ETC 2420/5242 Lab 6 2016

Di Cook

SOLUTION

A dictionary of variables that we will use further (in addition to the `math` variable we just created) is as follows:

Variable name	Description	Coding
ST04Q01	Gender	1=Female, 2=Male
ST06Q01	Age when started school	Actual age, 9997-9999 indicate missing values
ST57Q01	Out-of-School Study Time - Homework	Hours per week, 9997-9999 indicate missing values
ST15Q01	Mother Current Job Status	1=Full-time, 2=Part-time, 3=Not working, but looking for a job, 4=Other (inc stay-at-home), 7-9 indicate missing values
ST19Q01	Father Current Job Status	1=Full-time, 2=Part-time, 3=Not working, but looking for a job, 4=Other (inc stay-at-home), 7-9 indicate missing values
ST26Q01	Possessions - desk	1=Yes, 2=No, 7-9 indicate missing values
ST26Q02	Possessions - own room	1=Yes, 2=No, 7-9 indicate missing values
ST26Q04	Possessions - computer	1=Yes, 2=No, 7-9 indicate missing values
ST26Q06	Possessions - Internet	1=Yes, 2=No, 7-9 indicate missing values
ST27Q02	How many - televisions	1=None, 2=One, 3=Two, 4=Three or more, 7-9 indicate missing values
ST28Q01	How many books at home	1=0-10, 2=11-25, 3=26-100, 4=101-200, 5=201-500, 6=More than 500, 7-9 indicate missing values
SENWGT_STU	Weight	Reflects how the student represents other students in Australia based on socioeconomic and demographic characteristics

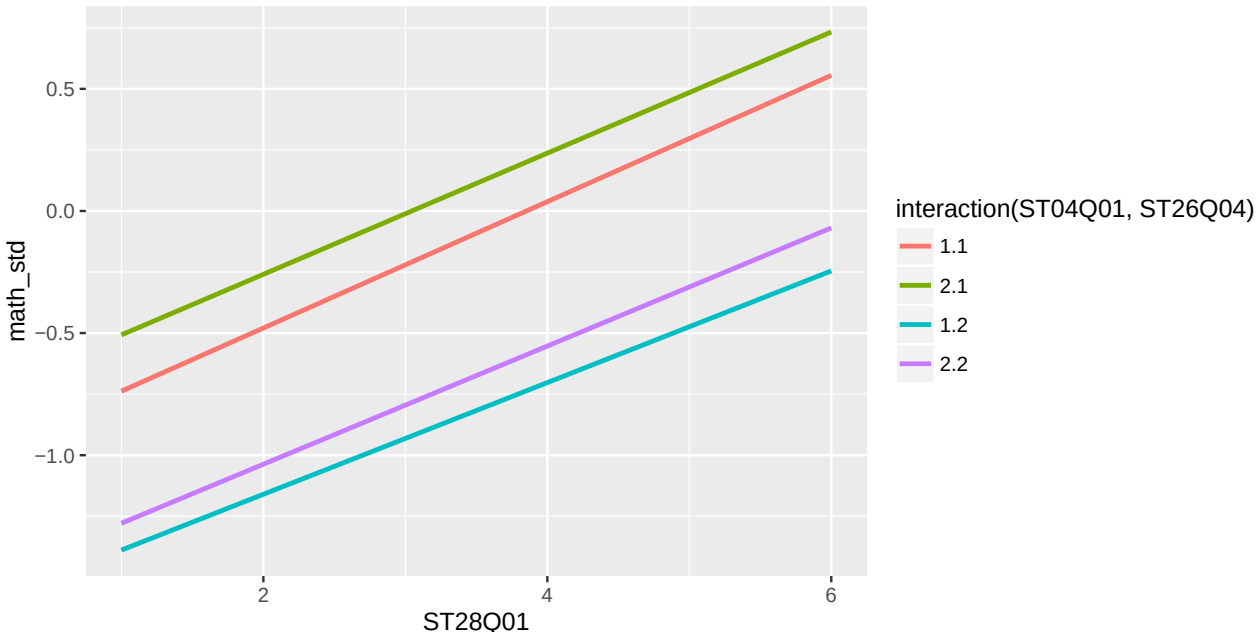
Model building will be done using:

- Response: `math` (standardised)
- Explanatory variables: `ST04Q01, ST06Q01, ST15Q01, ST19Q01, ST26Q01, ST26Q02, ST26Q04, ST26Q06, ST27Q02, ST28Q01`. Age at school start will be set to be 0 meaning age 4.

