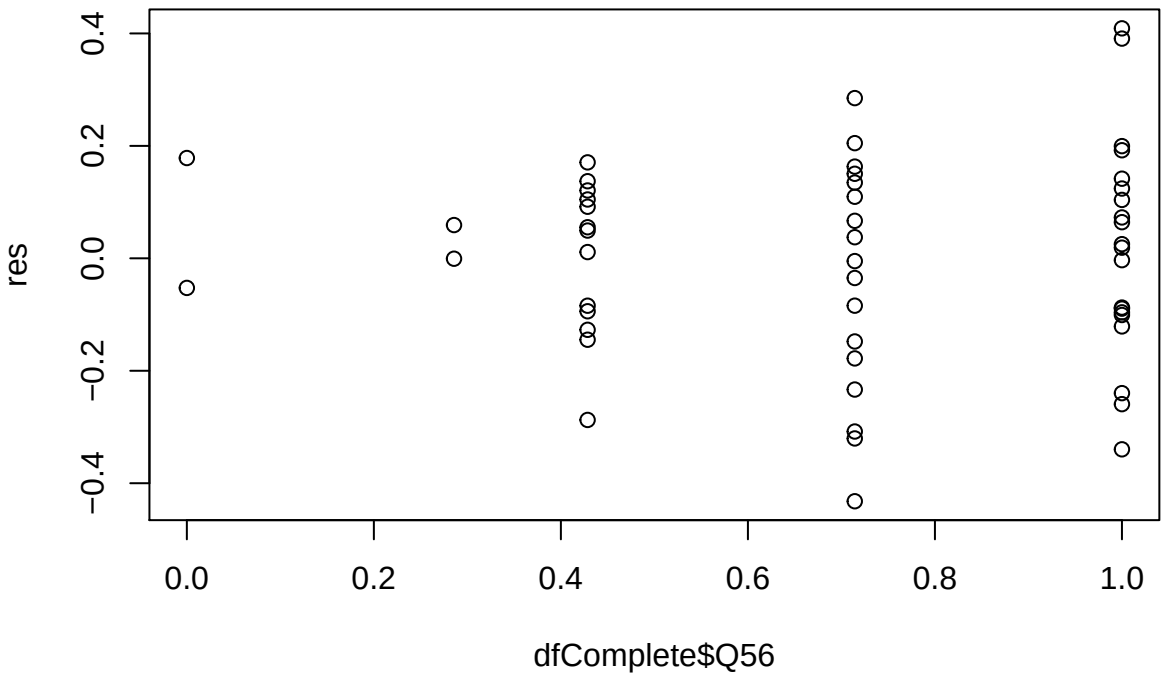

Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−0.4
−0.2
0.0
0.2
0.4
−2
−1
0
1
2

