Student Final Math Grade Regression Analysis

AUTHOR
Zeb Sorenson

Abstract

This report aims to see how various factors of a student's life can affect their grades in math. We want to see which variable(s) influence the students' final math scores in the most and how using variables such as "Study Time" as continuous vs. categorical will affect our model.

While the analysis can definitely be improved with techniques we haven't yet covered, our findings include the fact that previous class failures was the variable with the most effect on final grade and using variables like "Study Time" as categorical will provide the most accurate models.

1 Problem and Motivation

We obtained the data from the UC Irvine Machine Learning Repository. While the data is collected from a specific school, the inferences and analysis on this data set can give insights into how to help students from all walks of life. With this analysis, school systems, teachers, parents, and students themselves can see what area(s) the student can improve in to receive better grades. In addition, if the factors are mainly outside of school, the family or school system can find ways to improve the quality of life for the student.

Although we may find helpful information beyond this, our two main questions of interest with this analysis are, want to study are the effect on the final math grade when study time is continuous vs. categorical to see if there are any special trends and the effect of number of previous class failures on final math grade.

We have also induced the file student.txt which gives an explanation for all 33 variables in the dataset and their data types.

1.1 Data Description

Please see the attached student.txt file for information regarding each variable. This comes from the UC Irvine Machine learning repository. We will discuss relevant variables further into our analysis.

1.2 Questions of Interest

How does looking at some of the variables as numbers vs categories (continuous vs categorical) affect the models and predictions we make? We specifically looked at the study time variable to answer this question.

What effect does past class failures have on the current final math grade? (We will be using the failures variables for this question)

We can see below in section 1.3 with box_higher that students that wish to pursue higher education are earning a significantly higher final math grade than those that do not. This is what motivates us to investigate if this variable will be significant in our analysis.

1.3 Regression Methods

For each question of interest listed in Section 2, describe what regression models, techniques, and tools you will use to fully answer it. These should include any plots you will use for exploratory analysis or diagnostic checks, as well as methods of inference you will use.

Reading in the data and setting needed variables to factors

```
# Read in data
math <- read_csv2("student-mat.csv")
# Convert Needed Variables to Factors
columns_to_factor <- c("school", "sex", "address", "famsize", "Pstatus", "Medu", "Fedu", "Mjob", "Fjob", "reason", "
math[columns_to_factor] <- lapply(math[columns_to_factor], factor)

# Removing G1 and G2 so we can focus just on final grade
math <- math[, -c(31, 32)]
math <- as.data.frame(math)
```

Relevant EDA

In our EDA we found that there seems to be a significant relationship between previous class failures and final math score, which is why we made the effect of failures one of our primary questions.

```
# Exploratory Scatterplot of Failures
failures_plot <- ggplot(math, aes(x = failures, y = G3)) +
  geom_point() +
  labs(
    title = "Failures and Final Math Score",
    x = "Failures",
    y = "Final Math Score out of 20",
  ) +
  geom_smooth(mapping = aes(x = failures, y = G3),
              method = "lm",
              se = FALSE)
```

As you can see, there is a clear negative trend between failures and final math score, with students failing 3+ classes not scoring higher than 10, or 50% of the maximum possible score.

Model Selection

We went through a number of model selection processes (Which can be found in the Appendix), before deciding to move forward with the model produced with LASSO. Note, in order to save space, we have ommited the output of this model. You may see it by removing the eval=FALSE in the R code. In addition, we used the minimum lambda value instead of one standard error away because the sparser model only gave us one variable and not much predictive power.

```
base_mod <- lm(G3 ~ 1, data = math) # Intercept only model
full_mod <- lm(G3 ~ ., data = math) # Full model

math_y <- math[, 31]
math_x_2 <- model.matrix(full_mod) # turning full model into a matrix with base cases

set.seed(50)
math_lasso_cv <- cv.glmnet(x = math_x_2, y = math_y, type.measure = "mse", alpha = 1)
math_lasso_cv$lambda.min
coef(math_lasso_cv, s = "lambda.min") # Coefficients corresponding to min lambda
```

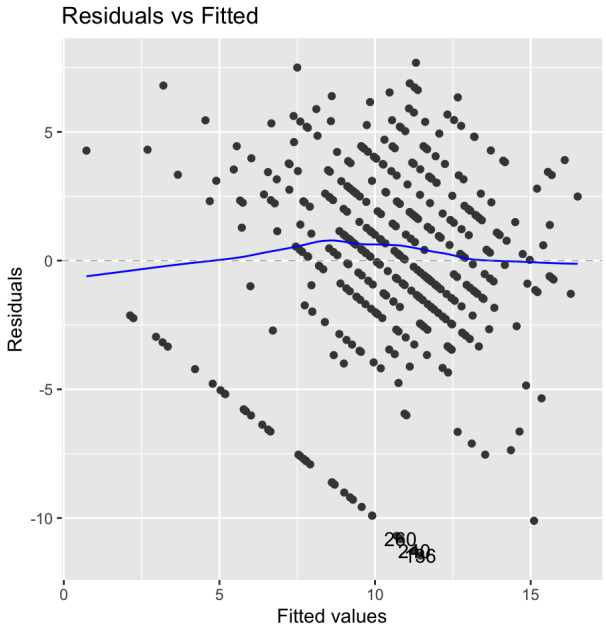
Linear Model

```
# Creating model
math_lm <- lm(G3 ~ sex + age + address + famsize + Pstatus + Medu + Mjob + Fjob + reason + traveltime + studytime +
# Adding the residuals and fits columns
math$residuals <- math_lm$residuals
math$fits <- math_lm$fitted.values
```

Diagnostic Checks

The X's vs Y are linear

```
autoplot(math_lm, which = 1, ncol = 1, nrow = 1) +
  theme(aspect.ratio = 1)
```

