Final Project Modern Statistical Computing

Martin Blasi and Jonas Wallstein

2023-03-15

Setup

```r
mat <- read.csv2('../data/student-mat.csv')
por <- read.csv2('../data/student-por.csv')
```

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0     v purrr   1.0.1
## v tibble  3.1.8     v dplyr   1.1.0
## v tidyr   1.2.1     v stringr 1.5.0
## v readr   2.1.3     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(boot)
library(coefplot)
library(modelr)
library(openintro)
```

```
## Loading required package: airports
## Loading required package: cherryblossom
## Loading required package: usdata
##
## Attaching package: 'openintro'
##
## The following object is masked from 'package:boot':
##
##     salinity
```

```r
library(brglm)
```

```
## Loading required package: profileModel
## 'brglm' will gradually be superseded by the 'brglm2' R package (https://cran.r-project.org/package=b
##  Methods for the detection of separation and infinite estimates in binomial-response models are prov
```

```r
library(mombf)
```

```
## Loading required package: mvtnorm
## Loading required package: ncvreg
## Loading required package: mgcv
## Loading required package: nlme
##
## Attaching package: 'nlme'
```

```
##
## The following object is masked from 'package:dplyr':
##
##     collapse
##
## This is mgcv 1.8-40. For overview type 'help("mgcv-package")'.
```

```
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
##
## Attaching package: 'pROC'
##
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(keras)
library(mlbench)
library(mgcv)
library(ggpubr)
library(huxtable)
```

```
##
## Attaching package: 'huxtable'
##
## The following object is masked from 'package:ggpubr':
##
##     font
##
## The following object is masked from 'package:dplyr':
##
##     add_rownames
##
## The following object is masked from 'package:ggplot2':
##
##     theme_grey
```

```
library(jtools)
```

```
##
## Attaching package: 'jtools'
##
## The following object is masked from 'package:keras':
##
##     get_weights
```

```
source('routines.R')
```

1. Explaining

1.1 Linear Regression

```
fitall <- lm(G3~.,data=mat)
summary(fitall)
```

Full linear model

```
##
## Call:
## lm(formula = G3 ~ ., data = mat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.9339 -0.5532  0.2680  0.9689  4.6461
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.115488   2.116958  -0.527 0.598573
## schoolMS          0.480742   0.366512   1.312 0.190485
## sexM              0.174396   0.233588   0.747 0.455805
## age              -0.173302   0.100780  -1.720 0.086380 .
## addressU          0.104455   0.270791   0.386 0.699922
## famsizeLE3        0.036512   0.226680   0.161 0.872128
## PstatusT         -0.127673   0.335626  -0.380 0.703875
## Medu              0.129685   0.149999   0.865 0.387859
## Fedu             -0.133940   0.128768  -1.040 0.298974
## Mjobhealth       -0.146426   0.518491  -0.282 0.777796
## Mjobother         0.074088   0.332044   0.223 0.823565
## Mjobservices      0.046956   0.369587   0.127 0.898973
## Mjobteacher      -0.026276   0.481632  -0.055 0.956522
## Fjobhealth        0.330948   0.666601   0.496 0.619871
## Fjobother        -0.083582   0.476796  -0.175 0.860945
## Fjobservices     -0.322142   0.493265  -0.653 0.514130
## Fjobteacher      -0.112364   0.601448  -0.187 0.851907
## reasonhome       -0.209183   0.256392  -0.816 0.415123
## reasonother       0.307554   0.380214   0.809 0.419120
## reasonreputation  0.129106   0.267254   0.483 0.629335
## guardianmother    0.195741   0.252672   0.775 0.439046
## guardianother     0.006565   0.463650   0.014 0.988710
## traveltime        0.096994   0.157800   0.615 0.539170
## studytime        -0.104754   0.134814  -0.777 0.437667
## failures         -0.160539   0.161006  -0.997 0.319399
## schoolsupyes      0.456448   0.319538   1.428 0.154043
## famsupyes         0.176870   0.224204   0.789 0.430710
## paidyes           0.075764   0.222100   0.341 0.733211
## activitiesyes    -0.346047   0.205938  -1.680 0.093774 .
## nurseryyes       -0.222716   0.254184  -0.876 0.381518
## higheryes         0.225921   0.500398   0.451 0.651919
## internetyes      -0.144462   0.287528  -0.502 0.615679
## romanticyes      -0.272008   0.219732  -1.238 0.216572
## famrel            0.356876   0.114124   3.127 0.001912 **
## freetime          0.047002   0.110209   0.426 0.670021
## goout             0.012007   0.105230   0.114 0.909224
## Dalc             -0.185019   0.153124  -1.208 0.227741
## Walc              0.176772   0.114943   1.538 0.124966
## health            0.062995   0.074800   0.842 0.400259
## absences          0.045879   0.013412   3.421 0.000698 ***
## G1                0.188847   0.062373   3.028 0.002645 **
## G2                0.957330   0.053460  17.907  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.901 on 353 degrees of freedom
## Multiple R-squared:  0.8458, Adjusted R-squared:  0.8279
## F-statistic: 47.21 on 41 and 353 DF,  p-value: < 2.2e-16
```

```
# bestBIC(G3~., data=mat)
# fit1 <- lm(G3~age +famrel+ absences+ G1+ G2,data=mat)
# summary(fit1)
```

```
dfnog <- dplyr::select(mat, -c("G1", "G2"))
fitallnog <- lm(G3~.,data=dfnog)
summary(fitallnog)
```

Full linear model without grades

```
##
## Call:
## lm(formula = G3 ~ ., data = dfnog)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.0442  -1.9028   0.4289   2.7570   8.8874
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      14.07769    4.48089   3.142  0.00182 **
## schoolMS          0.72555    0.79157   0.917  0.35997
## sexM              1.26236    0.50003   2.525  0.01202 *
## age              -0.37516    0.21721  -1.727  0.08501 .
## addressU          0.55135    0.58412   0.944  0.34586
## famsizeLE3        0.70281    0.48824   1.439  0.15090
## PstatusT         -0.32010    0.72390  -0.442  0.65862
## Medu              0.45687    0.32317   1.414  0.15833
## Fedu             -0.10458    0.27762  -0.377  0.70663
## Mjobhealth        0.99808    1.11819   0.893  0.37268
## Mjobother        -0.35900    0.71316  -0.503  0.61500
## Mjobservices      0.65832    0.79784   0.825  0.40985
## Mjobteacher      -1.24149    1.03821  -1.196  0.23257
## Fjobhealth        0.34767    1.43796   0.242  0.80909
## Fjobother        -0.61967    1.02304  -0.606  0.54509
## Fjobservices     -0.46577    1.05697  -0.441  0.65972
## Fjobteacher       1.32619    1.29654   1.023  0.30707
## reasonhome        0.07851    0.55380   0.142  0.88735
## reasonother       0.77707    0.81757   0.950  0.34252
## reasonreputation  0.61304    0.57657   1.063  0.28839
## guardianmother    0.06978    0.54560   0.128  0.89830
## guardianother     0.75010    0.99946   0.751  0.45345
## traveltime       -0.24027    0.33897  -0.709  0.47889
## studytime         0.54952    0.28765   1.910  0.05690 .
## failures         -1.72398    0.33291  -5.179 3.75e-07 ***
## schoolsupyes     -1.35058    0.66693  -2.025  0.04361 *
## famsupyes        -0.86182    0.47869  -1.800  0.07265 .
```

```
## paidyes           0.33975    0.47775   0.711  0.47746
## activitiesyes    -0.32953    0.44494  -0.741  0.45942
## nurseryyes       -0.17730    0.54931  -0.323  0.74706
## higheryes         1.37045    1.07780   1.272  0.20437
## internetyes       0.49813    0.61956   0.804  0.42192
## romanticyes      -1.09449    0.46925  -2.332  0.02024 *
## famrel            0.23155    0.24593   0.942  0.34706
## freetime          0.30242    0.23735   1.274  0.20345
## goout            -0.59367    0.22451  -2.644  0.00855 **
## Dalc             -0.27223    0.33087  -0.823  0.41120
## Walc              0.26339    0.24801   1.062  0.28896
## health           -0.17678    0.16101  -1.098  0.27297
## absences          0.05629    0.02897   1.943  0.05277 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.108 on 355 degrees of freedom
## Multiple R-squared:  0.2756, Adjusted R-squared:  0.196
## F-statistic: 3.463 on 39 and 355 DF,  p-value: 3.317e-10
```

```
#bestBIC(G3~., data=dfnog)
```

```
export_summs(fitall, fitallnog, digits=4, error_format = "({p.value})", model.names = c("Full Linear mod
```

Test table

1.2 Poisson Regression

```
fitallp <- glm(G3~.,data=mat, family=poisson())
summary(fitallp)
```

Full Poisson

```
##
## Call:
## glm(formula = G3 ~ ., family = poisson(), data = mat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.0133  -0.2120   0.1333   0.4646   1.5971
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       0.7342471  0.3564613   2.060 0.039415 *
## schoolMS          0.0685549  0.0610207   1.123 0.261239
## sexM              0.0244299  0.0377134   0.648 0.517128
## age               0.0002861  0.0169339   0.017 0.986520
## addressU          0.0266205  0.0454070   0.586 0.557697
## famsizeLE3        0.0050209  0.0365608   0.137 0.890770
## PstatusT          0.0089554  0.0534351   0.168 0.866903
## Medu             -0.0103479  0.0245054  -0.422 0.672828
## Fedu             -0.0067971  0.0207199  -0.328 0.742877
## Mjobhealth        0.0129923  0.0857216   0.152 0.879531
```

```
## Mjobother        -0.0051620  0.0585527  -0.088 0.929750
## Mjobservices     -0.0134616  0.0634765  -0.212 0.832051
## Mjobteacher       0.0328367  0.0820120   0.400 0.688870
## Fjobhealth        0.0443579  0.1088357   0.408 0.683591
## Fjobother        -0.0014016  0.0793579  -0.018 0.985909
## Fjobservices     -0.0023863  0.0825201  -0.029 0.976930
## Fjobteacher      -0.0299705  0.0978826  -0.306 0.759462
## reasonhome       -0.0154345  0.0429498  -0.359 0.719325
## reasonother       0.0236210  0.0610127   0.387 0.698646
## reasonreputation  0.0304291  0.0443217   0.687 0.492365
## guardianmother    0.0090397  0.0407455   0.222 0.824426
## guardianother    -0.0072947  0.0807424  -0.090 0.928012
## traveltime        0.0115322  0.0271297   0.425 0.670781
## studytime         0.0042116  0.0215942   0.195 0.845368
## failures         -0.0594046  0.0325244  -1.826 0.067780 .
## schoolsupyes      0.1021009  0.0545400   1.872 0.061201 .
## famsupyes         0.0020079  0.0369971   0.054 0.956718
## paidyes           0.0379493  0.0357939   1.060 0.289046
## activitiesyes    -0.0328670  0.0340127  -0.966 0.333885
## nurseryyes       -0.0254084  0.0427143  -0.595 0.551947
## higheryes         0.0695943  0.0981118   0.709 0.478115
## internetyes      -0.0517383  0.0490178  -1.055 0.291197
## romanticyes      -0.0107542  0.0367615  -0.293 0.769874
## famrel            0.0354547  0.0188780   1.878 0.060367 .
## freetime          0.0111734  0.0178293   0.627 0.530863
## goout            -0.0114724  0.0177116  -0.648 0.517160
## Dalc             -0.0110207  0.0255432  -0.431 0.666138
## Walc              0.0276017  0.0190919   1.446 0.148253
## health            0.0059326  0.0121156   0.490 0.624369
## absences          0.0078130  0.0021837   3.578 0.000346 ***
## G1               -0.0244982  0.0119561  -2.049 0.040460 *
## G2                0.1386421  0.0115365  12.018  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1159.13  on 394  degrees of freedom
## Residual deviance:  432.79  on 353  degrees of freedom
## AIC: 2036.2
##
## Number of Fisher Scoring iterations: 5
```