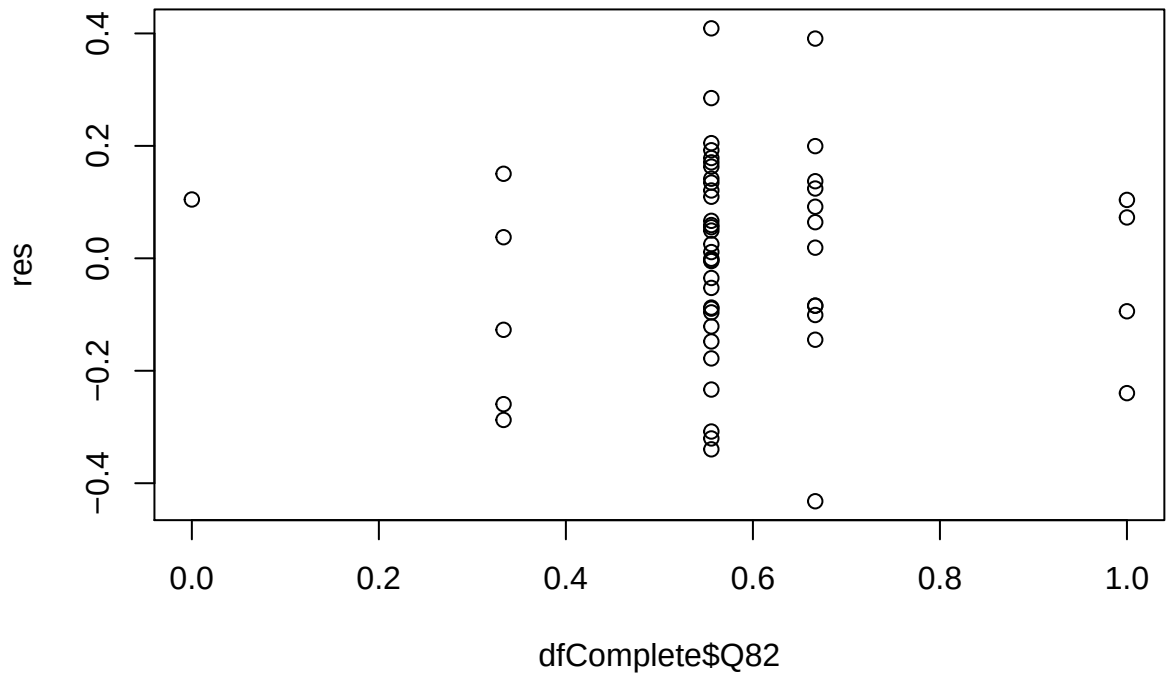

res
0.4
0.2
0.0
−0.2
−0.4
0
10
20
30
40
50
Index

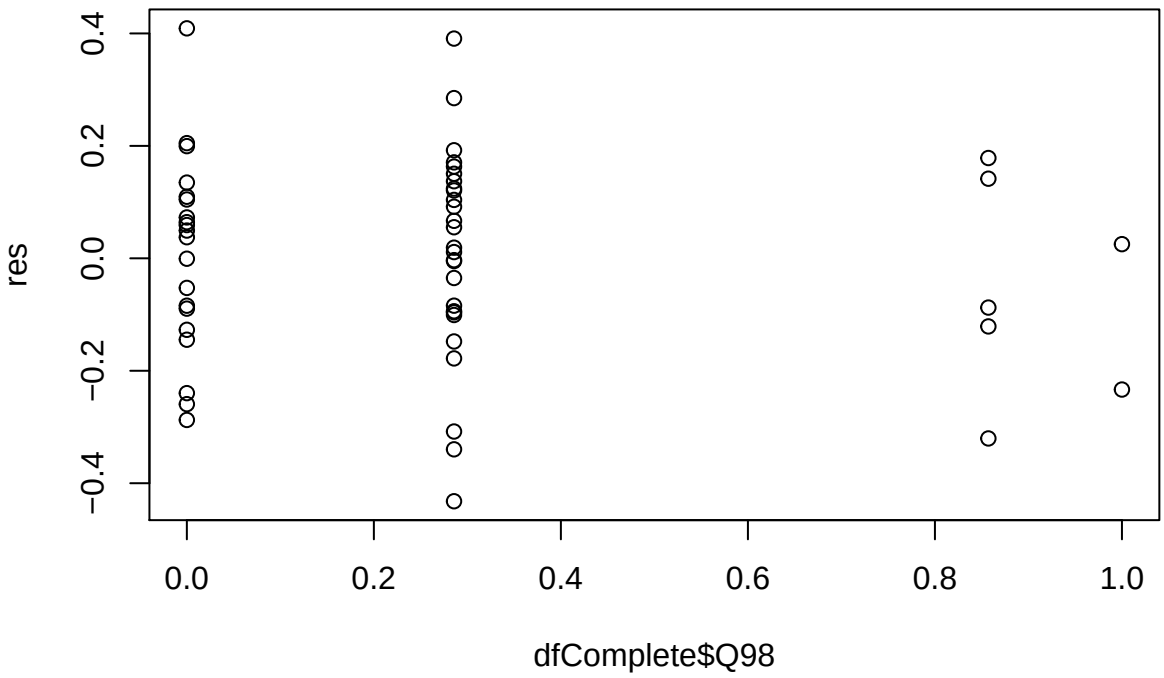

res
0.4
0.2
0.0
−0.2
−0.4
0.0
0.2
0.4
0.6
0.8
lmfittotal5$fitted.values

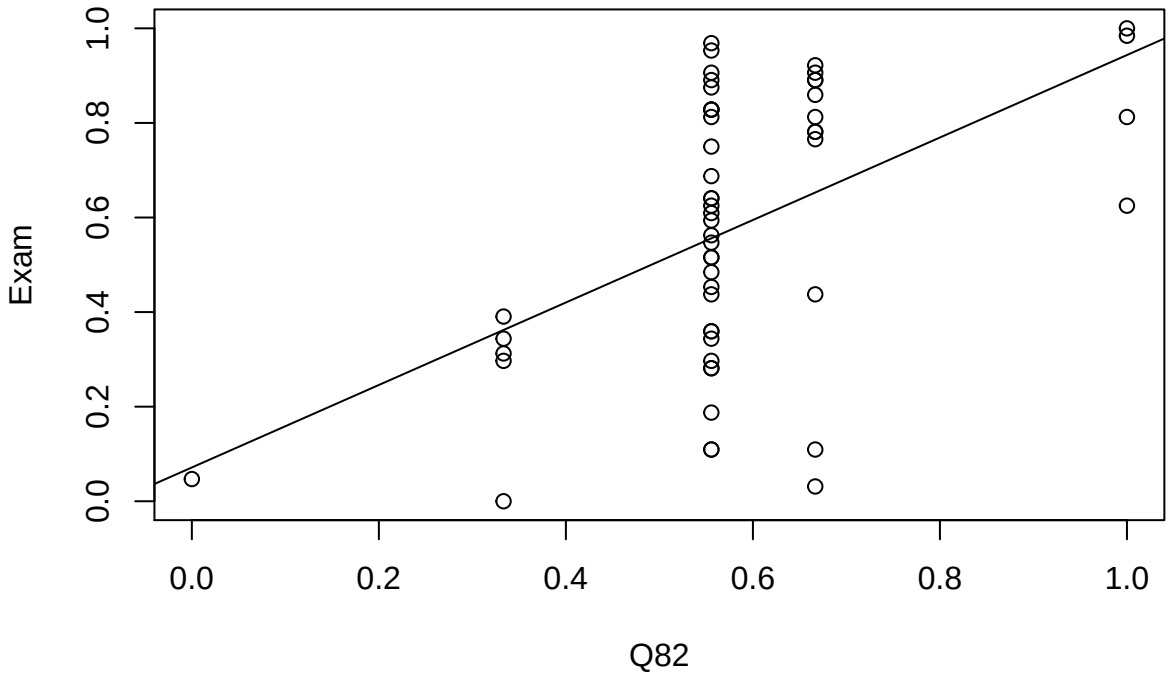

res
0.4
0.2
0.0
−0.2
−0.4
0.0
0.2
0.4
0.6
0.8
1.0
dfComplete$PeerSubmissionScore

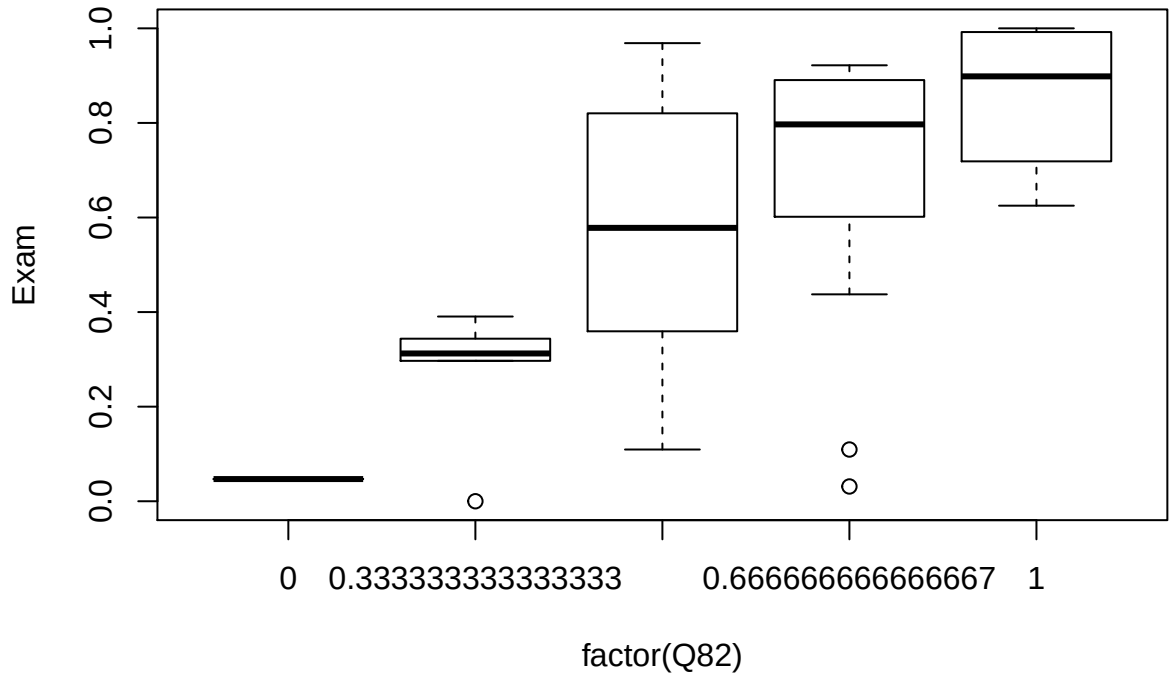

res
0.4
0.2
0.0
−0.2
−0.4
0.0