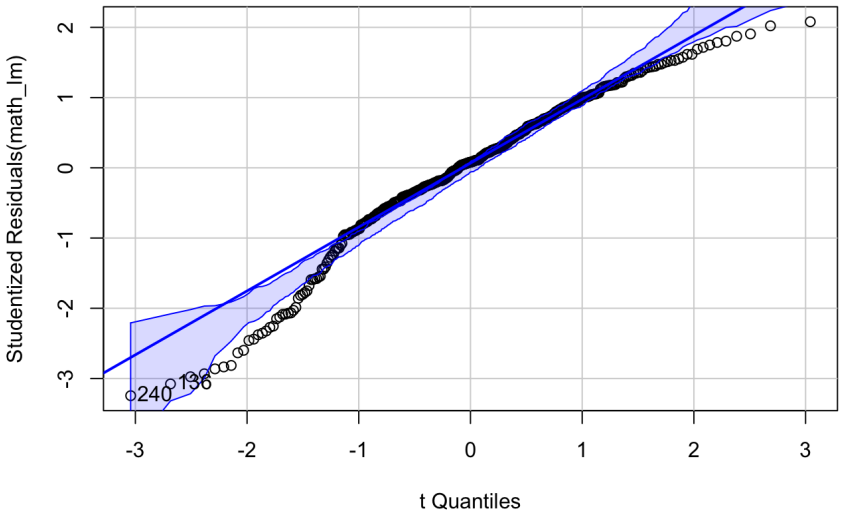


Linearity appears met. While there does seem to be a trend of negative diagonal lines, this is explained by the many categorical variables and the blue line is still horizontal around 0. Because of this, we have only included the residuals vs. fitted plot. You may view the others in the appendix.

The residuals are normally distributed

```
ggplot(data = math) +
  geom_histogram(aes(x = residuals, y = after_stat(density)),
                 binwidth = 2) +
  stat_function(fun = dnorm, color = "red", linewidth = 2,
                args = list(mean = mean(math$residuals),
                            sd = sd(math$residuals)))

shapiro.test(math$residuals)
```

```
qqPlot(math_lm, envelope = .99)
```

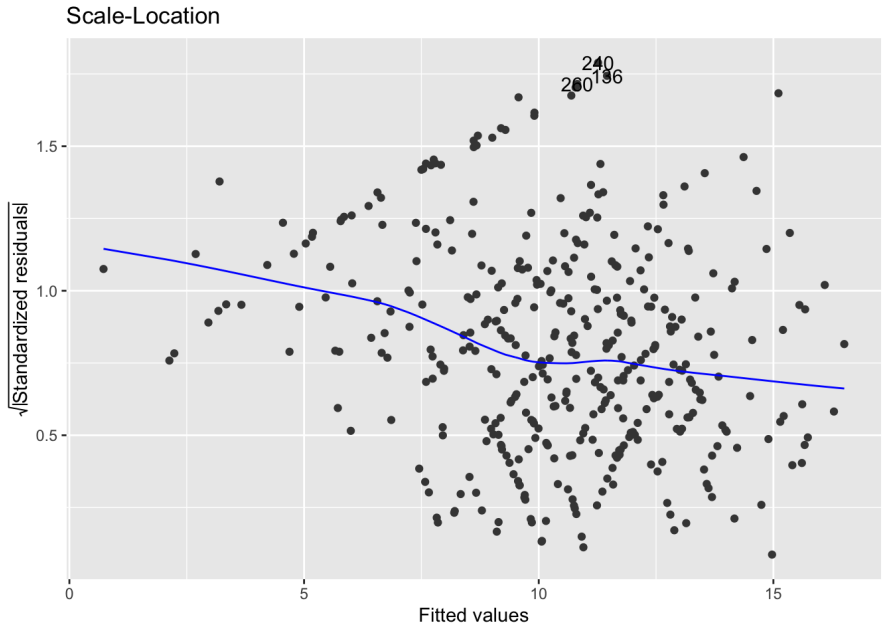



```
[1] 136 240
```

Two out of our three diagnostic checks appear to be sufficient. Our histogram does have slight skewness and there are some values that leave our boundaries in our QQ plot, but nothing appears too extreme. We have chosen to only show the QQ plot as we do reference it again in our transformations. We have omitted the histogram to conserve space, but you may always view these graphs/code output by removing the eval=FALSE from the attached R code. We do also have a significantly low P value of 4.613e-08 which is cause for concern. We will further investigate this assumption to see if improvements can be made via transformations.

The residuals have equal/constant variance across all values of X

```r
autoplot(math_lm, which = 3, nrow = 1, ncol = 1)
```

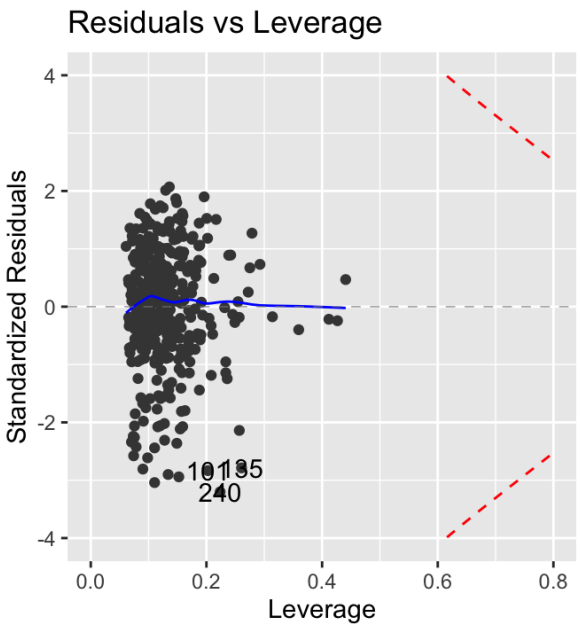


Here we have a consistent spread in both of plots without any blaring patterns. Our line mostly stays straight except for towards the end, however, no major curvature is introduced.

Again, we see some diagonal trends that aren't too extreme, so we will look into transformations.

No influential points

```
cd_cont_pos <- function(leverage, level, model) {sqrt(level*length(coef(model))*(1-leverage)/leverage)}
cd_cont_neg <- function(leverage, level, model) {-cd_cont_pos(leverage, level, model)}

cd_threshold <- 0.5
autoplot(math_lm, which = 5) +
  stat_function(fun = cd_cont_pos,
                args = list(level = cd_threshold, model = math_lm),
                xlim = c(0, 0.8), lty = 2, colour = "red") +
  stat_function(fun = cd_cont_neg,
                args = list(level = cd_threshold, model = math_lm),
                xlim = c(0, 0.8), lty = 2, colour = "red") +
  scale_y_continuous(limits = c(-4, 4))  +
 theme(aspect.ratio = 1)
```



There are no points outside of the 0.5 Cook's Distance limit, so this assumption is met.