Some variables need to be treated as categorical variables, so it is best if they are forced to be factors before modeling:

Test the model fitting, by fitting a model for math against gender, books at home and whether they own a computer.

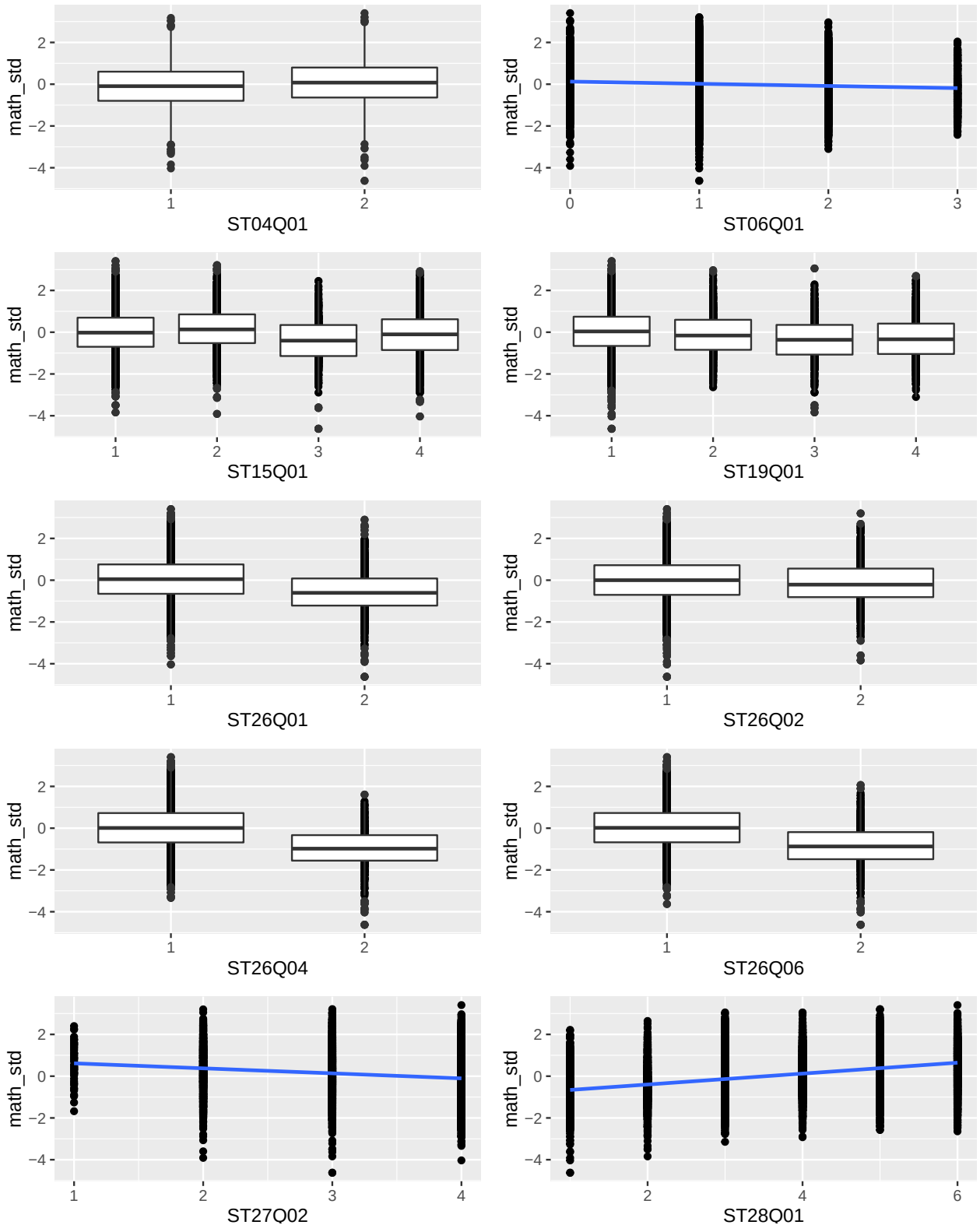
Sketch what this model looks like.