0.2
0.4
0.6
0.8
1.0
dfComplete$Q6


res
0.4
0.2
0.0
−0.2
−0.4
0.0
0.2
0.4
0.6
0.8
1.0
dfComplete$Q56

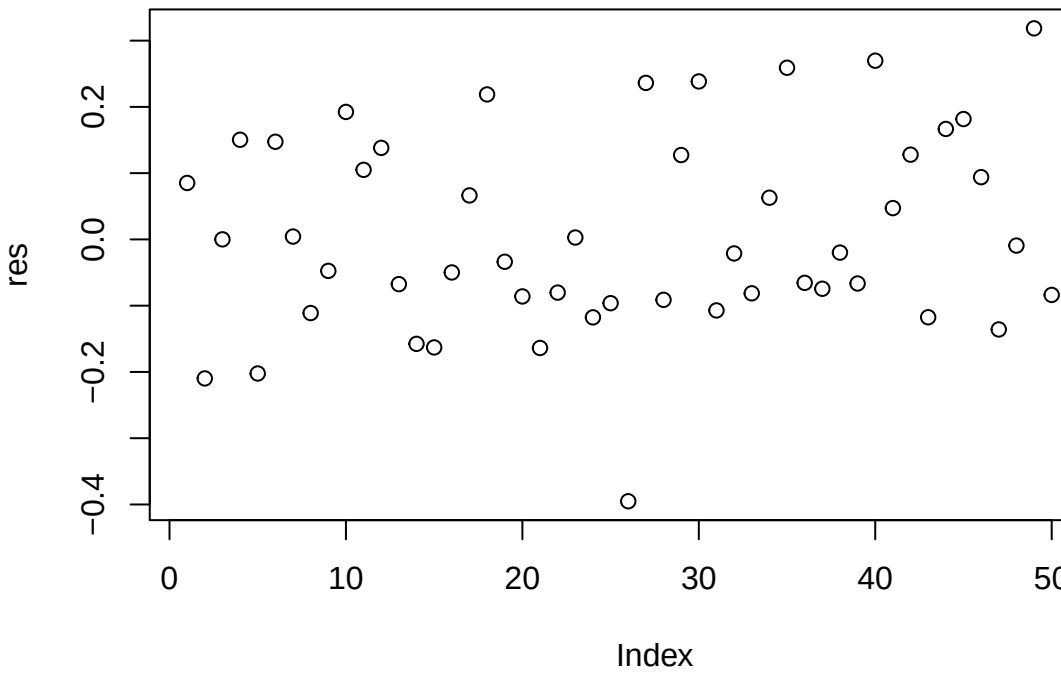

res
0.4
0.2
0.0
−0.2
−0.4
0.0
0.2
0.4
0.6
0.8
1.0
dfComplete$Q82


res
0.4
0.2
0.0
−0.2
−0.4
0.0
0.2
0.4
0.6
0.8
1.0
dfComplete$Q98

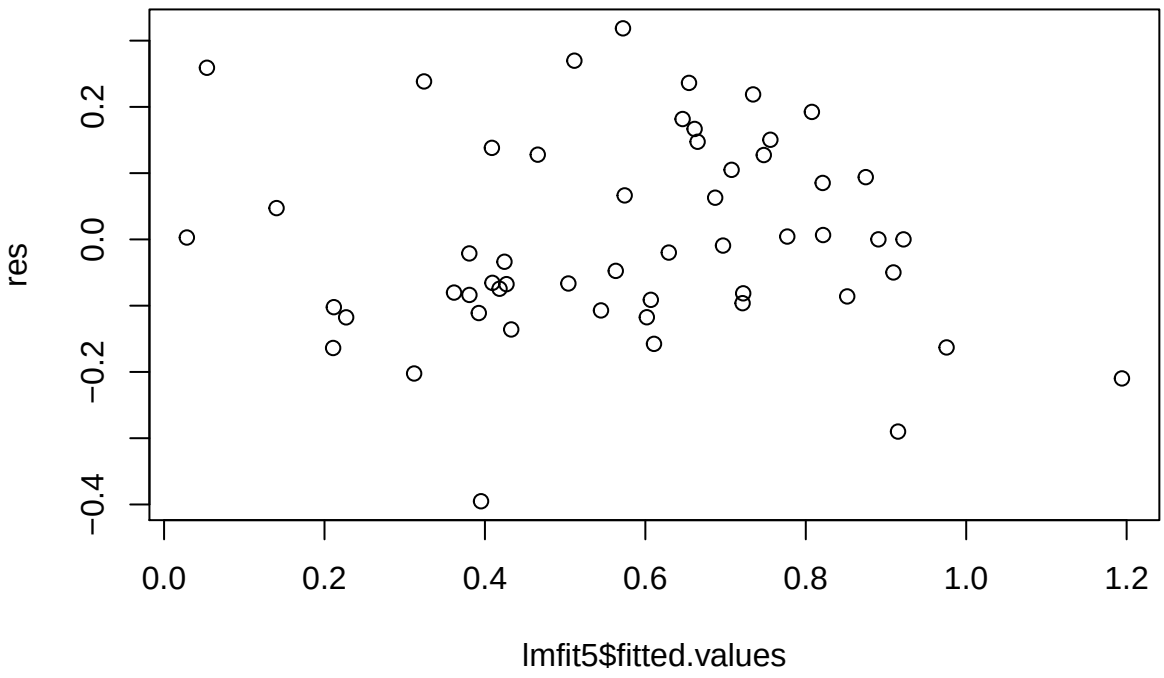

Exam
0.0
0.2
0.4
0.6
0.8
1.0
Q82
0.0
0.2
0.4
0.6
0.8
1.0

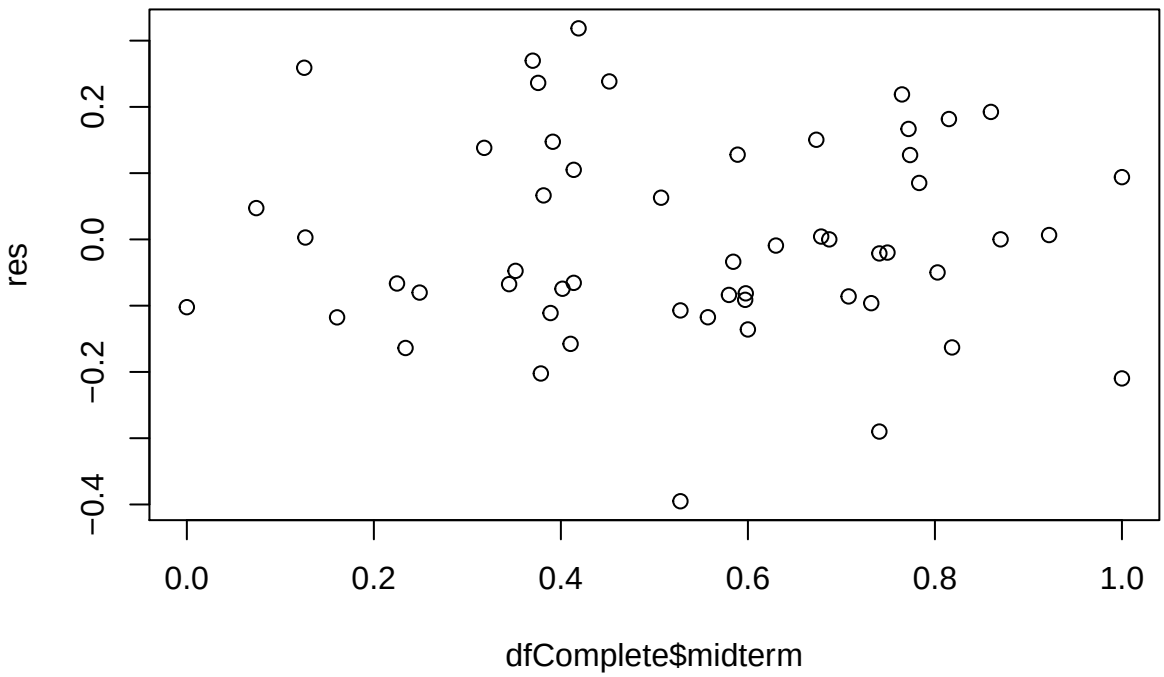