Check for extreme multicolinearity

```
vif(math_lm) |> max()
```

```
[1] 5.803096
```

```
vif(math_lm) |> mean()
```

```
[1] 1.790583
```

Our variance inflation factors pass the needed threshold to assume there is no extreme multicolinearity

The residuals are independent.

The data was "collected by using school reports and questionnaires". There isn't a specific element of randomness, but logic tells us that one student's grade shouldn't majorly effect another's, so we will continue with the analysis.

Transfomations

Please see the added code labeled Transformations in the appendix. We wished to investigate if we could improve the distribution of the residuals.

We found that the optimal lambda was roughly .5, suggesting a square root transformation. After performing such transformation on our model, the normality of the residuals in the QQ plot worsened. We also experimented with other transformations and saw no significant improvement. For this reason, we feel confident in continuing with our current model as is for this analysis. Further investigation may be recommended with methods that are beyond Stat 330

2 Analyses, Results, and Interpretation

Study Time Analysis

```
# Creating "new" students to use for predictions (each value is the most common outcome of each variable)
student_1 <- data.frame(sex = "F", age = 17, address = "U", famsize = "GT3", Pstatus = "T", Medu = "4", Mjob = "othe

# Changing studytime to 3
student_2 <- data.frame(sex = "F", age = 17, address = "U", famsize = "GT3", Pstatus = "T", Medu = "4", Mjob = "othe

s1_grade <- predict(math_lm, newdata = student_1, se.fit = TRUE)
s2_grade <- predict(math_lm, newdata = student_2, se.fit = TRUE)
s1_fit <- s1_grade$fit
s2_fit <- s2_grade$fit

s1_fit
```

```
       1
8.509723
```

```
s2_fit
```

```
       1
10.10556
```

```
confint(math_lm, "failures", level=0.95)
```

```
             2.5 %   97.5 %
failures -2.338804 -1.08056
```

Here, we create two student objects, both with all of the coefficients from our linear model. Each variable value is the most common value that variable takes on (for example, females are more common than males in the study, so both example students are female). Student 1 has study time set to 1 (representing study time of < 2 hours) and student 2 has study time set to 3 (representing 5-10 hours).

We can see that student one ends with a final grade of roughly 8.5/20 while student two ends with a final grade of 10.1/20.

In addition, to answer one of our primary questions, we are 95% confident that when the number of previously failed classes increases by one, the final math score out of 20 will decrease between 1.081 and 2.339 points on average (equivalent to ~5-11% decrease out of 100).

Second Linear Model

```
# This model will have "studytime" as continuous, not categorical
math2 <- read_csv2("student-mat.csv")
```

```r
columns_to_factor_2 <- c("school", "sex", "address", "famsize", "Pstatus", "Medu", "Fedu", "Mjob", "Fjob", "reason",
math2[columns_to_factor_2] <- lapply(math2[columns_to_factor_2], factor)
math2 <- math2[, -c(31, 32)]
math2 <- as.data.frame(math2)

math2_lm <- lm(G3 ~ sex + age + address + famsize + Pstatus + Medu + Mjob + Fjob + reason + traveltime + studytime +

student2_1 <- data.frame(sex = "F", age = 17, address = "U", famsize = "GT3", Pstatus = "T", Medu = "4", Mjob = "oth

student2_2 <- data.frame(sex = "F", age = 17, address = "U", famsize = "GT3", Pstatus = "T", Medu = "4", Mjob = "oth

s1_grade2 <- predict(math2_lm, newdata = student2_1, se.fit = TRUE)
s2_grade2 <- predict(math2_lm, newdata = student2_2, se.fit = TRUE)

s1_fit2 <- s1_grade2$fit
s2_fit2 <- s2_grade2$fit

s1_fit2 - s1_fit
```

```
       1
0.1891791
```

```r
s2_fit2 - s2_fit
```

```
        1
-0.5559961
```

Here we continue with our analysis of categorical vs continuous for the study time variable. This process is the same as before, only setting the data as continuous now.

We see that the difference in grade for student 1 from model 2 to model 1 is a 0.1892 point increase, whereas for student 2, it is actually a 0.5560 point decrease.

3 Conclusions

Our two main questions that we wanted to study were the effect on the final math grade when study time is continuous vs. categorical to see if there are any special trends and the effect of number of previous class failures on final math grade. To answer the first question, we made two new "students", who had the most common attribute of each variable in our model, but student one had a study time of 1 (representing study time of < 2 hours) and student two a study time of 3 (representing 5-10 hours) for the categorical model. Then, using a new model with study time as a continuous variable, we compared the differences between the two student "one"s and "two"s. The first difference in final grade of continuous vs categorical, keeping everything the same, was 0.1892 grade points out of 20. The difference for the second student, however, was -0.556 grade points out of 20, or around a 2.5% drop in grade. This means that there is some loss of information when using study time as a continuous variable, probably because in reality, a student in group 3 is studying more than 3 times more than a student in group one, which isn't represented when study time is continuous.