```r
#bestBIC(G3~.,data=mat, family="poisson")
```

```r
fitallpnog <- glm(G3~.,data=dfnog,family="poisson")
summary(fitallpnog)
```

Full Poisson without grades

```
##
## Call:
## glm(formula = G3 ~ ., family = "poisson", data = dfnog)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.2017  -0.6254   0.1057   0.8412   2.8588
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       2.633626   0.343177   7.674 1.66e-14 ***
## schoolMS          0.081723   0.060917   1.342 0.179740
## sexM              0.115098   0.037151   3.098 0.001948 **
## age              -0.035890   0.016520  -2.173 0.029815 *
## addressU          0.052338   0.045236   1.157 0.247279
## famsizeLE3        0.061290   0.036201   1.693 0.090452 .
## PstatusT         -0.026958   0.053287  -0.506 0.612921
## Medu              0.042600   0.024281   1.754 0.079349 .
## Fedu             -0.008111   0.020788  -0.390 0.696419
## Mjobhealth        0.089444   0.083963   1.065 0.286747
## Mjobother        -0.035508   0.057003  -0.623 0.533342
## Mjobservices      0.070618   0.062343   1.133 0.257325
## Mjobteacher      -0.113275   0.079283  -1.429 0.153082
## Fjobhealth        0.031155   0.107490   0.290 0.771937
## Fjobother        -0.060356   0.078606  -0.768 0.442592
## Fjobservices     -0.043903   0.081066  -0.542 0.588111
## Fjobteacher       0.115347   0.096281   1.198 0.230908
## reasonhome        0.005817   0.042655   0.136 0.891529
## reasonother       0.074529   0.060625   1.229 0.218939
## reasonreputation  0.053021   0.043437   1.221 0.222223
## guardianmother   -0.004933   0.040613  -0.121 0.903322
## guardianother     0.088615   0.078290   1.132 0.257683
## traveltime       -0.025305   0.026910  -0.940 0.347028
## studytime         0.052688   0.021249   2.480 0.013155 *
## failures         -0.224337   0.030872  -7.267 3.69e-13 ***
## schoolsupyes     -0.129216   0.052220  -2.474 0.013343 *
## famsupyes        -0.088907   0.036056  -2.466 0.013671 *
## paidyes           0.035369   0.035545   0.995 0.319721
## activitiesyes    -0.036272   0.033729  -1.075 0.282193
## nurseryyes       -0.010361   0.042202  -0.245 0.806070
## higheryes         0.195120   0.095918   2.034 0.041927 *
## internetyes       0.045241   0.048849   0.926 0.354372
## romanticyes      -0.106043   0.036371  -2.916 0.003550 **
## famrel            0.019952   0.018668   1.069 0.285154
## freetime          0.030301   0.017812   1.701 0.088913 .
## goout            -0.058131   0.017095  -3.400 0.000673 ***
## Dalc             -0.020912   0.025717  -0.813 0.416117
## Walc              0.023945   0.019125   1.252 0.210566
## health           -0.017740   0.012076  -1.469 0.141821
## absences          0.006355   0.002153   2.952 0.003158 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1159.13  on 394  degrees of freedom
## Residual deviance:  921.44  on 355  degrees of freedom
## AIC: 2520.9
```

```
##
## Number of Fisher Scoring iterations: 5
```

```
fitallqp <- glm(G3~.,data=mat,family="quasipoisson")
summary(fitallqp)
```

Quasipoisson - adjusting for overdispersion

```
##
## Call:
## glm(formula = G3 ~ ., family = "quasipoisson", data = mat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.0133  -0.2120   0.1333   0.4646   1.5971
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.7342471  0.3121591   2.352   0.0192 *
## schoolMS          0.0685549  0.0534369   1.283   0.2004
## sexM              0.0244299  0.0330262   0.740   0.4600
## age               0.0002861  0.0148293   0.019   0.9846
## addressU          0.0266205  0.0397636   0.669   0.5036
## famsizeLE3        0.0050209  0.0320169   0.157   0.8755
## PstatusT          0.0089554  0.0467940   0.191   0.8483
## Medu             -0.0103479  0.0214598  -0.482   0.6300
## Fedu             -0.0067971  0.0181447  -0.375   0.7082
## Mjobhealth        0.0129923  0.0750678   0.173   0.8627
## Mjobother        -0.0051620  0.0512756  -0.101   0.9199
## Mjobservices     -0.0134616  0.0555875  -0.242   0.8088
## Mjobteacher       0.0328367  0.0718192   0.457   0.6478
## Fjobhealth        0.0443579  0.0953092   0.465   0.6419
## Fjobother        -0.0014016  0.0694950  -0.020   0.9839
## Fjobservices     -0.0023863  0.0722643  -0.033   0.9737
## Fjobteacher      -0.0299705  0.0857174  -0.350   0.7268
## reasonhome       -0.0154345  0.0376119  -0.410   0.6818
## reasonother       0.0236210  0.0534298   0.442   0.6587
## reasonreputation  0.0304291  0.0388132   0.784   0.4336
## guardianmother    0.0090397  0.0356816   0.253   0.8002
## guardianother    -0.0072947  0.0707075  -0.103   0.9179
## traveltime        0.0115322  0.0237579   0.485   0.6277
## studytime         0.0042116  0.0189104   0.223   0.8239
## failures         -0.0594046  0.0284821  -2.086   0.0377 *
## schoolsupyes      0.1021009  0.0477616   2.138   0.0332 *
## famsupyes         0.0020079  0.0323990   0.062   0.9506
## paidyes           0.0379493  0.0313453   1.211   0.2268
## activitiesyes    -0.0328670  0.0297854  -1.103   0.2706
## nurseryyes       -0.0254084  0.0374056  -0.679   0.4974
## higheryes         0.0695943  0.0859182   0.810   0.4185
## internetyes      -0.0517383  0.0429257  -1.205   0.2289
## romanticyes      -0.0107542  0.0321926  -0.334   0.7385
## famrel            0.0354547  0.0165318   2.145   0.0327 *
## freetime          0.0111734  0.0156134   0.716   0.4747
## goout            -0.0114724  0.0155104  -0.740   0.4600
```

```
## Dalc              -0.0110207  0.0223686  -0.493   0.6225
## Walc               0.0276017  0.0167191   1.651   0.0996 .
## health             0.0059326  0.0106098   0.559   0.5764
## absences           0.0078130  0.0019123   4.086 5.44e-05 ***
## G1                -0.0244982  0.0104701  -2.340   0.0198 *
## G2                 0.1386421  0.0101027  13.723  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 0.7668797)
##
##     Null deviance: 1159.13  on 394  degrees of freedom
## Residual deviance:  432.79  on 353  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

```r
fitallqpnog <- glm(G3~.,data=dfnog,family="quasipoisson")
summary(fitallqpnog)
```

Quasipoisson without grades

```
##
## Call:
## glm(formula = G3 ~ ., family = "quasipoisson", data = dfnog)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.2017  -0.6254   0.1057   0.8412   2.8588
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.633626   0.463501   5.682 2.78e-08 ***
## schoolMS           0.081723   0.082275   0.993   0.3212
## sexM               0.115098   0.050177   2.294   0.0224 *
## age               -0.035890   0.022312  -1.609   0.1086
## addressU           0.052338   0.061097   0.857   0.3922
## famsizeLE3         0.061290   0.048894   1.254   0.2108
## PstatusT          -0.026958   0.071970  -0.375   0.7082
## Medu               0.042600   0.032794   1.299   0.1948
## Fedu              -0.008111   0.028077  -0.289   0.7728
## Mjobhealth         0.089444   0.113401   0.789   0.4308
## Mjobother         -0.035508   0.076990  -0.461   0.6449
## Mjobservices       0.070618   0.084201   0.839   0.4022
## Mjobteacher       -0.113275   0.107081  -1.058   0.2908
## Fjobhealth         0.031155   0.145177   0.215   0.8302
## Fjobother         -0.060356   0.106167  -0.568   0.5701
## Fjobservices      -0.043903   0.109489  -0.401   0.6887
## Fjobteacher        0.115347   0.130039   0.887   0.3757
## reasonhome         0.005817   0.057610   0.101   0.9196
## reasonother        0.074529   0.081881   0.910   0.3633
## reasonreputation   0.053021   0.058667   0.904   0.3667
## guardianmother    -0.004933   0.054853  -0.090   0.9284
## guardianother      0.088615   0.105739   0.838   0.4026
```

```
## traveltime       -0.025305   0.036345  -0.696   0.4867
## studytime         0.052688   0.028699   1.836   0.0672 .
## failures         -0.224337   0.041697  -5.380 1.35e-07 ***
## schoolsupyes     -0.129216   0.070529  -1.832   0.0678 .
## famsupyes        -0.088907   0.048698  -1.826   0.0687 .
## paidyes           0.035369   0.048008   0.737   0.4618
## activitiesyes    -0.036272   0.045555  -0.796   0.4264
## nurseryyes       -0.010361   0.056999  -0.182   0.8559
## higheryes         0.195120   0.129548   1.506   0.1329
## internetyes       0.045241   0.065976   0.686   0.4933
## romanticyes      -0.106043   0.049123  -2.159   0.0315 *
## famrel            0.019952   0.025213   0.791   0.4293
## freetime          0.030301   0.024057   1.260   0.2087
## goout            -0.058131   0.023089  -2.518   0.0123 *
## Dalc             -0.020912   0.034733  -0.602   0.5475
## Walc              0.023945   0.025831   0.927   0.3546
## health           -0.017740   0.016310  -1.088   0.2775
## absences          0.006355   0.002907   2.186   0.0295 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 1.824162)
##
##     Null deviance: 1159.13  on 394  degrees of freedom
## Residual deviance:  921.44  on 355  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

1.3 Binomial Regression

```
df <- mat
df$pass <- ifelse(df$G3>9, 1 ,0)
dfbin <- select(df, -c("G3"))
fitallb <- glm(pass~.,data=dfbin,family=binomial())
```

Pass/Fail Model

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(fitallb)
```

```
##
## Call:
## glm(formula = pass ~ ., family = binomial(), data = dfbin)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -3.327   0.000   0.000   0.000   2.954
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -28.42647   19.41993  -1.464   0.1433
## schoolMS          11.90561    5.33871   2.230   0.0257 *
## sexM              -3.96013    1.93631  -2.045   0.0408 *
```

```
## age              -3.24533    1.30419  -2.488   0.0128 *
## addressU          0.31464    1.90655   0.165   0.8689
## famsizeLE3       -6.93521    3.42095  -2.027   0.0426 *
## PstatusT         -1.55774    2.93266  -0.531   0.5953
## Medu              0.22905    1.23596   0.185   0.8530
## Fedu             -3.33163    1.73562  -1.920   0.0549 .
## Mjobhealth       -0.31256    4.00462  -0.078   0.9378
## Mjobother        -6.49684    2.94351  -2.207   0.0273 *
## Mjobservices     -1.12982    2.54999  -0.443   0.6577
## Mjobteacher      -4.43407    3.11305  -1.424   0.1543
## Fjobhealth        3.27050    5.35000   0.611   0.5410
## Fjobother         8.58553    3.78745   2.267   0.0234 *
## Fjobservices     -0.48006    2.77774  -0.173   0.8628
## Fjobteacher      17.09109    9.35060   1.828   0.0676 .
## reasonhome        4.54926    2.69789   1.686   0.0918 .
## reasonother      -3.49872    5.34202  -0.655   0.5125
## reasonreputation  0.11575    1.86001   0.062   0.9504
## guardianmother   -1.89593    1.84627  -1.027   0.3045
## guardianother    -7.74892    4.65347  -1.665   0.0959 .
## traveltime       -1.27683    1.13328  -1.127   0.2599
## studytime        -3.36944    1.36916  -2.461   0.0139 *
## failures          0.69418    0.82573   0.841   0.4005
## schoolsupyes     -0.20720    1.91005  -0.108   0.9136
## famsupyes        -0.76439    1.44360  -0.530   0.5965
## paidyes           1.10801    1.57753   0.702   0.4824
## activitiesyes    -2.30271    1.57795  -1.459   0.1445
## nurseryyes       -1.21066    1.76925  -0.684   0.4938
## higheryes        -4.59279    3.86246  -1.189   0.2344
## internetyes       2.65378    2.27343   1.167   0.2431
## romanticyes      -3.54763    1.86237  -1.905   0.0568 .
## famrel            4.22649    1.83052   2.309   0.0209 *
## freetime         -0.57928    0.91781  -0.631   0.5279
## goout            -1.13394    0.73992  -1.533   0.1254
## Dalc              2.17567    1.78741   1.217   0.2235
## Walc              1.61149    1.09972   1.465   0.1428
## health           -1.06596    0.79556  -1.340   0.1803
## absences          0.02243    0.07313   0.307   0.7591
## G1                0.95864    0.58514   1.638   0.1014
## G2                9.03102    3.62351   2.492   0.0127 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 500.504  on 394  degrees of freedom
## Residual deviance:  44.988  on 353  degrees of freedom
## AIC: 128.99
##
## Number of Fisher Scoring iterations: 13
```

```
#bestBIC(pass~.  ,data=dfbin, family="binomial")
#fitbin <- glm(pass~Fedu+ famrel+ goout+ Walc+ G2,data=dfbin,family="binomial")
#summary(fitbin)
```

```r
# df$gradelevel <- cut(df$G3, breaks=c(0,9,11,13,15,20), labels=c("Fail", "Sufficient", "Satisfactory",
# df$gradecat <- cut(df$G3, breaks=c(0,9,11,13,15,20), labels=c(0,1,2,3,4))
# dfless2 <- select(dfless, -c("gradelevel"))
# fitcatall <- glm(gradecat~.,data=dfless2,family=poisson())
# summary(fitcatall)
```