Question 1

- Make plots of the response variable `math` against each of the possible explanatory variables.
- Which variables look like they should be most important for predicting math score?

```
p1 <- ggplot(aus_nomiss, aes(x=ST04Q01, y=math_std)) + geom_boxplot()
p2 <- ggplot(aus_nomiss, aes(x=ST06Q01, y=math_std)) +
  geom_point() + geom_smooth(method="lm", se=FALSE)
p3 <- ggplot(aus_nomiss, aes(x=ST15Q01, y=math_std)) +
  geom_point() + geom_boxplot()
p4 <- ggplot(aus_nomiss, aes(x=ST19Q01, y=math_std)) +
  geom_point() + geom_boxplot()
p5 <- ggplot(aus_nomiss, aes(x=ST26Q01, y=math_std)) +
  geom_point() + geom_boxplot()
p6 <- ggplot(aus_nomiss, aes(x=ST26Q02, y=math_std)) +
  geom_point() + geom_boxplot()
p7 <- ggplot(aus_nomiss, aes(x=ST26Q04, y=math_std)) +
  geom_point() + geom_boxplot()
p8 <- ggplot(aus_nomiss, aes(x=ST26Q06, y=math_std)) +
  geom_point() + geom_boxplot()
p9 <- ggplot(aus_nomiss, aes(x=ST27Q02, y=math_std)) +
  geom_point() + geom_smooth(method="lm", se=FALSE)
p10 <- ggplot(aus_nomiss, aes(x=ST28Q01, y=math_std)) +
  geom_point() + geom_smooth(method="lm", se=FALSE)
library("gridExtra")
grid.arrange(p1, p2, p3, p4, p5, p6, p7, p8, p9, p10, ncol=2)
```



All of the variables look like they are associated with math scores, except for ST26Q02 (Possessions - own room)

Question 2

- Fit the weighted multiple regression model to all the explanatory variables.

```
#
# Call:
# glm(formula = math_std ~ ST04Q01 + ST06Q01 + ST15Q01 + ST19Q01 +
#     ST26Q01 + ST26Q02 + ST26Q04 + ST26Q06 + ST27Q02 + ST28Q01,
#     data = aus_nomiss, weights = SENWGT_STU)
#
# Deviance Residuals:
#      Min        1Q    Median        3Q       Max
# -1.14733  -0.14537  -0.01362   0.13083   0.93207
#
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)  0.16453    0.05643   2.916  0.00356 **
# ST04Q012     0.21914    0.01649  13.288  < 2e-16 ***
# ST06Q01     -0.10698    0.01215  -8.807  < 2e-16 ***
# ST15Q012     0.08964    0.02000   4.482 7.45e-06 ***
# ST15Q013    -0.21109    0.04432  -4.763 1.93e-06 ***
# ST15Q014    -0.01979    0.02163  -0.915  0.36013
# ST19Q012    -0.07725    0.03460  -2.233  0.02559 *
# ST19Q013    -0.08302    0.05081  -1.634  0.10231
# ST19Q014    -0.17022    0.03404  -5.000 5.80e-07 ***
# ST26Q012    -0.28770    0.03147  -9.143  < 2e-16 ***
# ST26Q022     0.02074    0.03238   0.640  0.52199
# ST26Q042    -0.37038    0.06673  -5.551 2.90e-08 ***
# ST26Q062    -0.40925    0.05864  -6.979 3.13e-12 ***
# ST27Q02     -0.21696    0.01246 -17.416  < 2e-16 ***
# ST28Q01      0.21307    0.00606  35.160  < 2e-16 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# (Dispersion parameter for gaussian family taken to be 0.05668027)
#
#     Null deviance: 819.22  on 11917  degrees of freedom
# Residual deviance: 674.67  on 11903  degrees of freedom
# AIC: 34332
#
# Number of Fisher Scoring iterations: 2
```

- Summarise the coefficients for the model fit.