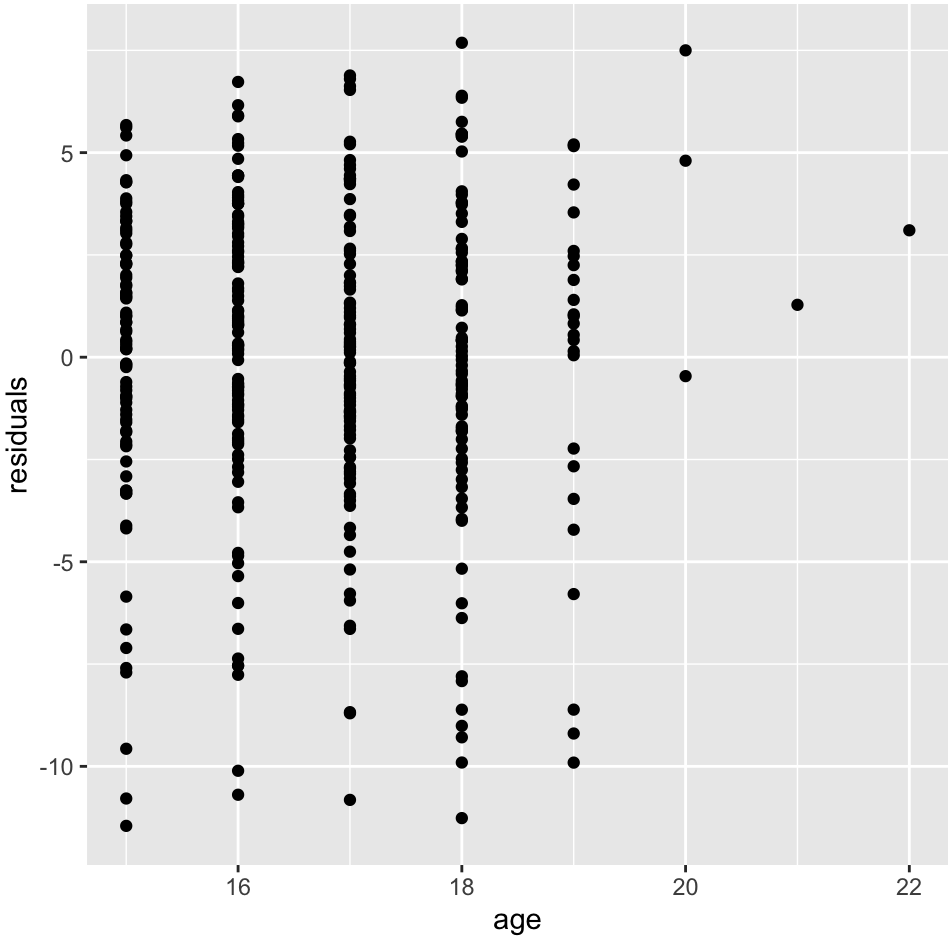
Our second model gives failures a coefficient of -1.7097 and a extremely low p-value of 1.65e-07. This means that holding all else constant, for each additional class failed in the past, a student's final grade will decrease on average by 1.7097 points out of 20, or around an 8.55% drop in grade. Failures was by far the variable with the most effect on final grade. In addition, we are 95% confident that when the number of previously failed classes increases by one, the final math score out of 20 will decrease between 1.081 and 2.339 points on average (equivalent to ~5-11% decrease out of 100).

There are some weakness in the model, however. To begin, the adjusted R-squared value is 0.2398, which is pretty low and means there is a lot of noise in the data, so we don't have great predictability. Also, the residuals could follow a normal curve a little better, but a log transformation was not possible because some values in our response were 0 and the lambda value from the BoxCox transformation was ~0.5, but using a square root transformation made the distribution of the residuals even worse. Thus, we proceeded without transformations because all other assumptions were met. While there are some weaknesses, the significant variables in our model do make logical sense, so we are confident there is value to the model.