Optional: Gradelevels

1.4 Sub-sample analysis of passing/failing students

```r
dfpass <- subset(df,pass==1)
#dfpass <- select(dfpass, -c("pass", "resl", "resbn", "predl"))
dffail <- subset(df, pass==0)
#dffail <- select(dffail, -c("pass", "resl", "resbn", "predl"))
fitallpass <- lm(G3~.,data=dfpass)
summary(fitallpass)
```

Sub-sample analysis of students who pass

```
##
## Call:
## lm(formula = G3 ~ ., data = dfpass)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.93667 -0.45622 -0.03427  0.45080  1.84095
##
## Coefficients: (1 not defined because of singularities)
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.653448   1.115667   0.586 0.558668
## schoolMS         -0.250633   0.187192  -1.339 0.181964
## sexM             -0.024198   0.112970  -0.214 0.830589
## age               0.003502   0.052854   0.066 0.947231
## addressU          0.003725   0.141532   0.026 0.979026
## famsizeLE3        0.007962   0.110158   0.072 0.942445
## PstatusT         -0.047544   0.158709  -0.300 0.764785
## Medu             -0.041880   0.075271  -0.556 0.578504
## Fedu             -0.062279   0.062422  -0.998 0.319495
## Mjobhealth        0.545161   0.267498   2.038 0.042730 *
## Mjobother         0.111220   0.183999   0.604 0.546154
## Mjobservices      0.174862   0.192070   0.910 0.363590
## Mjobteacher       0.496627   0.256098   1.939 0.053737 .
## Fjobhealth       -0.230863   0.348351  -0.663 0.508188
## Fjobother         0.166935   0.257958   0.647 0.518207
## Fjobservices      0.095919   0.265790   0.361 0.718528
## Fjobteacher       0.172909   0.308417   0.561 0.575609
## reasonhome       -0.112774   0.129441  -0.871 0.384558
## reasonother      -0.028207   0.184303  -0.153 0.878501
## reasonreputation  0.002571   0.136236   0.019 0.984959
## guardianmother    0.050054   0.120403   0.416 0.678014
## guardianother    -0.655441   0.262934  -2.493 0.013401 *
## traveltime       -0.028367   0.080610  -0.352 0.725241
## studytime         0.043637   0.065685   0.664 0.507158
```

```
## failures          0.237626   0.111422   2.133 0.034045 *
## schoolsupyes      0.230191   0.174704   1.318 0.188987
## famsupyes         0.164224   0.113043   1.453 0.147697
## paidyes          -0.190585   0.109098  -1.747 0.082029 .
## activitiesyes    -0.116865   0.104537  -1.118 0.264801
## nurseryyes       -0.165216   0.127743  -1.293 0.197228
## higheryes        -0.105594   0.328269  -0.322 0.748006
## internetyes      -0.044121   0.148931  -0.296 0.767314
## romanticyes       0.109575   0.113665   0.964 0.336083
## famrel            0.217002   0.058368   3.718 0.000254 ***
## freetime         -0.010918   0.054952  -0.199 0.842694
## goout            -0.009276   0.054569  -0.170 0.865181
## Dalc              0.030025   0.076335   0.393 0.694455
## Walc             -0.068396   0.059404  -1.151 0.250811
## health           -0.083907   0.036343  -2.309 0.021874 *
## absences         -0.004201   0.008588  -0.489 0.625170
## G1                0.086378   0.039268   2.200 0.028855 *
## G2                0.868497   0.042525  20.423  < 2e-16 ***
## pass                    NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7568 on 223 degrees of freedom
## Multiple R-squared:  0.9225, Adjusted R-squared:  0.9082
## F-statistic: 64.72 on 41 and 223 DF,  p-value: < 2.2e-16
```

```
#bestBIC(G3~., data=dfpass)
```

–> age, absences and G1 out –>rather consider significant effects in full model than bestBIC? –>R^2 over 90%

```
fitallfail <- lm(G3~.,data=dffail)
summary(fitallfail)
```

Sub-sample analysis of students who fail

```
##
## Call:
## lm(formula = G3 ~ ., data = dffail)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2210 -1.5748  0.3484  1.6023  5.0210
##
## Coefficients: (1 not defined because of singularities)
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -2.03644    6.29383  -0.324 0.747038
## schoolMS          1.54380    1.11673   1.382 0.170338
## sexM             -0.21170    0.77529  -0.273 0.785450
## age              -0.24890    0.29341  -0.848 0.398577
## addressU          0.57653    0.81137   0.711 0.479234
## famsizeLE3        0.33192    0.70402   0.471 0.638470
## PstatusT          0.86715    1.09840   0.789 0.431962
## Medu              0.66513    0.48302   1.377 0.171997
```

```
## Fedu                -0.54232    0.40624  -1.335 0.185332
## Mjobhealth          -0.81596    1.49560  -0.546 0.586739
## Mjobother           -0.35380    0.89001  -0.398 0.691941
## Mjobservices        -0.61932    1.03779  -0.597 0.552198
## Mjobteacher         -0.78792    1.39620  -0.564 0.573966
## Fjobhealth           1.65507    1.74636   0.948 0.345866
## Fjobother           -0.57557    1.27088  -0.453 0.651742
## Fjobservices        -0.55372    1.33395  -0.415 0.679080
## Fjobteacher          0.59131    1.71526   0.345 0.731115
## reasonhome          -0.77937    0.70285  -1.109 0.270510
## reasonother         -0.09359    1.35668  -0.069 0.945156
## reasonreputation     0.09594    0.80110   0.120 0.904948
## guardianmother       0.36261    0.82234   0.441 0.660328
## guardianother        1.32476    1.19833   1.105 0.271956
## traveltime          -0.13482    0.45808  -0.294 0.769217
## studytime           -0.64856    0.44372  -1.462 0.147406
## failures            -0.18064    0.35999  -0.502 0.617069
## schoolsupyes         0.92002    0.88486   1.040 0.301314
## famsupyes            0.27274    0.67038   0.407 0.685107
## paidyes              0.14694    0.76786   0.191 0.848680
## activitiesyes       -0.65688    0.63436  -1.035 0.303275
## nurseryyes          -0.44069    0.79573  -0.554 0.581105
## higheryes            0.37062    1.13587   0.326 0.744980
## internetyes         -0.47894    0.78490  -0.610 0.543307
## romanticyes         -0.85908    0.62574  -1.373 0.173273
## famrel               0.40665    0.34527   1.178 0.242050
## freetime             0.31295    0.33470   0.935 0.352346
## goout                0.20412    0.30554   0.668 0.505838
## Dalc                -0.67639    0.47582  -1.422 0.158695
## Walc                 0.50618    0.32466   1.559 0.122561
## health               0.55231    0.22719   2.431 0.017085 *
## absences             0.10296    0.02990   3.444 0.000881 ***
## G1                   0.17489    0.20783   0.841 0.402350
## G2                   0.75415    0.13082   5.765 1.19e-07 ***
## pass                      NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.829 on 88 degrees of freedom
## Multiple R-squared:  0.5903, Adjusted R-squared:  0.3995
## F-statistic: 3.093 on 41 and 88 DF,  p-value: 4.74e-06
```

```
#bestBIC(G3~.,data=dffail)
```

–>also absences and previous performance –>R^2 much lower though (close to 60%)

1.5 Grade difference

What students have improved their grades over the course of the year? What role did support from the family/school play?

```
df.diff <- df %>%
  mutate(gradediff13 = G3 - G1) %>%
  select(-c("G1", "G2", "G3", "pass")) %>%
```

```
  mutate(improvement = ifelse(gradediff13 >= 0, 1, 0))
```

Creating dataframe

```
fitdiff1 <- lm(gradediff13 ~ ., data = df.diff)
summary(fitdiff1)
```

Linear regression on grade difference

```
##
## Call:
## lm(formula = gradediff13 ~ ., data = df.diff)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4135 -0.8124  0.2612  1.1856  3.7682
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -0.3243180  2.0578134  -0.158  0.87486
## schoolMS          0.7035945  0.3623768   1.942  0.05298 .
## sexM              0.2553044  0.2289919   1.115  0.26565
## age              -0.2144842  0.0995595  -2.154  0.03189 *
## addressU          0.3042122  0.2674579   1.137  0.25613
## famsizeLE3        0.0854238  0.2237448   0.382  0.70285
## PstatusT          0.0302109  0.3325120   0.091  0.92766
## Medu              0.2495887  0.1480239   1.686  0.09265 .
## Fedu             -0.2135568  0.1271068  -1.680  0.09381 .
## Mjobhealth       -0.6707698  0.5134646  -1.306  0.19228
## Mjobother        -0.0131682  0.3273301  -0.040  0.96793
## Mjobservices     -0.0982754  0.3655819  -0.269  0.78823
## Mjobteacher      -0.2392037  0.4753068  -0.503  0.61509
## Fjobhealth        0.8735055  0.6582941   1.327  0.18539
## Fjobother         0.1403961  0.4687794   0.299  0.76474
## Fjobservices     -0.1384633  0.4852109  -0.285  0.77553
## Fjobteacher      -0.4817632  0.5944934  -0.810  0.41827
## reasonhome        0.0727647  0.2536753   0.287  0.77440
## reasonother       0.6040290  0.3747677   1.612  0.10791
## reasonreputation  0.1778373  0.2639497   0.674  0.50091
## guardianmother   -0.1219058  0.2498885  -0.488  0.62596
## guardianother    -0.2521422  0.4576087  -0.551  0.58198
## traveltime       -0.2540650  0.1551910  -1.637  0.10250
## studytime        -0.0525449  0.1316872  -0.399  0.69012
## failures         -0.3317940  0.1524609  -2.176  0.03020 *
## schoolsupyes      0.6276433  0.3054668   2.055  0.04064 *
## famsupyes        -0.0561686  0.2193407  -0.256  0.79804
## paidyes           0.0254209  0.2198622   0.116  0.90802
## activitiesyes    -0.0752253  0.2039824  -0.369  0.71251
## nurseryyes       -0.1486564  0.2514920  -0.591  0.55483
## higheryes         0.2845832  0.4934196   0.577  0.56447
## internetyes       0.0001162  0.2839313   0.000  0.99967
## romanticyes      -0.6002199  0.2153616  -2.787  0.00561 **
## famrel            0.0911353  0.1127533   0.808  0.41948
```

```
## freetime          0.0933221  0.1086874   0.859  0.39113
## goout            -0.0672447  0.1029581  -0.653  0.51410
## Dalc             -0.0155442  0.1518322  -0.102  0.91851
## Walc              0.1439313  0.1138060   1.265  0.20681
## health            0.0265653  0.0737357   0.360  0.71885
## absences          0.0599183  0.0132884   4.509 8.87e-06 ***
## improvement       3.7733875  0.2034490  18.547  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.881 on 354 degrees of freedom
## Multiple R-squared:  0.5836, Adjusted R-squared:  0.5365
## F-statistic:  12.4 on 40 and 354 DF,  p-value: < 2.2e-16
```

```
bestBIC(gradediff13 ~. , data = df.diff)
```

```
## Greedy searching posterior mode... Done.
## Running Gibbs sampler.......... Done.
```

```
## icfit object
##
## Model with best BIC : age failures romanticyes absences improvement
##
## Use summary(), coef() and predict() to get inference for the top model
## Use coef(object$msfit) and predict(object$msfit) to get BMA estimates and predictions
```

```
fitdiff2 <- glm(improvement ~ . -gradediff13, data = df.diff, family = "binomial")
summary(fitdiff2)
```