E	stimate Std	. Error	t value Pr	(>|t|)
(Intercept)	0.1645311	0.0564323	2.9155452	0.0035573
ST04Q012	0.2191394	0.0164917	13.2878772	0.0000000
ST06Q01	-0.1069802	0.0121467	-8.8073534	0.0000000
ST15Q012	0.0896426	0.0199988	4.4824046	0.0000074
ST15Q013	-0.2110885	0.0443179	-4.7630588	0.0000019
ST15Q014	-0.0197907	0.0216257	-0.9151488	0.3601320
ST19Q012	-0.0772524	0.0346001	-2.2327213	0.0255859
ST19Q013	-0.0830172	0.0508098	-1.6338801	0.1023105

E	stimate Std	. Error	t value Pr	(>|t|)
ST19Q014	-0.1702251	0.0340427	-5.0003430	0.0000006
ST26Q012	-0.2877042	0.0314671	-9.1430291	0.0000000
ST26Q022	0.0207360	0.0323851	0.6402923	0.5219949
ST26Q042	-0.3703821	0.0667246	-5.5509093	0.0000000
ST26Q062	-0.4092519	0.0586416	-6.9788713	0.0000000
ST27Q02	-0.2169571	0.0124572	-17.4161820	0.0000000
ST28Q01	0.2130722	0.0060601	35.1597246	0.0000000

- Not all variables are significant in the model. What variables can be dropped? Re-fit the model with this subset.

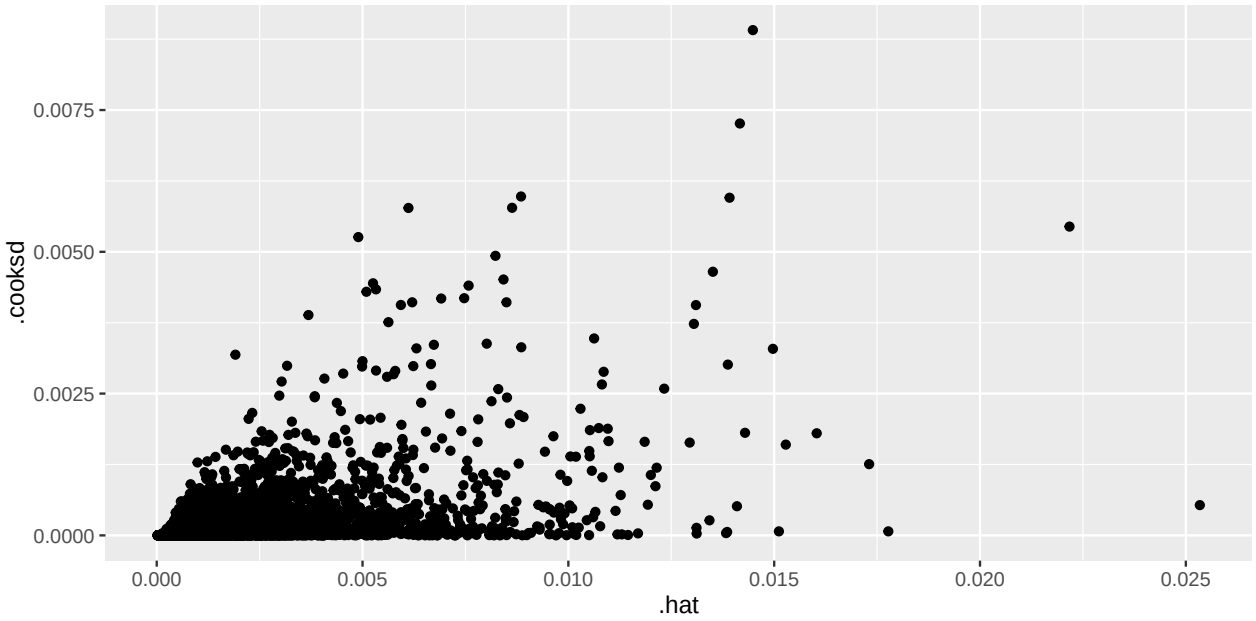
ST26Q02 (Possessions - own room) does not significantly add to the model. We re-fit the model without this variable. (You might think that ST15Q014, ST19Q013 should also be dropped, but these are dummy variables representing different levels of ST15Q01 and ST19Q01, and other levels have significant contributions. You could collapse the category with the base level for each variable, though, and re-fit.)

```
#
# Call:
# glm(formula = math_std ~ ST04Q01 + ST06Q01 + ST15Q01 + ST19Q01 +
#     ST26Q01 + ST26Q04 + ST26Q06 + ST27Q02 + ST28Q01, data = aus_nomiss,
#     weights = SENWGT_STU)
#
# Deviance Residuals:
#      Min        1Q    Median        3Q       Max
# -1.14769  -0.14545  -0.01388   0.13061   0.93154
#
# Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
# (Intercept)  0.168205   0.056139   2.996  0.00274 **
# ST04Q012     0.219131   0.016491  13.288  < 2e-16 ***
# ST06Q01     -0.107002   0.012146  -8.809  < 2e-16 ***
# ST15Q012     0.089772   0.019997   4.489 7.22e-06 ***
# ST15Q013    -0.210685   0.044312  -4.755 2.01e-06 ***
# ST15Q014    -0.018866   0.021577  -0.874  0.38194
# ST19Q012    -0.076803   0.034592  -2.220  0.02642 *
# ST19Q013    -0.081732   0.050769  -1.610  0.10745
# ST19Q014    -0.169310   0.034012  -4.978 6.52e-07 ***
# ST26Q012    -0.286384   0.031399  -9.121  < 2e-16 ***
# ST26Q042    -0.370127   0.066722  -5.547 2.96e-08 ***
# ST26Q062    -0.408397   0.058625  -6.966 3.43e-12 ***
# ST27Q02     -0.217575   0.012419 -17.519  < 2e-16 ***
# ST28Q01      0.212939   0.006056  35.159  < 2e-16 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# (Dispersion parameter for gaussian family taken to be 0.05667747)
#
#     Null deviance: 819.22  on 11917  degrees of freedom
# Residual deviance: 674.69  on 11904  degrees of freedom
# AIC: 34331
```