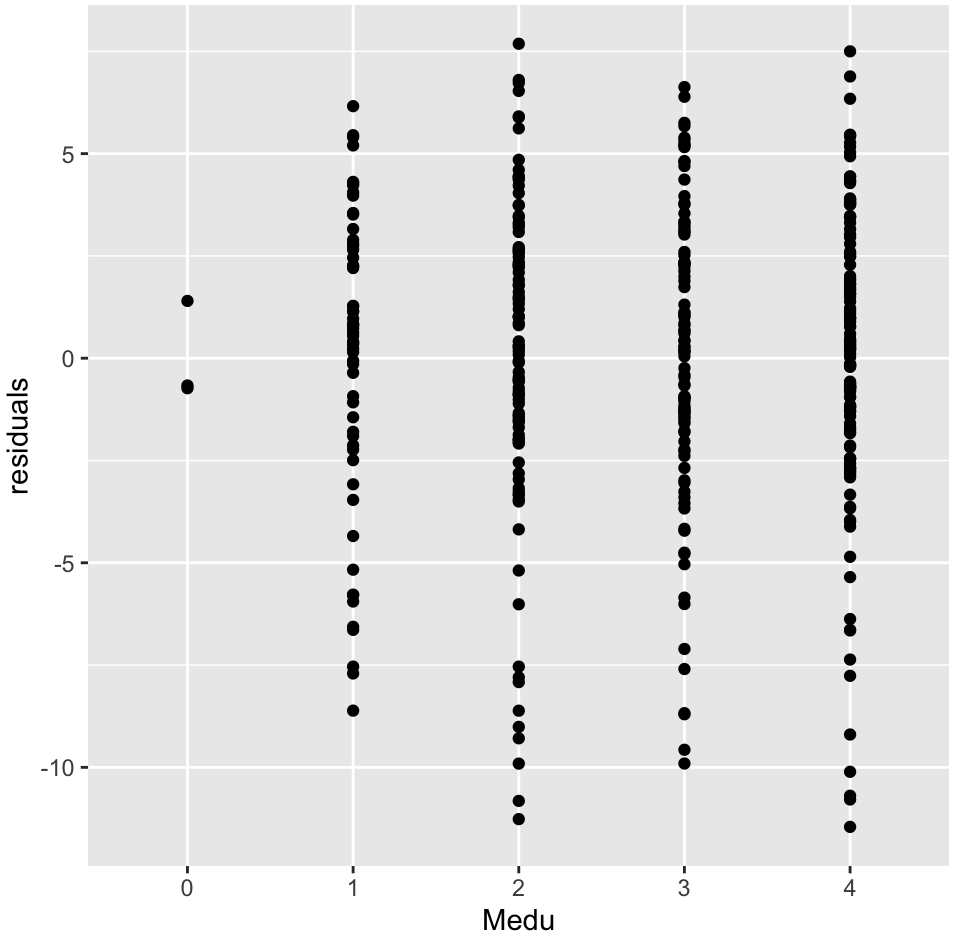
APPENDIX

Other Diagnostic Checks

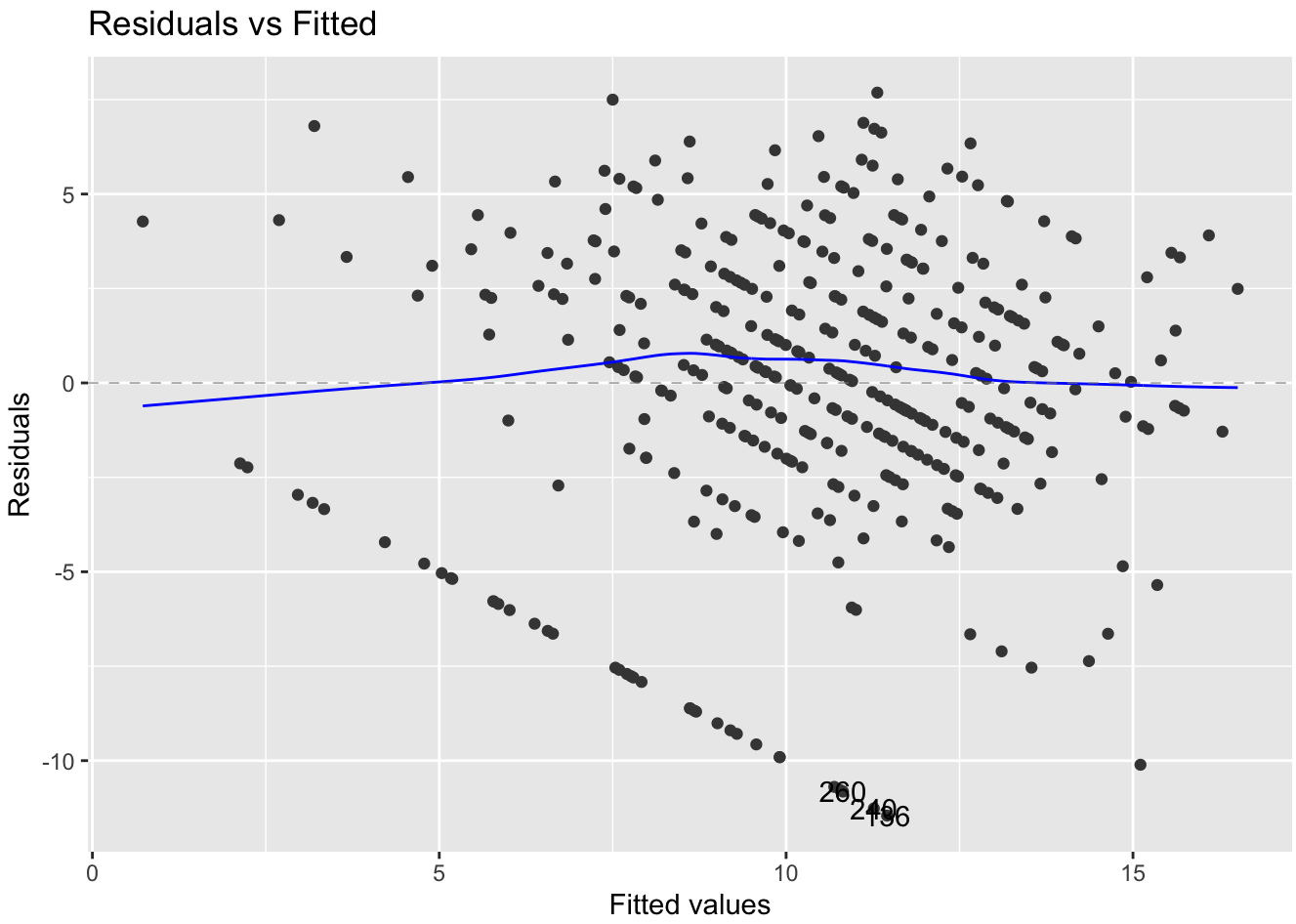
```
# Linearity (One continuous variable and one categorical to not overwhelm the page count)
resid_vs_age <- ggplot(data = math) +
  geom_point(mapping = aes(x = age, y = residuals)) +
  theme(aspect.ratio = 1)
resid_vs_age
```




```
resid_vs_Medu <- ggplot(data = math) +
  geom_point(mapping = aes(x = Medu, y = residuals)) +
  theme(aspect.ratio = 1)
resid_vs_Medu
```




```r
# Equal/Constant Variance
autoplot(math_lm, which = 1, nrow = 1, ncol = 1)
```



Model Selection

We went through a number of different model selection processes, most of which produced models that we did not feel confident in moving forward with until we decided on LASSO. We ran both BIC and AIC. you may modify the eval=FALSE in our R code in order to see the results of these model selection processes. Which we have chosen to exclude to save space.

```
base_mod <- lm(G3 ~ 1, data = math) # Intercept only model
full_mod <- lm(G3 ~ ., data = math)

math_int_lm <- lm(G3 ~ .^2, data = math)
summary(math_int_lm)


forw_AIC <- step(base_mod,
     direction = "forward",
     scope=list(lower= base_mod, upper= full_mod))
summary(forw_AIC)

back_BIC <- step(full_mod,
     direction = "backward",
     k = log(nrow(math)),
     scope=list(lower= base_mod, upper= full_mod))
summary(back_BIC)

step_BIC_base <- step(base_mod,
     direction = "both",
     k = log(nrow(math)),
     scope=list(lower= base_mod, upper= full_mod))
summary(step_BIC_base)
```