Binomial regression: improvement yes/no

```
##
## Call:
## glm(formula = improvement ~ . - gradediff13, family = "binomial",
##     data = df.diff)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0733  -1.1547   0.6891   0.9829   1.6526
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       1.660102   2.340033   0.709   0.4781
## schoolMS          0.007561   0.409094   0.018   0.9853
## sexM              0.138670   0.263241   0.527   0.5983
## age              -0.118934   0.113621  -1.047   0.2952
## addressU          0.096723   0.303617   0.319   0.7501
## famsizeLE3        0.245913   0.259981   0.946   0.3442
## PstatusT         -0.637462   0.402690  -1.583   0.1134
## Medu              0.101530   0.170832   0.594   0.5523
## Fedu             -0.037278   0.144799  -0.257   0.7968
## Mjobhealth        1.017995   0.619109   1.644   0.1001
## Mjobother         0.525260   0.367041   1.431   0.1524
## Mjobservices      0.341816   0.410977   0.832   0.4056
## Mjobteacher      -0.104827   0.534472  -0.196   0.8445
```

```
## Fjobhealth       -0.051666   0.756529  -0.068   0.9456
## Fjobother         0.409252   0.535061   0.765   0.4443
## Fjobservices      0.782443   0.552215   1.417   0.1565
## Fjobteacher       0.690087   0.676291   1.020   0.3075
## reasonhome       -0.199395   0.289384  -0.689   0.4908
## reasonother       0.444574   0.443633   1.002   0.3163
## reasonreputation -0.010528   0.302257  -0.035   0.9722
## guardianmother    0.187857   0.285676   0.658   0.5108
## guardianother     0.191534   0.515908   0.371   0.7104
## traveltime        0.037641   0.174319   0.216   0.8290
## studytime        -0.017160   0.151115  -0.114   0.9096
## failures         -0.090979   0.172348  -0.528   0.5976
## schoolsupyes      0.222796   0.356050   0.626   0.5315
## famsupyes         0.208431   0.248721   0.838   0.4020
## paidyes           0.528467   0.252678   2.091   0.0365 *
## activitiesyes    -0.259266   0.233849  -1.109   0.2676
## nurseryyes       -0.064662   0.286118  -0.226   0.8212
## higheryes        -0.075906   0.546976  -0.139   0.8896
## internetyes       0.297115   0.318754   0.932   0.3513
## romanticyes      -0.350910   0.244638  -1.434   0.1515
## famrel            0.144202   0.129031   1.118   0.2637
## freetime         -0.070609   0.125488  -0.563   0.5737
## goout            -0.136595   0.118011  -1.157   0.2471
## Dalc             -0.222122   0.170987  -1.299   0.1939
## Walc              0.170712   0.129336   1.320   0.1869
## health           -0.043910   0.085839  -0.512   0.6090
## absences         -0.019391   0.014833  -1.307   0.1911
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 532.48  on 394  degrees of freedom
## Residual deviance: 490.17  on 355  degrees of freedom
## AIC: 570.17
##
## Number of Fisher Scoring iterations: 4
```

```
bestBIC(improvement ~. -gradediff13, data = df.diff)
```

```
## Greedy searching posterior mode... Done.
## Running Gibbs sampler.......... Done.
```

```
## icfit object
##
## Model with best BIC : (Intercept) paidyes
##
## Use summary(), coef() and predict() to get inference for the top model
## Use coef(object$msfit) and predict(object$msfit) to get BMA estimates and predictions
```

- failures, romantic relationships and absences seem to be important factors.
- Interestingly, no type of support has a significant effect, are there heterogeneous effects and reverse causality? Can we test that somehow?
- With binary outcome improvement yes/no absences and resonother seem important, although bestBIC suggests only age as predictor
- If improvement is relaxed to >= 0 instead of >0, paidyes becomes significant and positive -> interesting!

```r
table(df.diff$improvement)
```

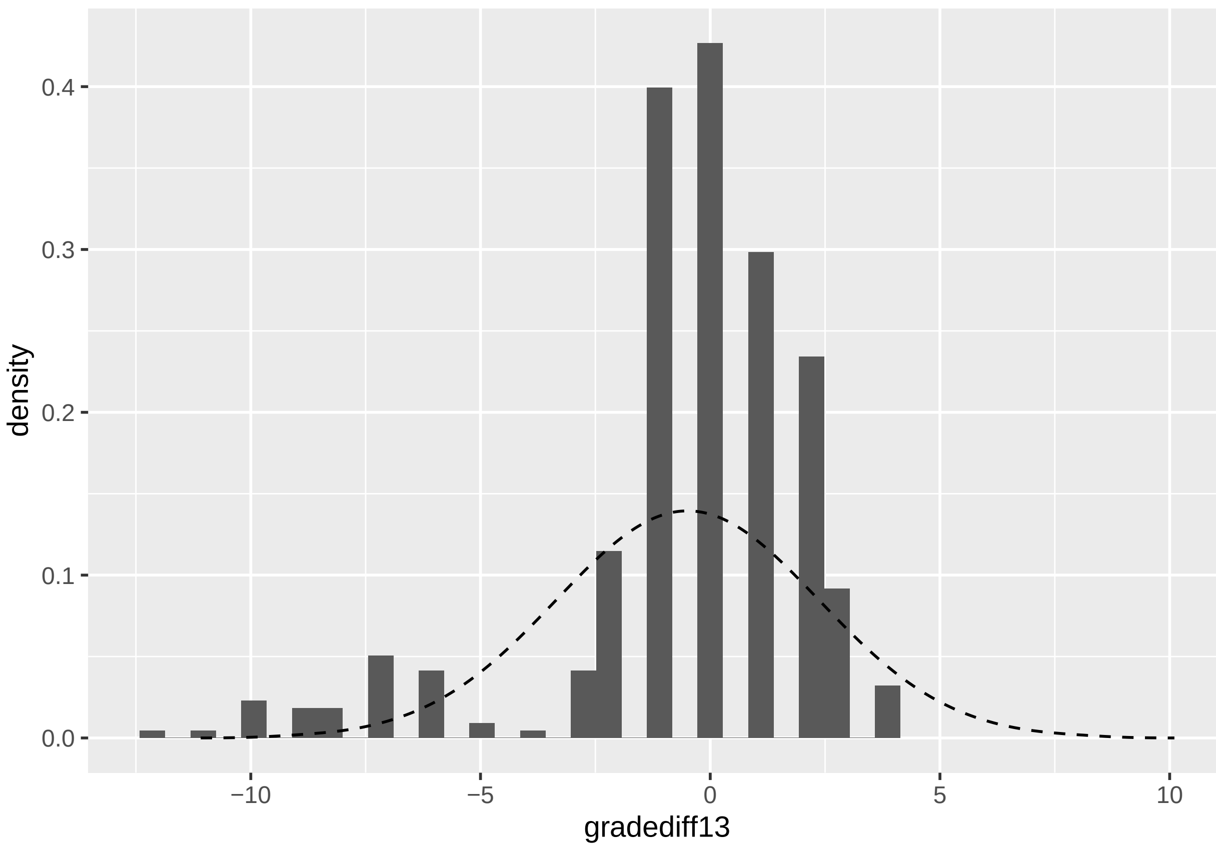
Plots

```
##
##   0   1
## 159 236
```

```r
library(ggpubr)
ggplot(data = df.diff, aes(x = gradediff13)) +
  geom_histogram(aes(y = ..density..)) +
  stat_overlay_normal_density(linetype = "dashed")
```