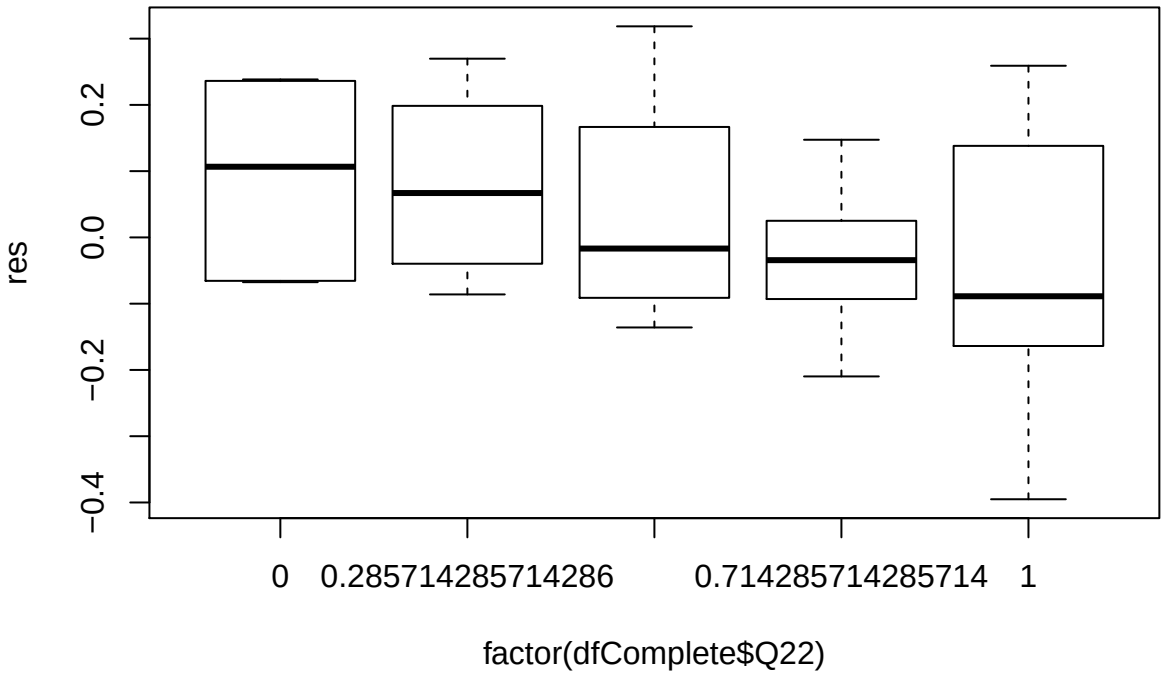

```
##
## Call:
## lm(formula = Exam ~ factor(Q82), data = dfComplete)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65104 -0.19238  0.04873  0.17830  0.39746
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   0.04687    0.25558   0.183  0.85524
## factor(Q82)0.333333333333333  0.22188    0.27997   0.792  0.43190
## factor(Q82)0.555555555555556  0.52441    0.25954   2.021  0.04881 *
## factor(Q82)0.666666666666667  0.63542    0.26601   2.389  0.02081 *
## factor(Q82)1                  0.80859    0.28575   2.830  0.00673 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2556 on 49 degrees of freedom
## Multiple R-squared:  0.2727, Adjusted R-squared:  0.2133
## F-statistic: 4.593 on 4 and 49 DF,  p-value: 0.003143
```