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

Question 3

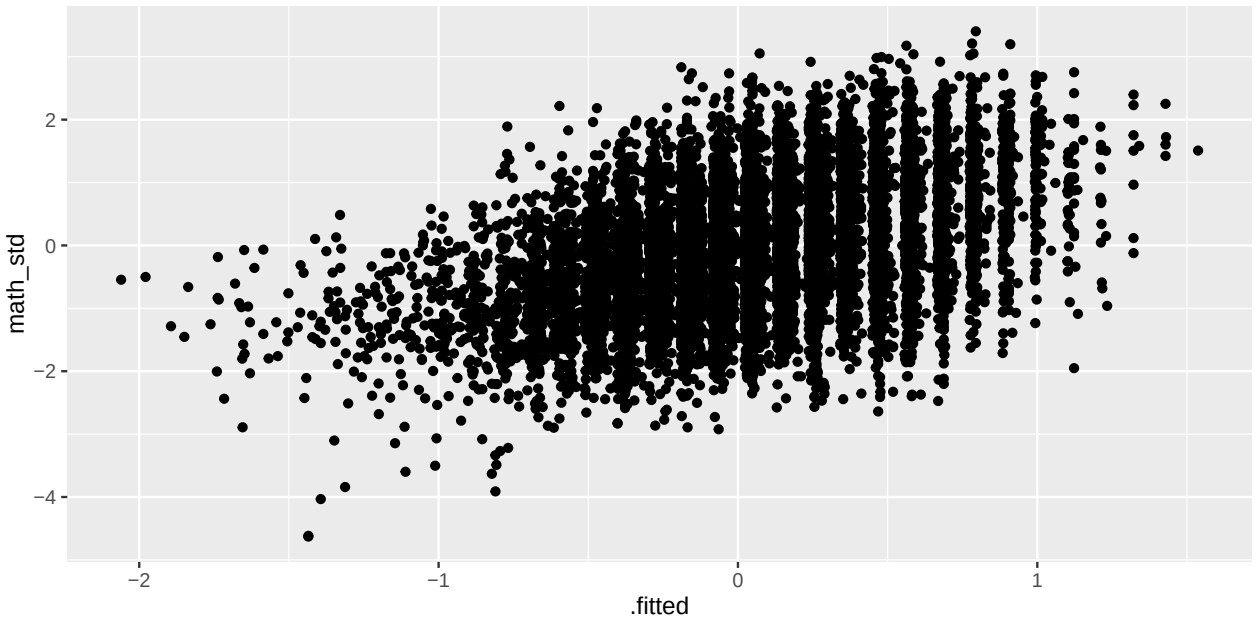
- Compute the leverage and influence statistics.