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



1.6 Subsample analysis of students that improved vs. those that did not

```r
df.posdiff <- subset(df.diff, improvement == 1)
fitdiff3 <- glm(gradediff13 ~ . -improvement, data = df.posdiff)
summary(fitdiff3)
```

```
##
## Call:
## glm(formula = gradediff13 ~ . - improvement, data = df.posdiff)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2386  -0.6687  -0.1078   0.6631   2.6645
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4.448269   1.614799   2.755  0.00643 **
## schoolMS         -0.403752   0.281349  -1.435  0.15286
## sexM             -0.015117   0.164364  -0.092  0.92682
## age              -0.184544   0.071594  -2.578  0.01068 *
## addressU         -0.062814   0.206555  -0.304  0.76137
## famsizeLE3        0.039611   0.162022   0.244  0.80711
## PstatusT          0.080042   0.225534   0.355  0.72305
## Medu              0.125436   0.109459   1.146  0.25321
## Fedu             -0.014591   0.094213  -0.155  0.87708
## Mjobhealth       -0.455259   0.370138  -1.230  0.22018
## Mjobother        -0.055222   0.256643  -0.215  0.82986
## Mjobservices     -0.089114   0.294994  -0.302  0.76290
## Mjobteacher      -0.185132   0.372726  -0.497  0.61996
## Fjobhealth        0.511097   0.531669   0.961  0.33758
## Fjobother         0.391915   0.403632   0.971  0.33276
## Fjobservices      0.096628   0.418349   0.231  0.81758
## Fjobteacher      -0.502212   0.492623  -1.019  0.30924
## reasonhome        0.148356   0.189211   0.784  0.43394
## reasonother       0.378052   0.258997   1.460  0.14598
## reasonreputation -0.133883   0.197290  -0.679  0.49819
## guardianmother   -0.045548   0.176790  -0.258  0.79696
## guardianother    -0.275545   0.356268  -0.773  0.44020
## traveltime       -0.128635   0.117202  -1.098  0.27375
## studytime        -0.110223   0.098184  -1.123  0.26297
## failures          0.256222   0.122408   2.093  0.03762 *
## schoolsupyes      0.284940   0.218794   1.302  0.19434
## famsupyes         0.052582   0.172837   0.304  0.76127
## paidyes          -0.276869   0.162182  -1.707  0.08938 .
## activitiesyes     0.034467   0.148962   0.231  0.81726
## nurseryyes       -0.307618   0.186313  -1.651  0.10032
## higheryes         0.200624   0.408691   0.491  0.62405
## internetyes      -0.167008   0.224790  -0.743  0.45840
## romanticyes      -0.070779   0.158164  -0.448  0.65500
## famrel            0.008674   0.080816   0.107  0.91463
## freetime         -0.022536   0.081129  -0.278  0.78147
## goout            -0.001872   0.080812  -0.023  0.98154
## Dalc              0.252903   0.115415   2.191  0.02961 *
## Walc             -0.107046   0.086857  -1.232  0.21926
## health           -0.010399   0.052817  -0.197  0.84413
## absences         -0.026675   0.012206  -2.185  0.03005 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 1.069063)
##
##     Null deviance: 285.47  on 235  degrees of freedom
## Residual deviance: 209.54  on 196  degrees of freedom
```

```
## AIC: 723.67
##
## Number of Fisher Scoring iterations: 2
```

```r
#bestBIC(gradediff13 ~. - improvement, data = df.posdiff)

df.negdiff <- subset(df.diff, improvement == 0)
fitdiff4 <- glm(gradediff13 ~ . -improvement, data = df.negdiff)
summary(fitdiff4)
```

```
##
## Call:
## glm(formula = gradediff13 ~ . - improvement, data = df.negdiff)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.9794  -0.7026   0.5124   1.3339   3.9598
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1.62016    4.25830  -0.380  0.70427
## schoolMS           1.66847    0.75227   2.218  0.02846 *
## sexM               0.87951    0.50965   1.726  0.08699 .
## age               -0.14338    0.23465  -0.611  0.54233
## addressU           0.54305    0.57703   0.941  0.34855
## famsizeLE3         0.30996    0.50484   0.614  0.54040
## PstatusT           0.29379    0.91724   0.320  0.74930
## Medu               0.66265    0.32861   2.017  0.04599 *
## Fedu              -0.60362    0.29229  -2.065  0.04108 *
## Mjobhealth        -2.00284    1.26778  -1.580  0.11681
## Mjobother         -0.52910    0.65580  -0.807  0.42139
## Mjobservices      -1.02550    0.71421  -1.436  0.15367
## Mjobteacher       -1.24441    0.96382  -1.291  0.19916
## Fjobhealth         0.89665    1.29465   0.693  0.48992
## Fjobother         -1.13867    0.89500  -1.272  0.20576
## Fjobservices      -0.94638    0.89791  -1.054  0.29403
## Fjobteacher       -1.16186    1.21967  -0.953  0.34272
## reasonhome        -0.33626    0.56379  -0.596  0.55202
## reasonother        0.83422    0.96015   0.869  0.38668
## reasonreputation   0.58200    0.57640   1.010  0.31468
## guardianmother    -0.23688    0.57135  -0.415  0.67919
## guardianother     -0.12584    1.00930  -0.125  0.90099
## traveltime        -0.65998    0.33489  -1.971  0.05108 .
## studytime         -0.32482    0.29576  -1.098  0.27432
## failures          -0.99823    0.31339  -3.185  0.00185 **
## schoolsupyes       0.28270    0.71657   0.395  0.69390
## famsupyes         -0.10118    0.48866  -0.207  0.83632
## paidyes            0.46480    0.49246   0.944  0.34716
## activitiesyes     -0.31410    0.44842  -0.700  0.48501
## nurseryyes        -0.35165    0.56202  -0.626  0.53271
## higheryes          1.52306    0.97107   1.568  0.11944
## internetyes        0.46742    0.58630   0.797  0.42690
## romanticyes       -1.12619    0.47386  -2.377  0.01907 *
## famrel            -0.06091    0.26897  -0.226  0.82124
## freetime           0.22661    0.23574   0.961  0.33835
```

```
## goout              0.02477    0.21702   0.114  0.90933
## Dalc              -0.28187    0.31842  -0.885  0.37782
## Walc               0.26375    0.24910   1.059  0.29183
## health             0.19689    0.17609   1.118  0.26576
## absences           0.12049    0.02350   5.128 1.15e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 5.993003)
##
##     Null deviance: 1268.42  on 158  degrees of freedom
## Residual deviance:  713.17  on 119  degrees of freedom
## AIC: 771.85
##
## Number of Fisher Scoring iterations: 2
```

```
#bestBIC(gradediff13 ~. - improvement, data = df.negdiff)
#fitdiff5 <- glm(gradediff13 ~ school + age + traveltime + failures + romantic + Walc + absences, data 
#summary(fitdiff5)
```

-> mean-center it? For improvers

2. Prediction model

–>for full linear model and the binary case

2.1 Training-Test Split

```
## 75% of the sample size
smp_size <- floor(0.90 * nrow(dfbin))

## set the seed to make your partition reproducible
set.seed(123)
train_ind <- sample(seq_len(nrow(dfbin)), size = smp_size)

train <- dfbin[train_ind, ]
test <- dfbin[-train_ind, ]
```

```
fitbin2= glm(pass~Fedu+ famrel+ goout+ Walc+ G2,data=train,family="binomial")
```

Rerun the models on the training data

```
pibintest= predict(fitbin2, type='response', newdata=test)
table(pibintest > 0.5, test$pass)
```

Make predictions on test data

```
##
##          0  1
##   FALSE 15  7
##   TRUE   2 16
```

```
cost_misclass= function(yobs, ypred) {
  err1= (ypred > 0.5) & (yobs==0)
  err2= (ypred < 0.5) & (yobs==1)
  ans= sum(err1 | err2) / length(yobs)
  return(ans)
}
misclas= c(cost_misclass(test$pass, pibintest))
names(misclas)= c('model 1')
misclas
```

Assess mis-classification in test data

```
## model 1
##   0.225
```

```
pibin= predict(fitbin2, type='response', data = train) # data??
loss.insample= c(cost_misclass(train$pass, pibin))
names(loss.insample)= c('model 1')
loss.insample
```

Compare to misclassification in training data

```
##    model 1
## 0.06478873
```

```
table(pibin > 0.5, train$pass)
```

```
##
##           0   1
##   FALSE 100  10
##   TRUE   13 232
```

2.2 Cross-validation

```
fitbin3= glm(pass~Fedu+ famrel+ goout+ Walc+ G2,data=dfbin,family="binomial")
fitbin3cv= cv.glm(dfbin, fitbin3, cost=cost_loglik_logistic, K=10)
loss= sqrt(fitbin3cv$delta)
loss
```

```
## [1] 2.635270 2.554554
```

- We should compare that to another model!

APPENDIX

A.1 Validating Assumptions

A.1.1 LM-model check –>Models: fit1=bestBIC (fitall auch?)

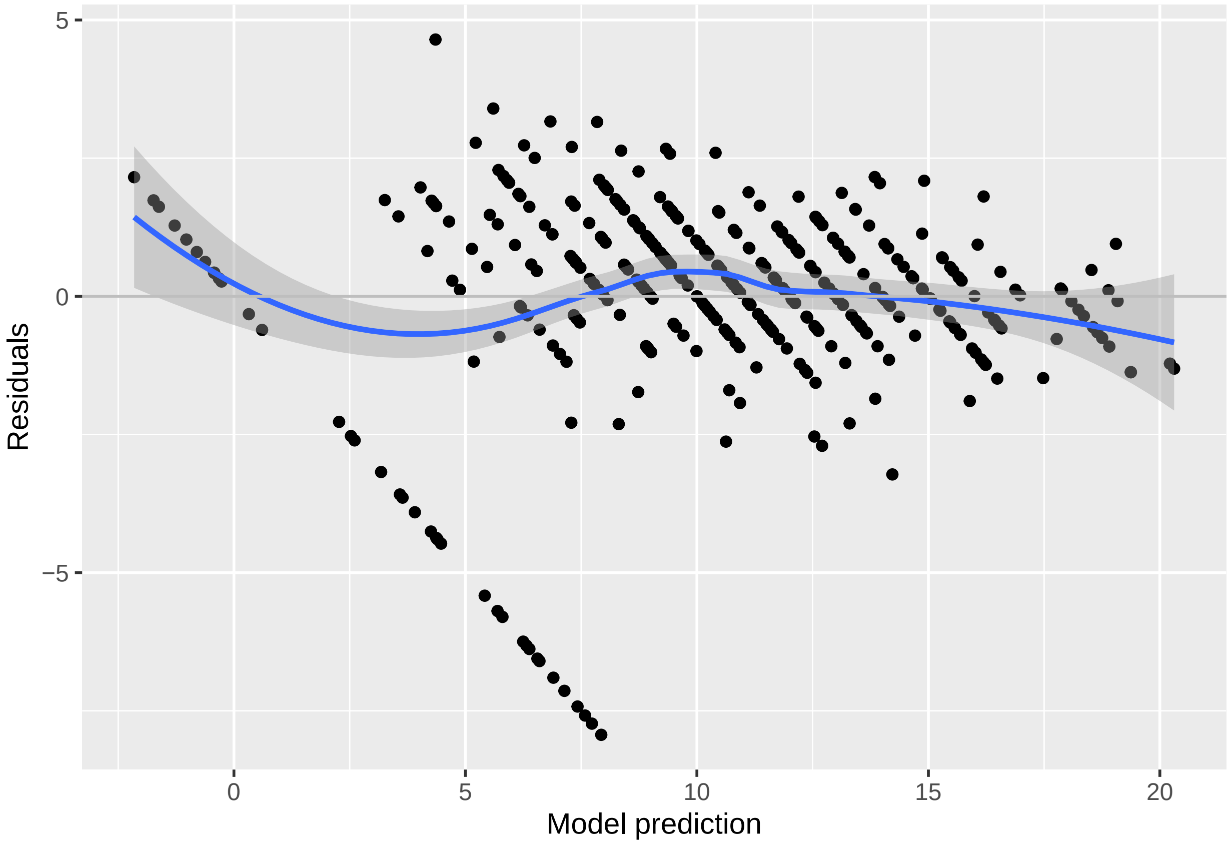
```
df$predl= predict(fitall)
df$resl= residuals(fitall)
```

Linearity

```
ggplot(df, aes(predl, resl)) +
  geom_point() +
```

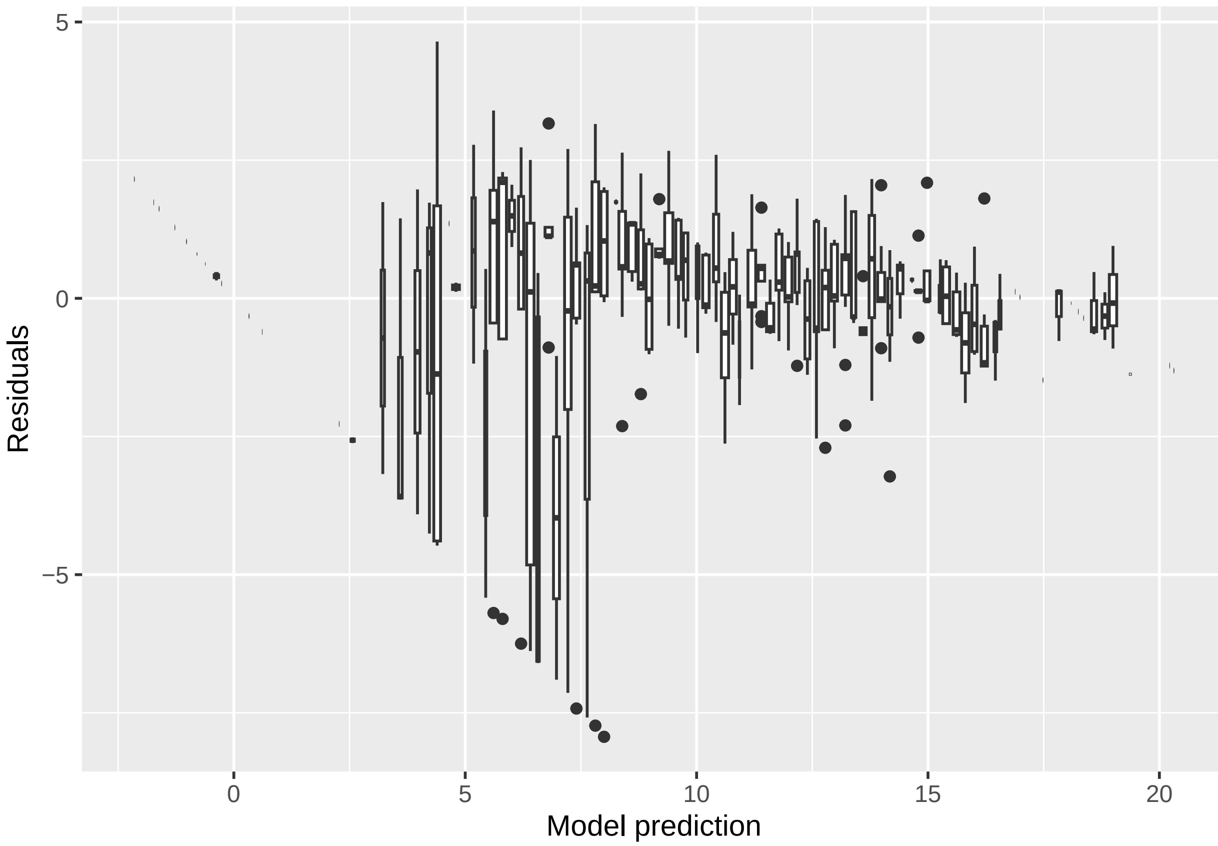
```
  geom_smooth() +
  geom_abline(slope=0, intercept=0, col='gray') +
  labs(x='Model prediction', y='Residuals')
```