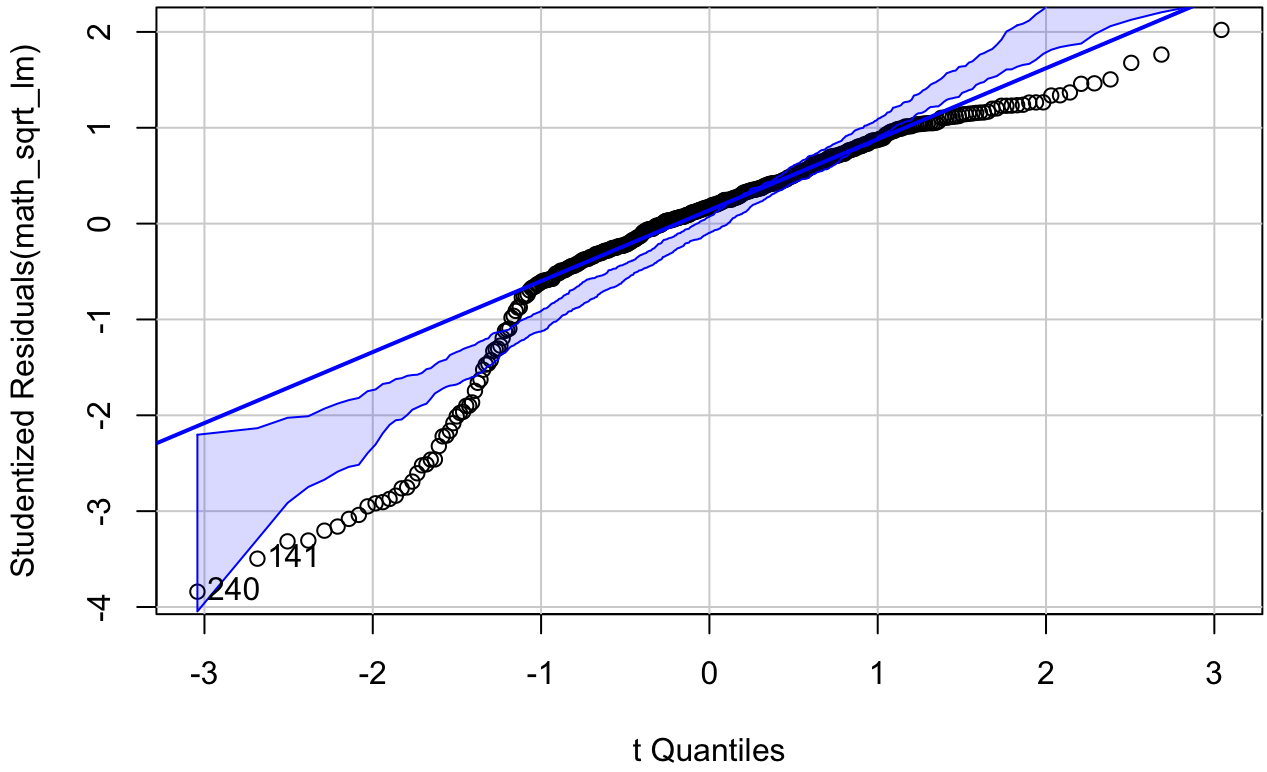
Transformations

```
math_sqrt_lm <- lm(sqrt(G3) ~ sex + age + address + famsize + Pstatus + Medu + Mjob + Fjob + reason + traveltime + s

summary(math_sqrt_lm)
```

```
Call:
lm(formula = sqrt(G3) ~ sex + age + address + famsize + Pstatus +
    Medu + Mjob + Fjob + reason + traveltime + studytime + failures +
    schoolsup + famsup + paid + higher + internet + romantic +
    freetime + goout + Dalc + Walc + absences, data = math)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2202 -0.2880  0.1565  0.5984  1.7509

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.793e+00  1.127e+00   3.366 0.000848 ***
sexM              2.233e-01  1.201e-01   1.859 0.063814 .
age              -5.121e-02  4.628e-02  -1.107 0.269270
addressU          7.606e-02  1.375e-01   0.553 0.580505
famsizeLE3        2.175e-01  1.186e-01   1.833 0.067640 .
PstatusT         -7.603e-02  1.750e-01  -0.434 0.664337
Medu1            -1.046e+00  5.991e-01  -1.746 0.081690 .
Medu2            -1.075e+00  6.013e-01  -1.788 0.074651 .
Medu3            -9.334e-01  6.052e-01  -1.542 0.123925
Medu4            -6.763e-01  6.184e-01  -1.094 0.274862
Mjobhealth       -6.568e-02  2.687e-01  -0.244 0.807010
Mjobother        -1.946e-02  1.696e-01  -0.115 0.908725
Mjobservices      8.303e-02  1.910e-01   0.435 0.664066
Mjobteacher      -4.134e-01  2.504e-01  -1.651 0.099692 .
Fjobhealth        2.472e-01  3.412e-01   0.724 0.469265
Fjobother         1.726e-02  2.452e-01   0.070 0.943927
Fjobservices      6.463e-02  2.534e-01   0.255 0.798826
Fjobteacher       1.894e-01  3.054e-01   0.620 0.535596
reasonhome       -1.235e-02  1.327e-01  -0.093 0.925922
reasonother       2.854e-01  1.952e-01   1.462 0.144600
reasonreputation  1.254e-01  1.357e-01   0.924 0.356191
traveltime2      -1.117e-01  1.193e-01  -0.936 0.349676
traveltime3       7.396e-02  2.342e-01   0.316 0.752393
traveltime4       3.664e-02  3.966e-01   0.092 0.926447
studytime2        1.461e-01  1.328e-01   1.100 0.272196
studytime3        3.457e-01  1.834e-01   1.885 0.060331 .
studytime4        3.785e-02  2.338e-01   0.162 0.871481
failures         -3.897e-01  7.772e-02  -5.015 8.52e-07 ***
schoolsupyes     -2.214e-02  1.609e-01  -0.138 0.890636
famsupyes        -1.919e-01  1.143e-01  -1.679 0.094048 .
paidyes           2.097e-01  1.164e-01   1.801 0.072523 .
higheryes         2.857e-01  2.606e-01   1.096 0.273789
internetyes       1.048e-01  1.486e-01   0.705 0.481171
romanticyes      -3.603e-01  1.134e-01  -3.177 0.001624 **
freetime2         4.045e-01  2.758e-01   1.467 0.143315
freetime3         7.497e-02  2.611e-01   0.287 0.774203
freetime4         3.346e-01  2.699e-01   1.240 0.215996
freetime5         6.960e-01  3.028e-01   2.299 0.022110 *
goout2            3.770e-01  2.390e-01   1.577 0.115634
goout3            3.010e-01  2.375e-01   1.267 0.205923
goout4            4.219e-05  2.492e-01   0.000 0.999865
goout5           -1.787e-01  2.693e-01  -0.664 0.507403
```

```
Dalc2           -3.384e-01  1.567e-01  -2.160 0.031451 *
Dalc3           -9.929e-02  2.460e-01  -0.404 0.686783
Dalc4           -5.194e-01  3.847e-01  -1.350 0.177878
Dalc5           -2.263e-01  4.437e-01  -0.510 0.610389
Walc2           -5.929e-02  1.474e-01  -0.402 0.687694
Walc3            2.707e-01  1.643e-01   1.648 0.100293
Walc4            1.081e-01  2.095e-01   0.516 0.606250
Walc5            5.720e-01  3.171e-01   1.804 0.072073 .
absences         2.464e-02  6.891e-03   3.576 0.000399 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9706 on 344 degrees of freedom
Multiple R-squared:  0.3119,    Adjusted R-squared:  0.2118
F-statistic: 3.118 on 50 and 344 DF,  p-value: 4.72e-10
```

```r
qqPlot(math_sqrt_lm, envelope = .99)
```