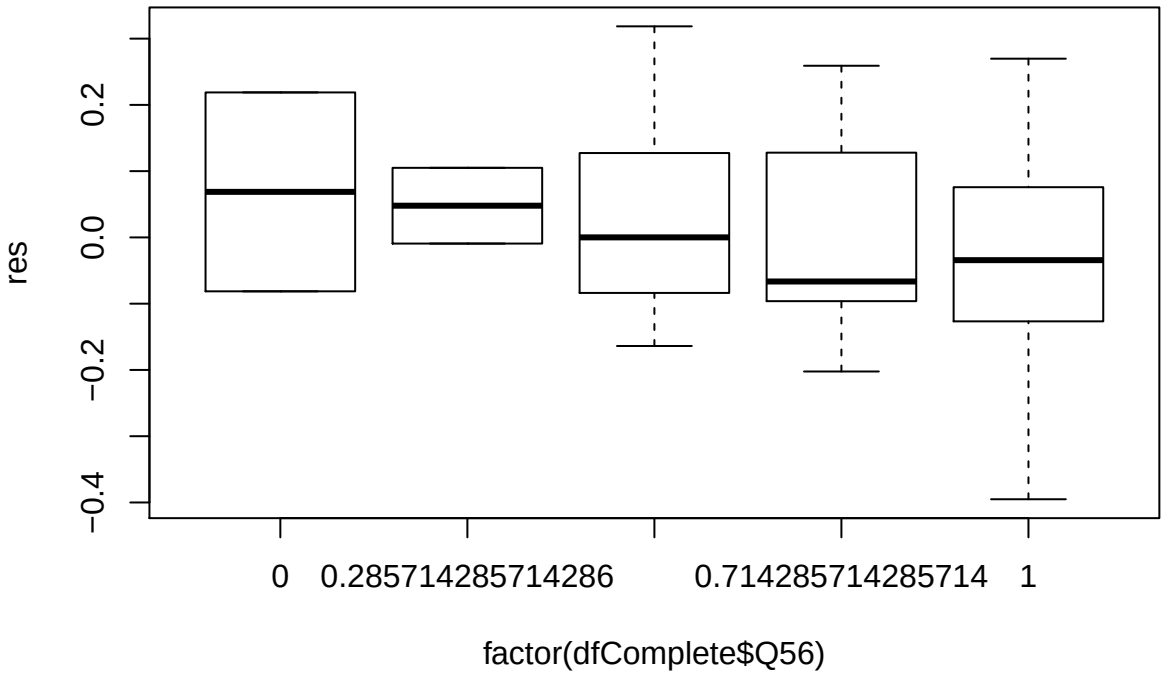


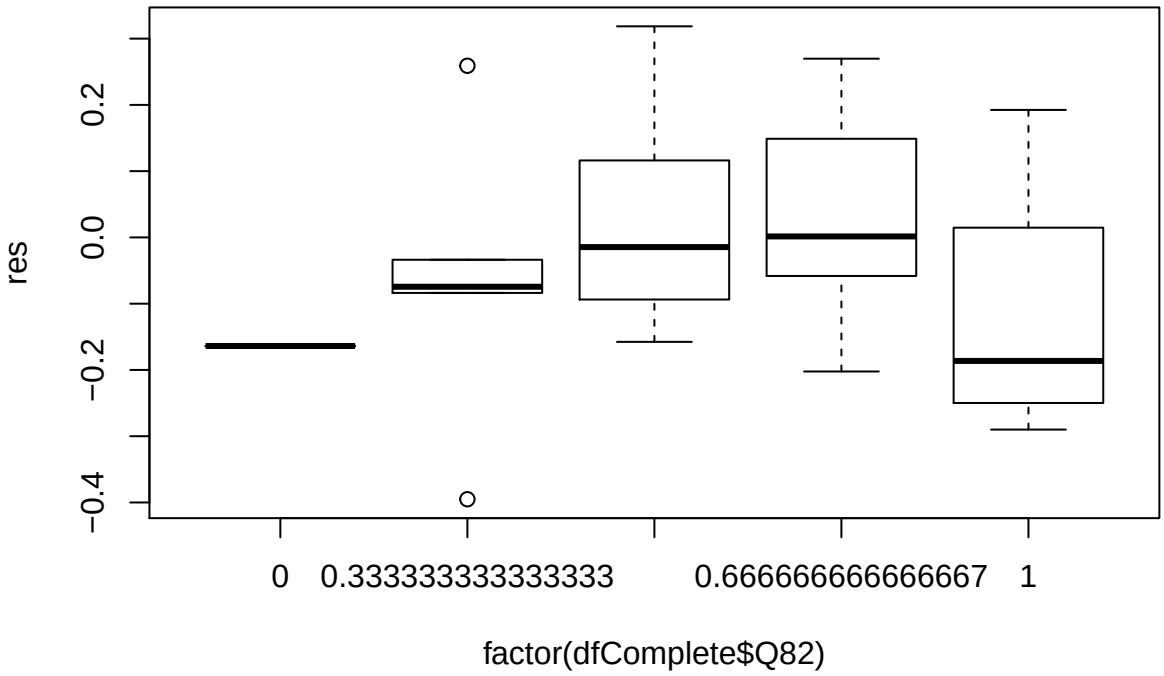
Checking residuals for one model

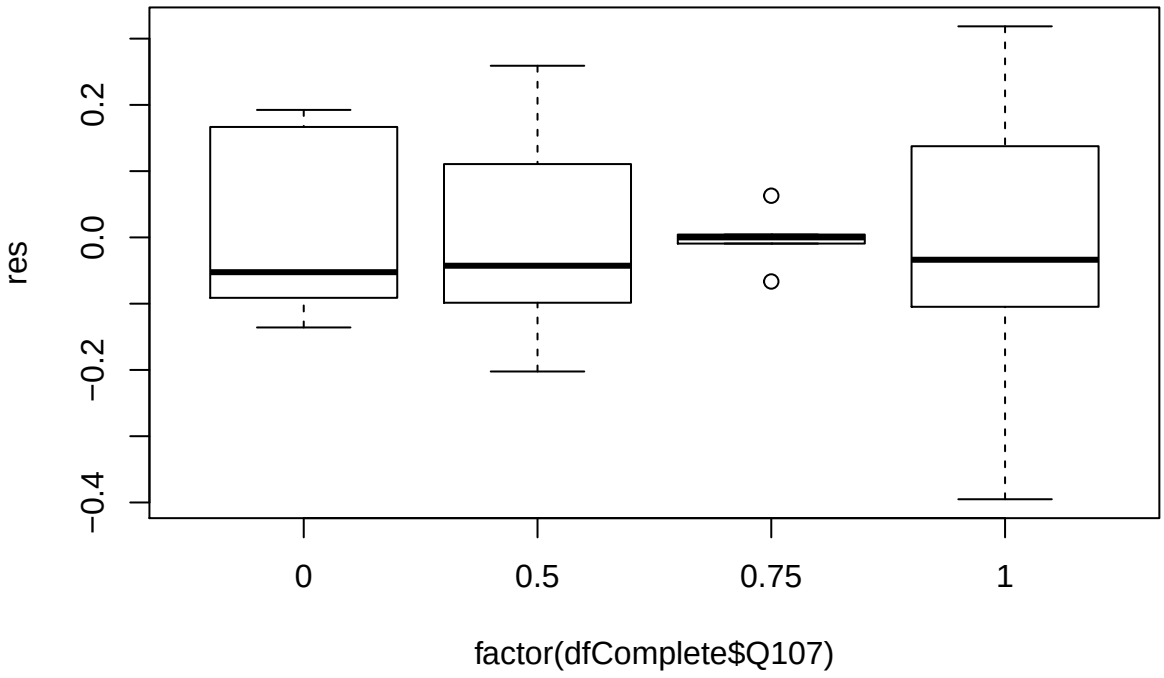

res
0.2
0.0
−0.2
−0.4
0.0
0.2
0.4
0.6
0.8
1.0
1.2
lmfit5$fitted.values

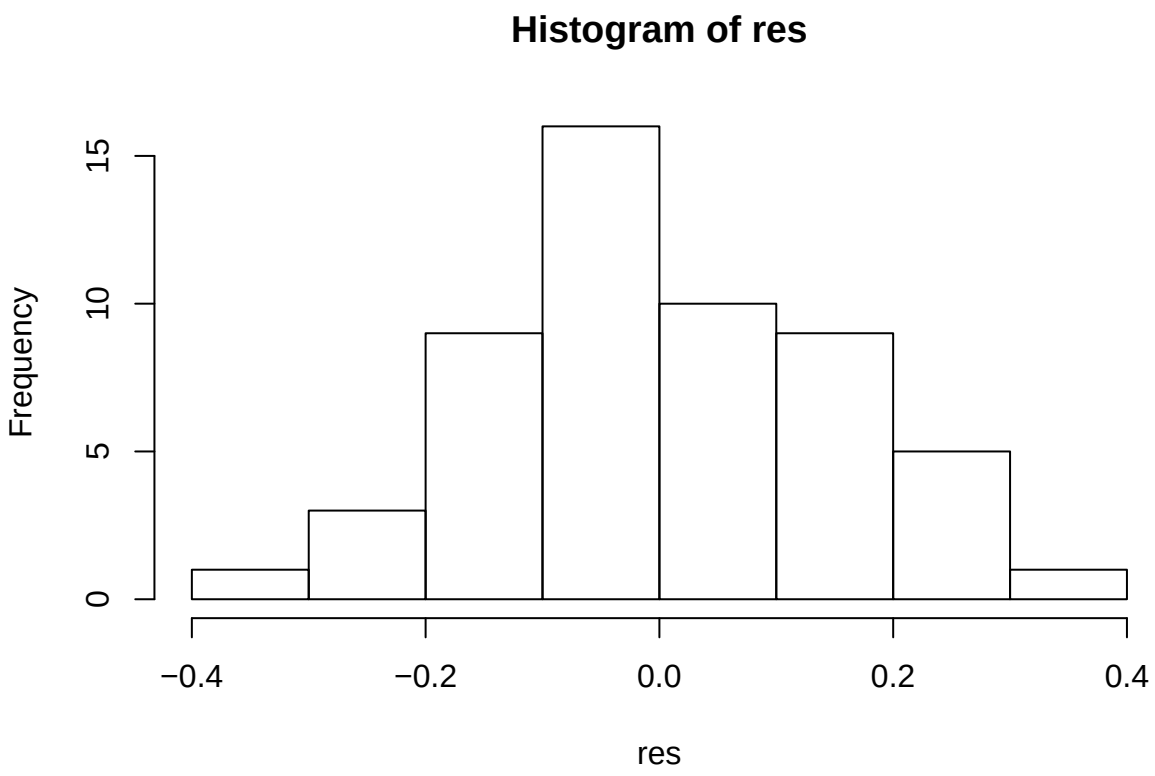

res
0.2
0.0
−0.2
−0.4
0.0
0.2
0.4
0.6
0.8
1.0
dfComplete$midterm

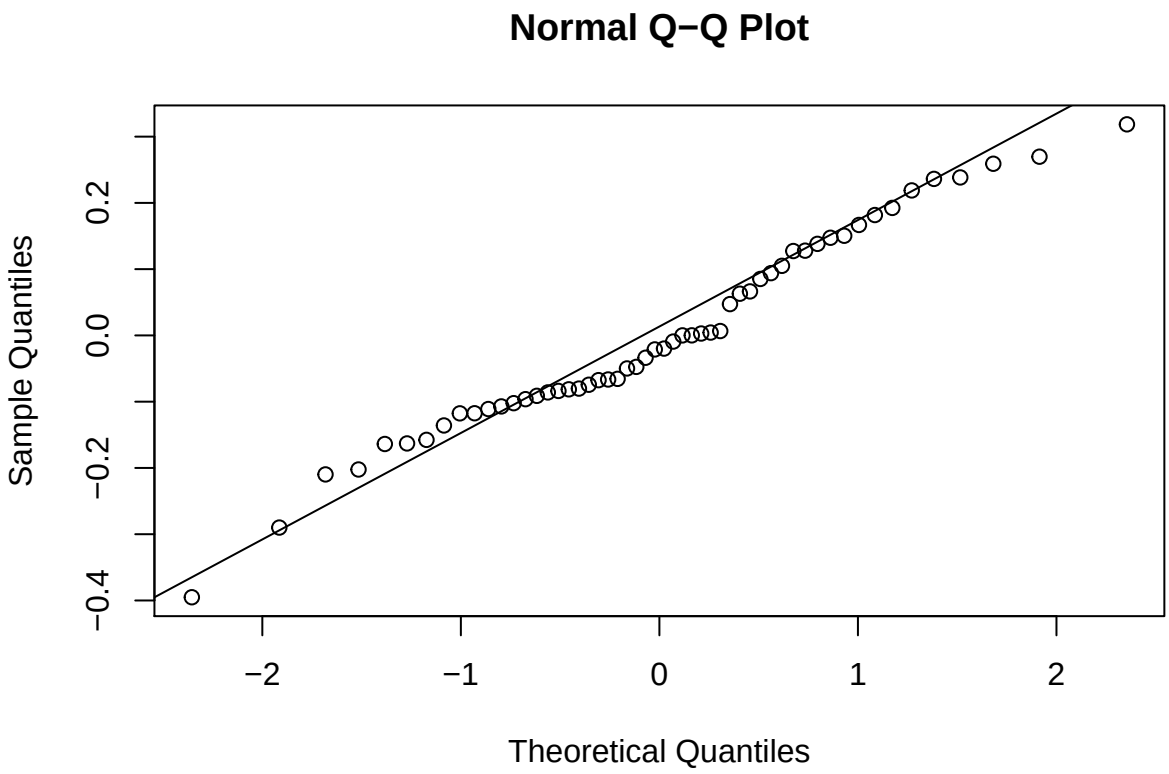

res
0.2
0.0
−0.2
−0.4
0
0.285714285714286
0.714285714285714
1
factor(dfComplete$Q22)


res
0.2
0.0
−0.2
−0.4
0
0.285714285714286
0.714285714285714
1
factor(dfComplete$Q56)


res
0.2
0.0
−0.2
−0.4
0
0.333333333333333
0.666666666666667
1
factor(dfComplete$Q82)


res
0.2
0.0
−0.2
−0.4
0
0.5
0.75
1
factor(dfComplete$Q107)


Histogram of res
Frequency
0
5
10
15
−0.4
−0.2
0.0
0.2
0.4
res



Predicting dropouts