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



Constant residual variance

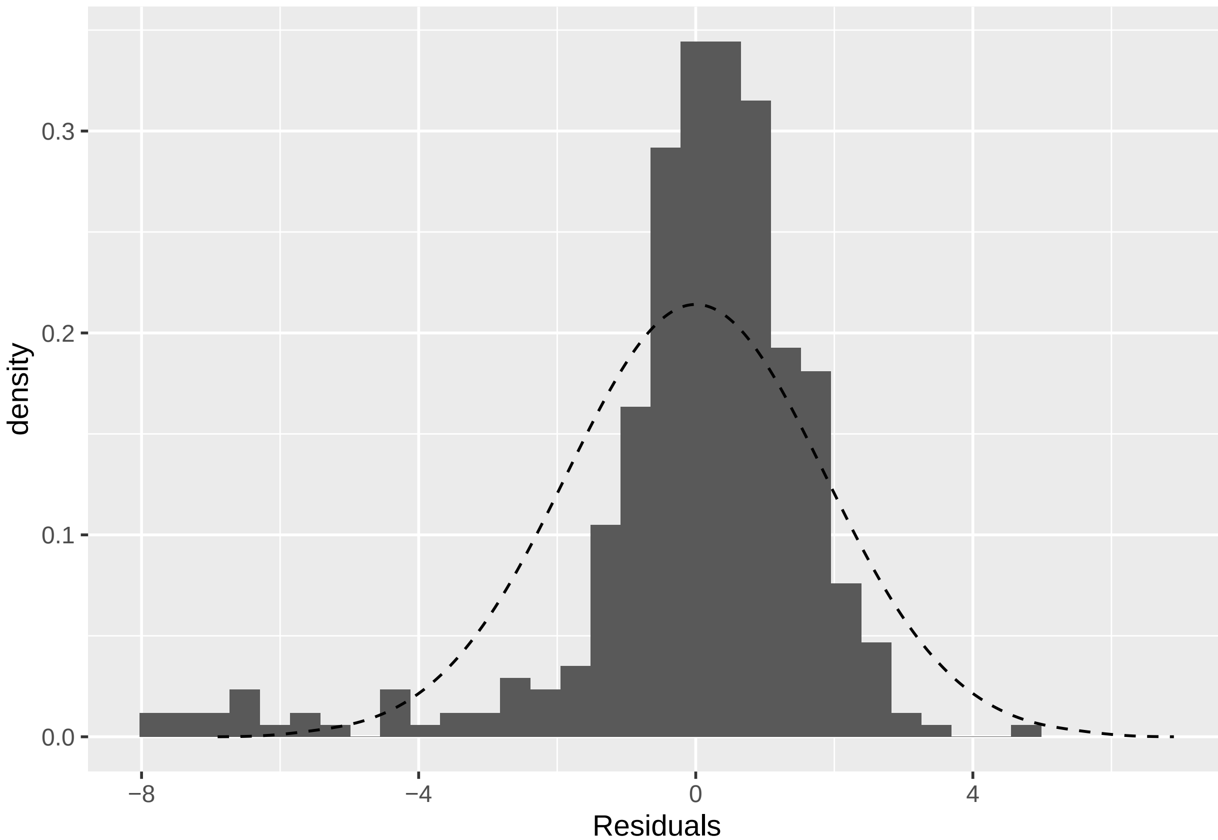
```
ggplot(df, aes(x=predl, y=resl)) +
  geom_boxplot(mapping = aes(group = cut_width(predl, 0.2))) +
  labs(x='Model prediction', y='Residuals')
```



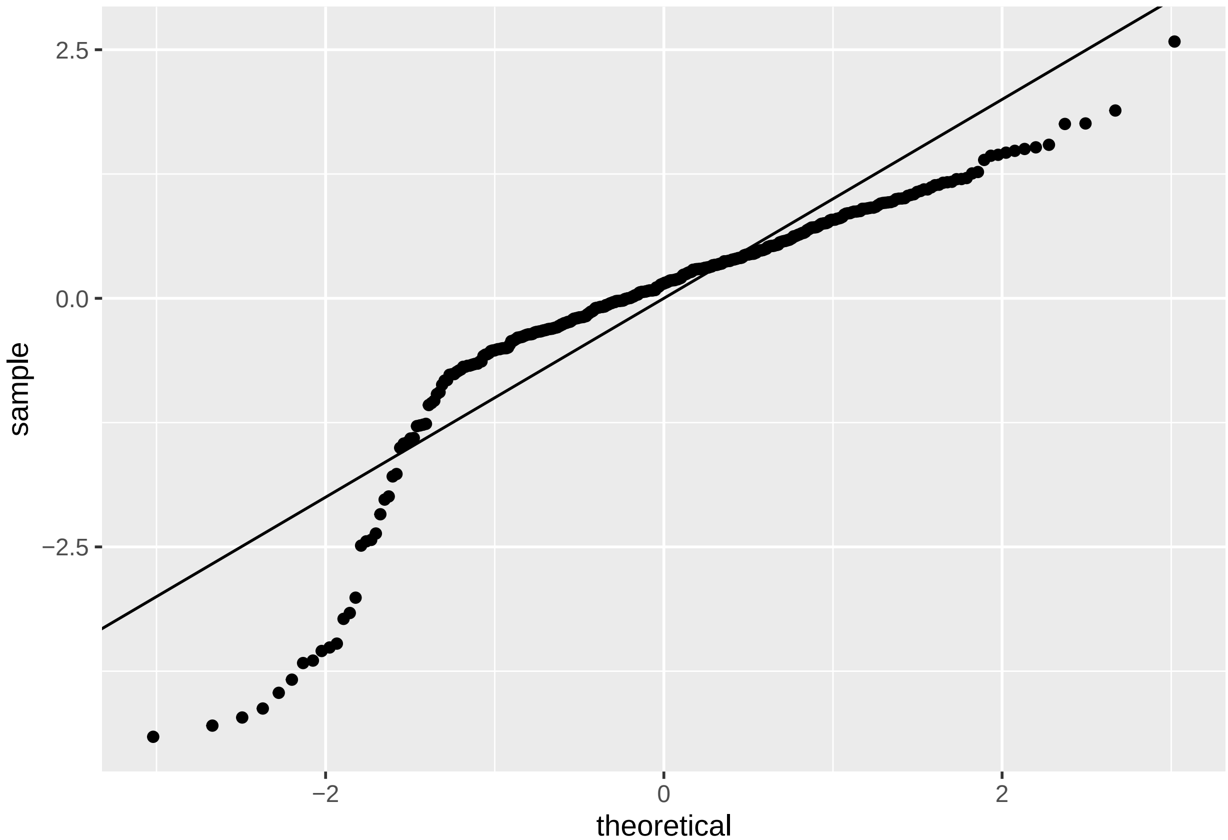
Error normality

```
ggplot(df, aes(x=resl)) +
  geom_histogram(aes(y= ..density..)) +
  stat_overlay_normal_density(linetype = "dashed") +
  labs(x='Residuals')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(df, aes(sample=scale(resl))) +
  geom_qq() +
  geom_abline(slope=1, intercept=0)
```



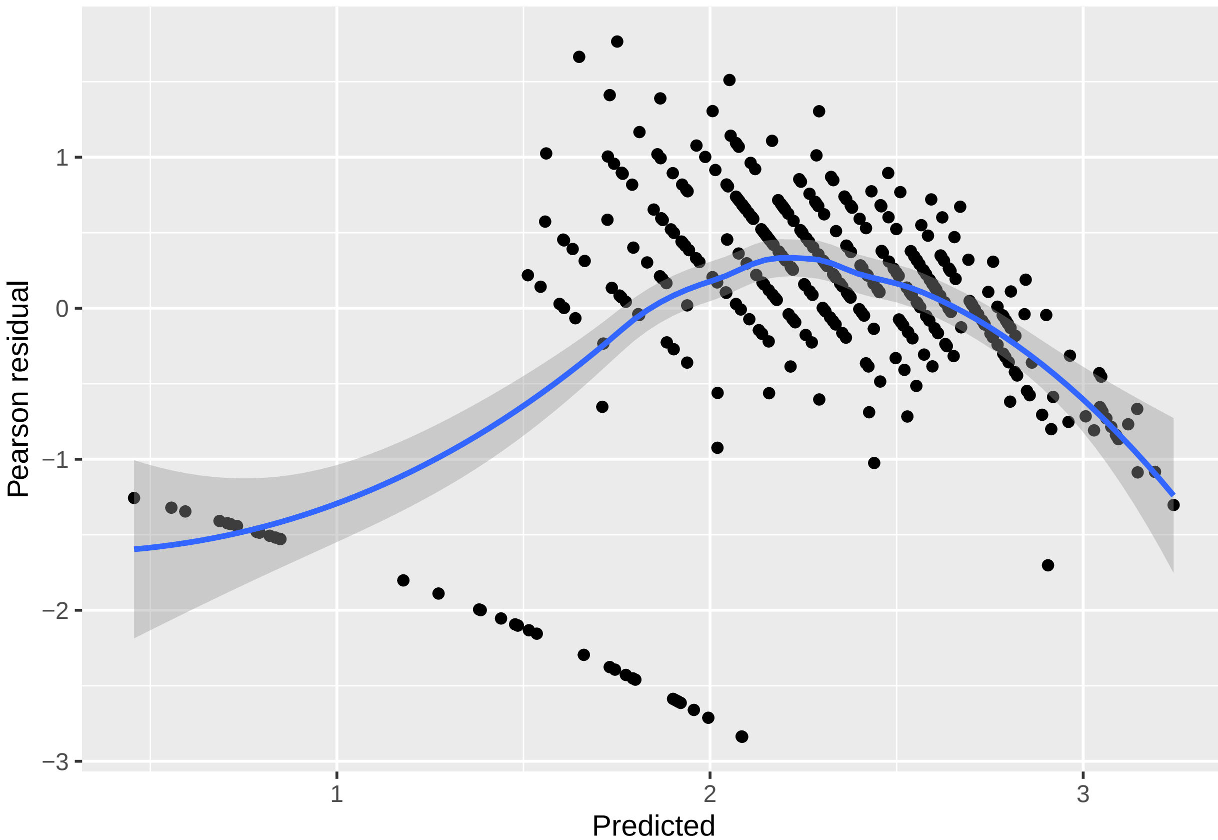
−>Errors are not normal and variance is not constant!! =>Apply robust standard errors

```
poires= mutate(df, pred= predict(fitallp), resdev= residuals(fitallp, type='deviance'), respearson= resi
```