- What value would be considered to tbe the cutoff for considering a case to have high leverage? `2*p/n=2*13/11918=0.00218`
- How many cases have high influence? 1669 `cases would be considered to have high leverage. No cases would be considered to have high influence, the CooksD values are all very small.`

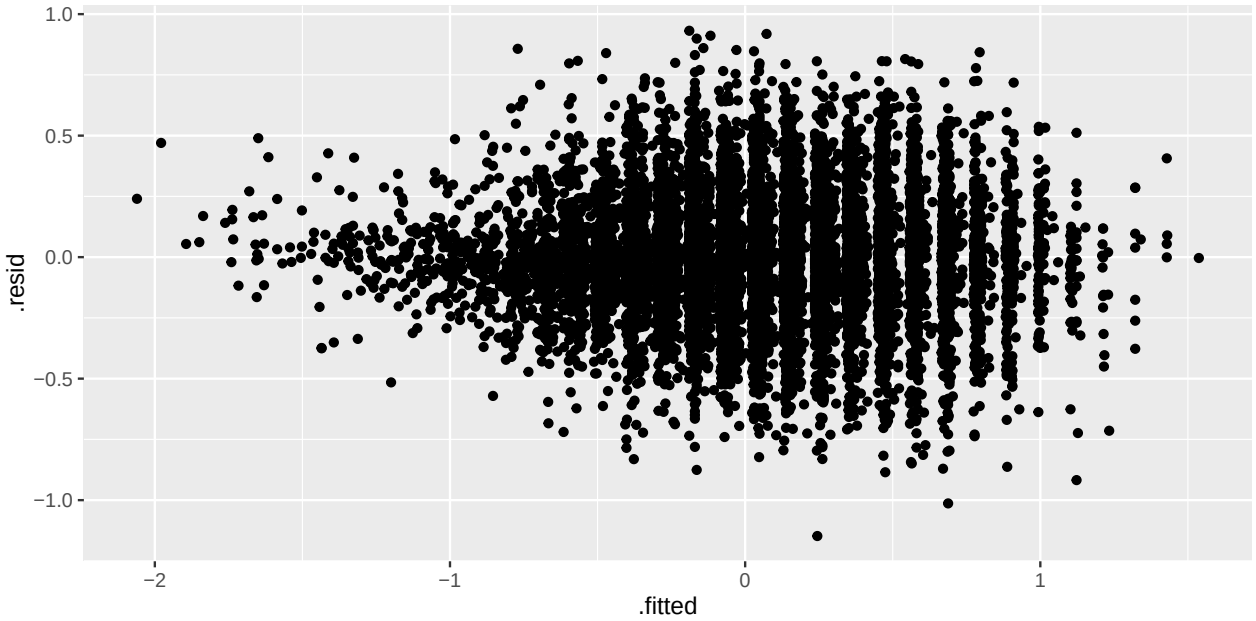
Question 4

- Plot the observed vs fitted values. How good is the model for predicting math score? (Is it weak, moderate or strong?)