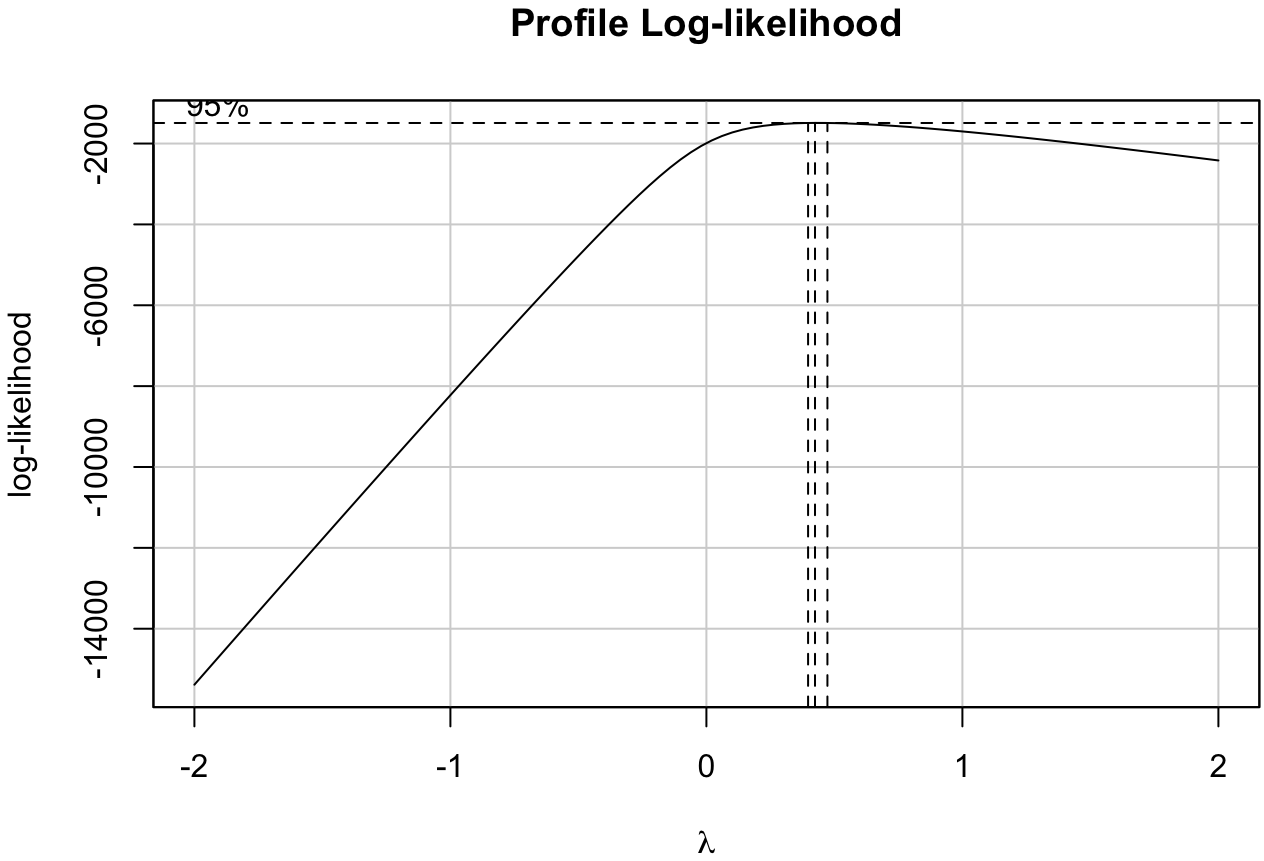

```
[1] 141 240
```

```r
shapiro.test(math$residuals)
```

```
    Shapiro-Wilk normality test

data:  math$residuals
W = 0.96528, p-value = 4.613e-08
```

```r
math$G3_positive <- math$G3 + 0.00000001
lm_model <- lm(G3_positive ~ sex + age + address + famsize + Pstatus + Medu + Mjob + Fjob + reason + traveltime + st
```

```
bc_math <- boxCox(lm_model)
```



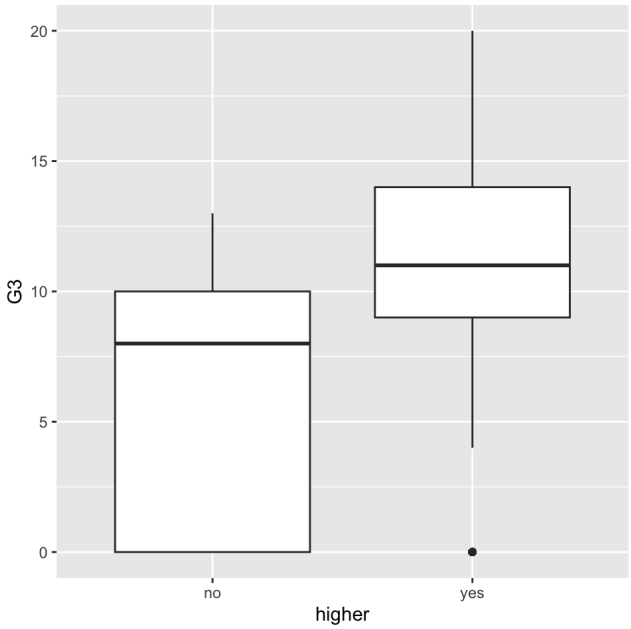

```
#bc_math <-  boxCox(math_lm)
lambda.opt = bc_math$x[which.max(bc_math$y)]
lambda.opt
```

```
[1] 0.4242424
```