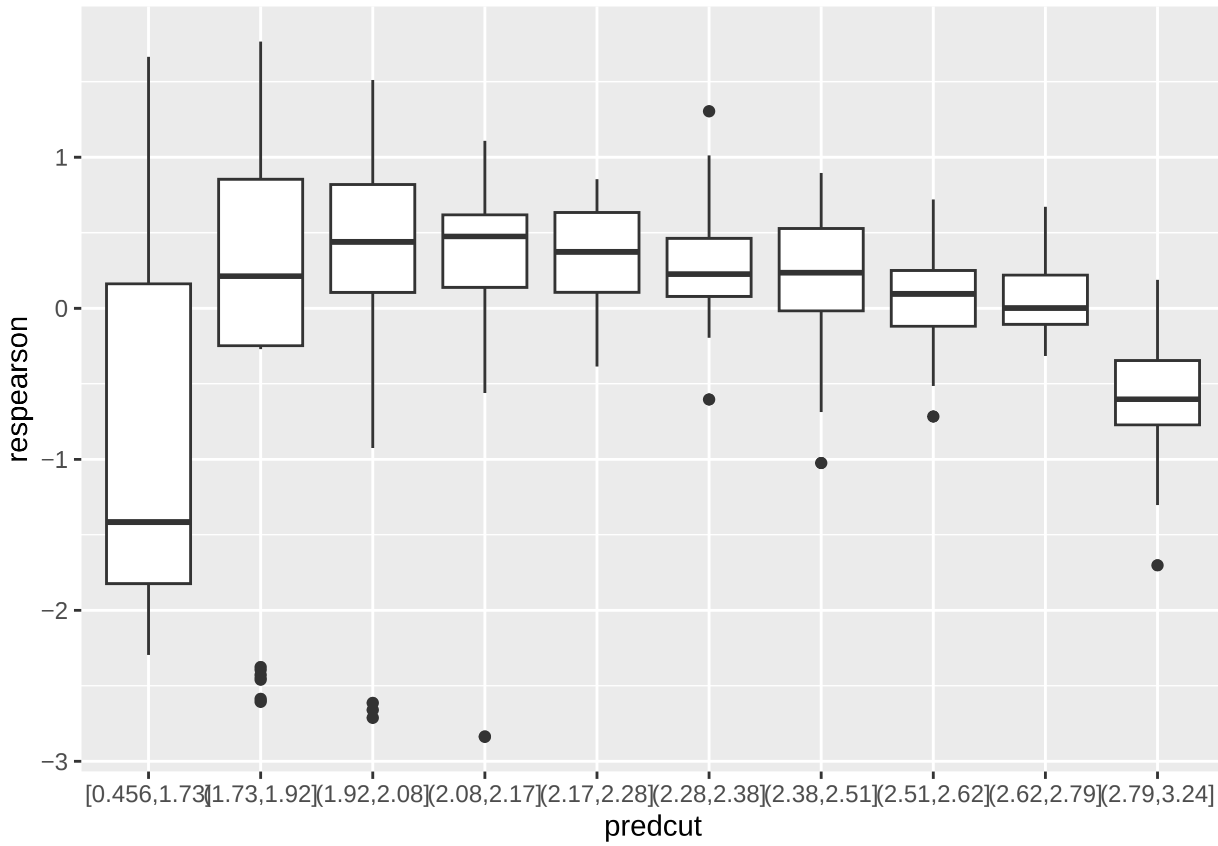
```
ggplot(poires, aes(pred, respearson)) + geom_point() + geom_smooth() + labs(x='Predicted', y='Pearson re
```

A.1.2 Poisson-check

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




```
poires2= mutate(poires, predcut= cut_number(pred, 10))
ggplot(poires2, aes(x=predcut, y=respearson)) + geom_boxplot()
```




```
mean(poires$respearson)
```

```
## [1] -0.0167188
```

```
sd(poires$respearson)
```

```
## [1] 0.8287321
```

```
mean(df$G3)
```

```
## [1] 10.41519
```

```
var(df$G3)
```

```
## [1] 20.98962
```

=>Huge overdispersion and variance not constant, errors not normal

Error normality?

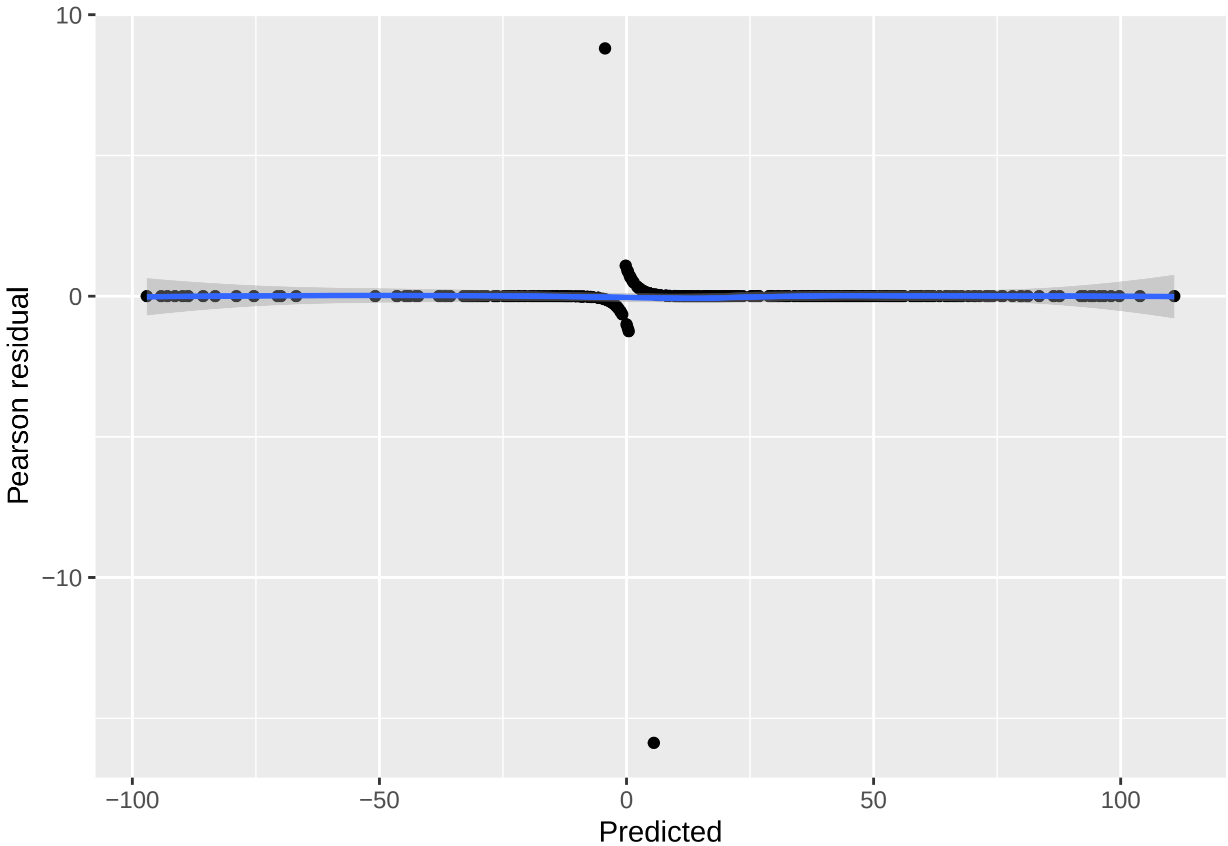
```
# ggplot(poires, aes(x=respearson)) +
#   geom_histogram(aes(y= ..density..)) +
#   stat_overlay_normal_density(linetype = "dashed") +
#   labs(x='Residuals')
# ggplot(poires, aes(sample=scale(respearson))) +
#   geom_qq() +
#   geom_abline(slope=1, intercept=0)
```

```
binres= mutate(df, pred= predict(fitallb), resdev= residuals(fitallb, type='deviance'), respearson= resi
```

```
ggplot(binres, aes(pred, respearson)) + geom_point() + geom_smooth() + labs(x='Predicted', y='Pearson re
```

A.1.3 Binomial check

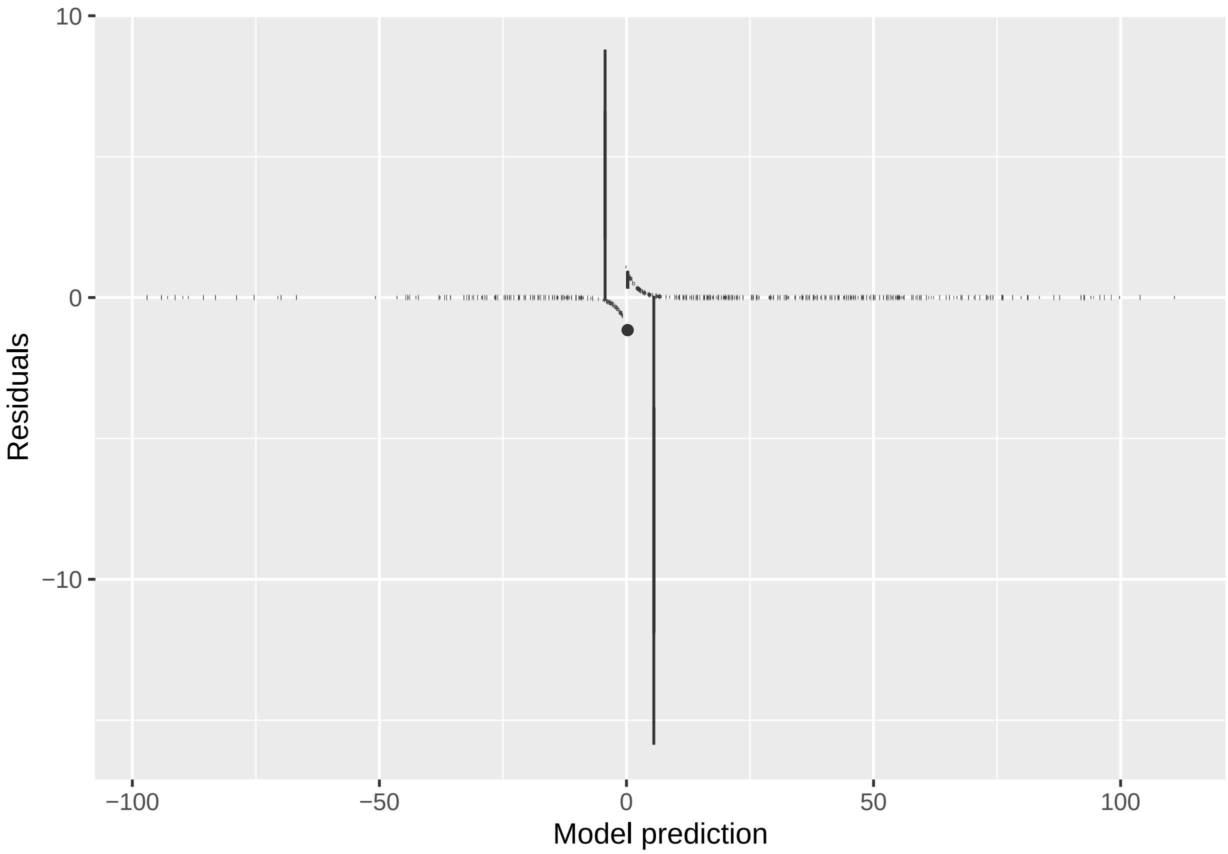
```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



Residuals seem roughly centered at zero

Constant residual variance