Trying to predict dropout based on exam results. Not many who have taken the exam has dropped out yet:

```
##
##  active dropout    Q999
##      65       2      10
```

```
##
## Call:
## glm(formula = dropout ~ X1_T_mGPA, family = binomial, data = MedData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9528  -1.0618   0.5672   0.9273   2.0954
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.62142    0.08020   7.749  9.3e-15 ***
## X1_T_mGPA   -0.22490    0.01967 -11.435  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1256.4  on 914  degrees of freedom
```

```
## Residual deviance: 1089.0  on 913  degrees of freedom
##   (92 observations deleted due to missingness)
## AIC: 1093
##
## Number of Fisher Scoring iterations: 4
```

```
## % latex table generated in R 3.5.1 by xtable 1.8-3 package
## % Tue Dec 11 22:58:21 2018
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrrr}
##   \hline
##  & active & dropout & Q999 \\
##   \hline
## 0 &  40 &   0 &   1 \\
##   1 &  25 &   2 &   9 \\
##    \hline
## \end{tabular}
## \end{table}
```

ROC, confusion matrix, and cost function

Analyses the best model without MT scores (i.e. lmfittotal5).

Prepare the data:

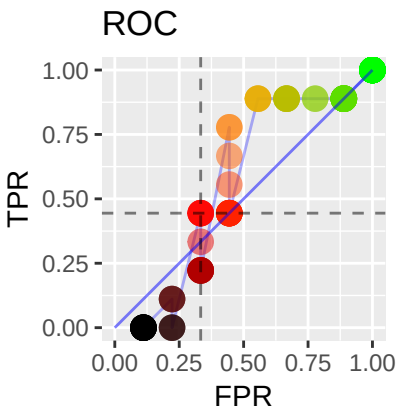
Confusion matrix:

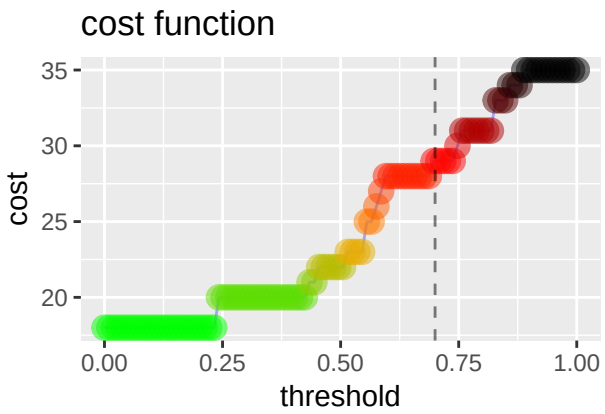
```
##     obs
## pred 0 1
##    0 5 6
##    1 4 3
## attr(,"class")
## [1] "confusion.matrix"
```

Distribution of the predictions

Calculate ROC and cost function

Plot ROC and cost function





'eshold at 0.70 − cost of FP = 1, cost of FN = 2

Area under the curve

```
## [1] 0.5555556
```