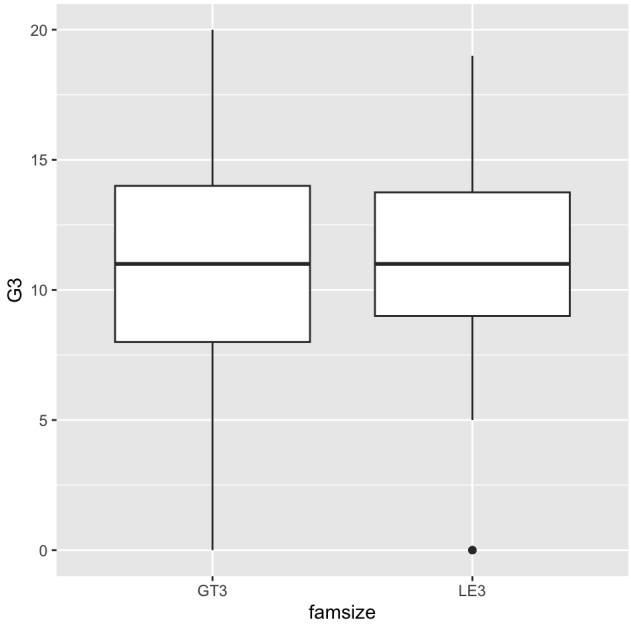
EDA boxplots

```
box_higher<- ggplot(data = math) +

  geom_boxplot(mapping = aes(x = higher, y = G3)) +

  theme(aspect.ratio = 1)


box_famsize<- ggplot(data = math) +

  geom_boxplot(mapping = aes(x = famsize, y = G3)) +

  theme(aspect.ratio = 1)

box_pStatus <- ggplot(data = math) +

  geom_boxplot(mapping = aes(x = Pstatus, y = G3)) +
```

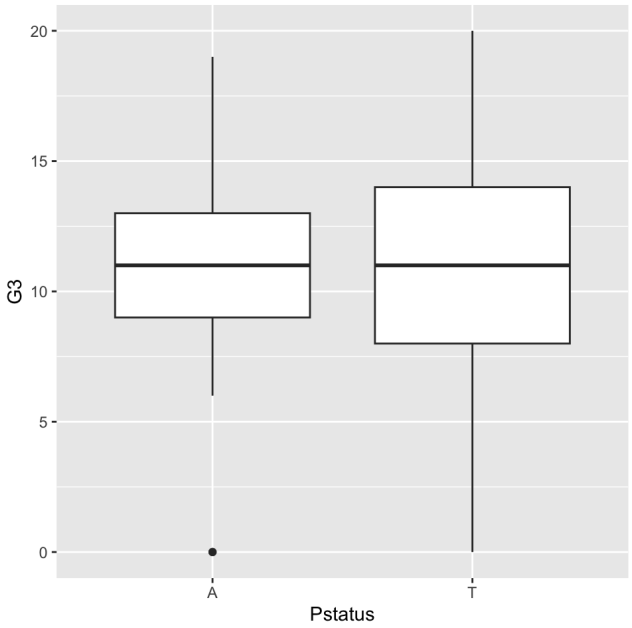
```
  theme(aspect.ratio = 1)

box_Mjob <- ggplot(data = math) +

  geom_boxplot(mapping = aes(x = Mjob, y = G3)) +

  theme(aspect.ratio = 1)

box_Fjob <- ggplot(data = math) +

  geom_boxplot(mapping = aes(x = Fjob, y = G3)) +

  theme(aspect.ratio = 1)

box_higher
```



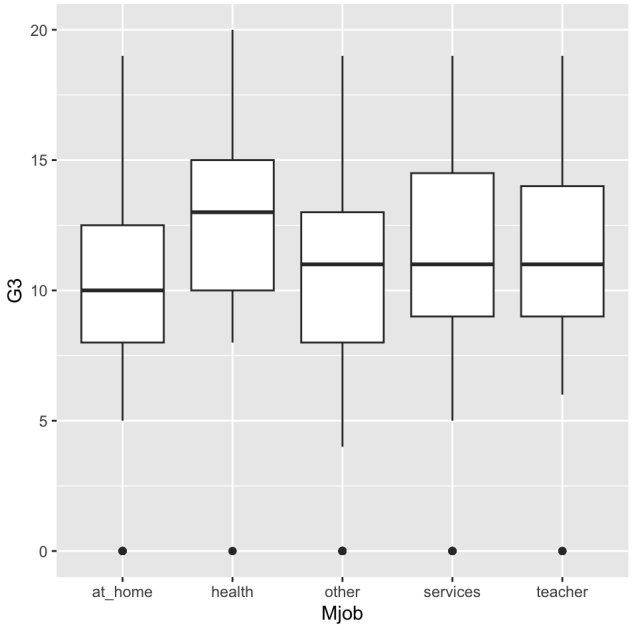

```
box_famsize
```



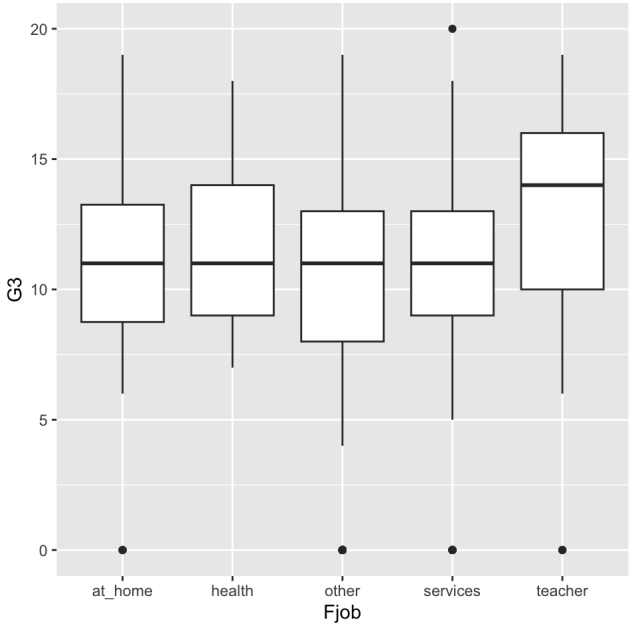
```
box_pStatus
```




```
box_Mjob
```




```
box_Fjob
```


20
15
10
5
0
G3
at_home
health
other
services
teacher
Fjob