```
ggplot(binres, aes(x=pred, y=respearson)) +
  geom_boxplot(mapping = aes(group = cut_width(pred, 0.2))) +
  labs(x='Model prediction', y='Residuals')
```

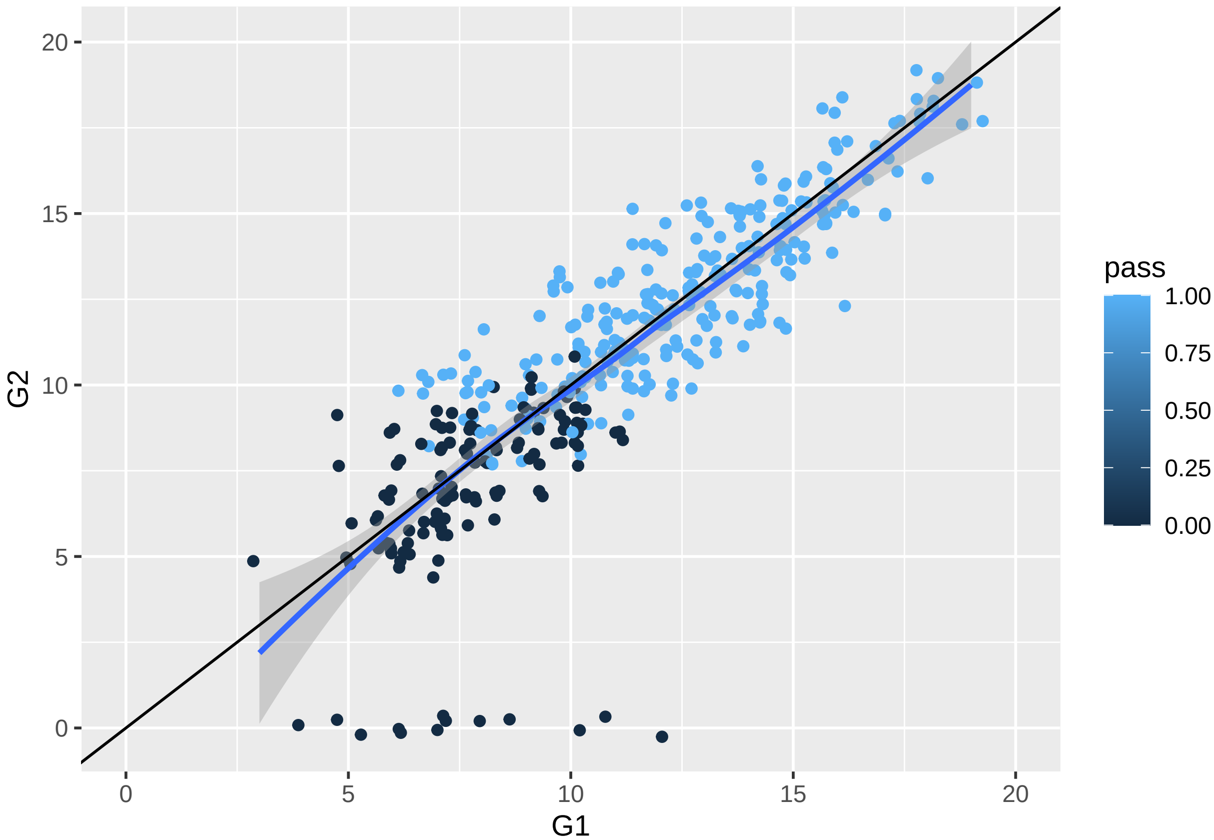


A.2 Grade difference Plots

```
ggplot(data=df, aes(G1,G2,color=pass))+
  geom_point(position="jitter")+
  geom_smooth()+
  geom_abline(slope=1)+
  coord_cartesian(xlim=c(0,20))
```

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

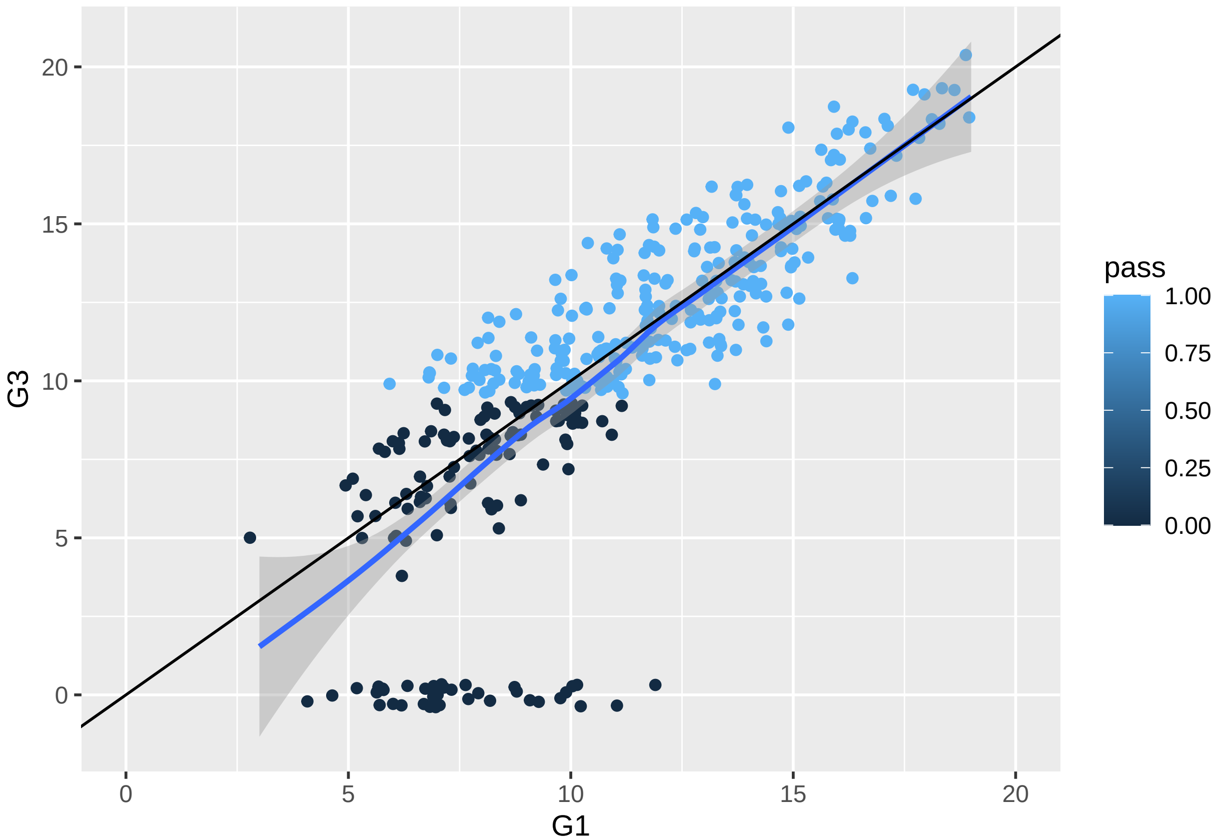
```
## Warning: The following aesthetics were dropped during statistical transformation: colour
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```




```r
ggplot(data=df, aes(G1,G3,color=pass,scale))+
  geom_point(position="jitter")+
  geom_smooth()+
  geom_abline(slope=1)+
  coord_cartesian(xlim=c(0,20))
```

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

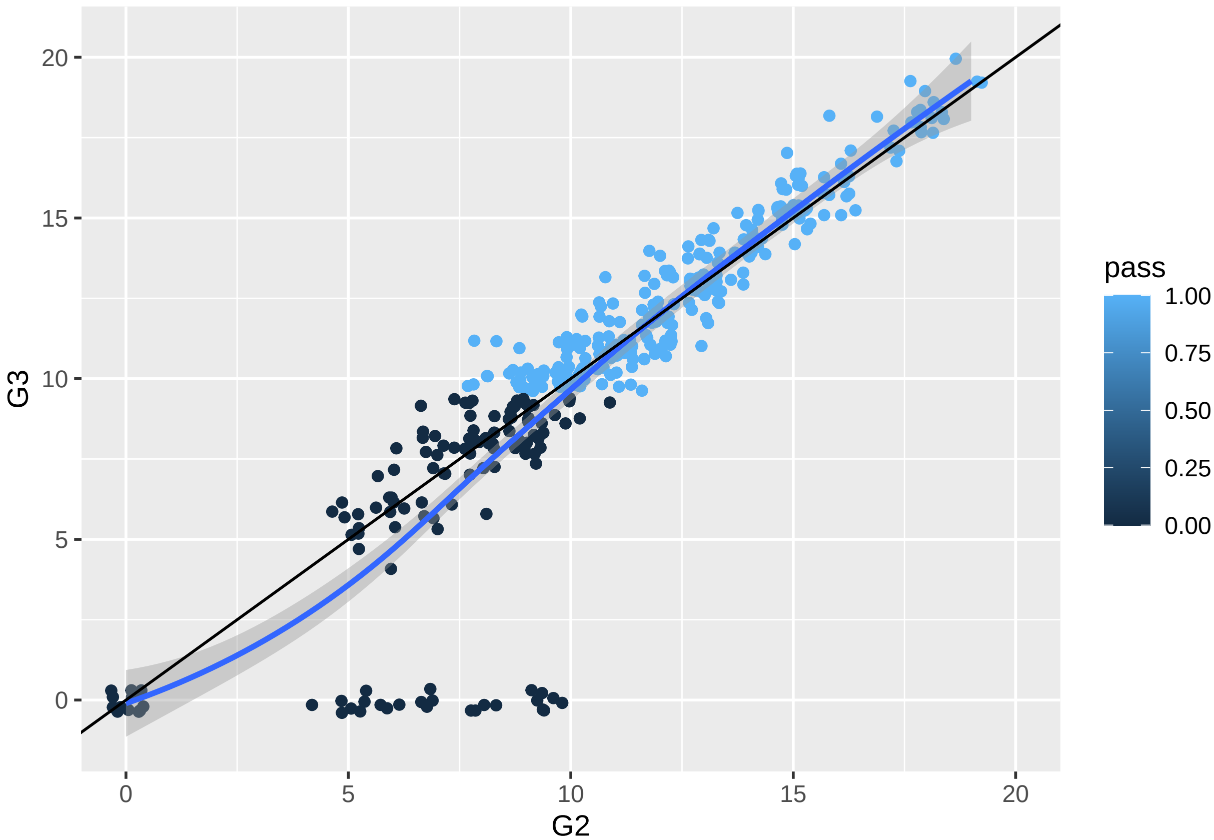
```
## Warning: The following aesthetics were dropped during statistical transformation: colour
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```




```
ggplot(data=df, aes(G2,G3,color=pass))+
  geom_point(position="jitter")+
  geom_smooth()+
  geom_abline(slope=1)+
  coord_cartesian(xlim=c(0,20))
```

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

```
## Warning: The following aesthetics were dropped during statistical transformation: colour
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```



G3
G2
pass
1.00
0.75
0.50
0.25
0.00
0
5
10
15
20

	Full Linear model	Full linear model without grades
(Intercept)	-1.1155	14.0777 **
	(0.5986)	(0.0018)
schoolMS	0.4807	0.7256
	(0.1905)	(0.3600)
sexM	0.1744	1.2624 *
	(0.4558)	(0.0120)
age	-0.1733	-0.3752
	(0.0864)	(0.0850)
addressU	0.1045	0.5513
	(0.6999)	(0.3459)
famsizeLE3	0.0365	0.7028
	(0.8721)	(0.1509)
PstatusT	-0.1277	-0.3201
	(0.7039)	(0.6586)
Medu	0.1297	0.4569
	(0.3879)	(0.1583)
Fedu	-0.1339	-0.1046
	(0.2990)	(0.7066)
Mjobhealth	-0.1464	0.9981
	(0.7778)	(0.3727)
Mjobother	0.0741	-0.3590
	(0.8236)	(0.6150)
Mjobservices	0.0470	0.6583
	(0.8990)	(0.4099)
Mjobteacher	-0.0263	-1.2415
	(0.9565)	(0.2326)
Fjobhealth	0.3309	0.3477
	(0.6199)	(0.8091)
Fjobother	-0.0836	-0.6197
	(0.8609)	(0.5451)
Fjobservices	-0.3221	-0.4658
	(0.5141)	(0.6597)
Fjobteacher	0.1134	-1.3262