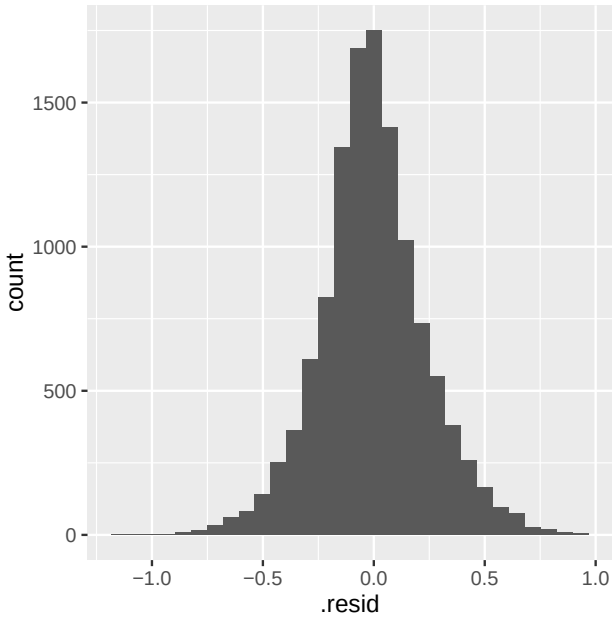

```
The model moderately explains math scores. There is a lot of variation in the scores that
is clearly not explained by the predictors.
```

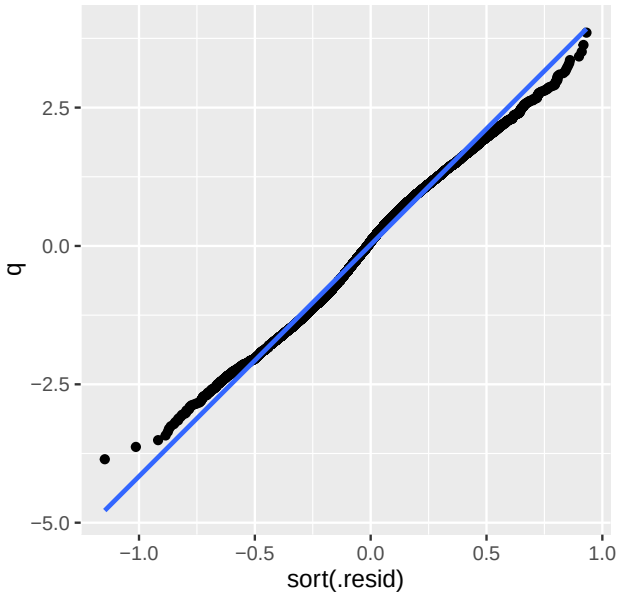
- Plot residuals vs fitted. What do you learn about the model fit by looking at this plot?




```
There is some heteroskedasticity. The variation in residuals at smaller fitted values is
smaller than at larger values.
```

- Make a histogram of residuals, and a qqplot (normal probability plot). Do these look like a sample from a normal model?






```
Doesn't look entirely normal. The tails are a bit long. Small residuals are higher than
expected and larger residuals are lower than expected.
```

Question 5

Compute the variance inflation factors. Do these indicate collinearity between predictors that needs to be addressed?

```
#             GVIF Df GVIF^(1/(2*Df))
# ST04Q01 1.011719  1        1.005842
# ST06Q01 1.010130  1        1.005052
# ST15Q01 1.035484  3        1.005828
# ST19Q01 1.042007  3        1.006882
# ST26Q01 1.060148  1        1.029635
# ST26Q04 1.138064  1        1.066801
# ST26Q06 1.144925  1        1.070012
# ST27Q02 1.025098  1        1.012471
# ST28Q01 1.057237  1        1.028220
```

```
There is no multicollinearity problem. All the VIFs are low.
```

Question 6

Interpret the model:

- For male students how much does math score increase or decrease on average?
- For each year delayed starting school what happens to average math score?
- For a student who's mother is part-time, looking for work or other, how does the average math score change?

- ....
- The base model is for female, started school at age 4, both parents worked full-time, had a desk, computer and internet, no TVs in the household, and 0-10 books. These students had an average standardised math score of 0.168205
- A male student's, with otherwise the same demographics, had an average standardised math score of 0.168205 + 0.219131 = 0.387336
- For each extra year delay in starting school average standardised math score decreased by 0.107002
- A student who's Mum worked part-time saw an increased average of 0.089772, but who's Mum was looking for work saw a decrease of 0.210685. No difference for a Mum who was in the other category.
- A student who's Dad was in any other category than full-time work saw a decrease in average math score of 0.076803, 0.081732, 0.169310 respectively.
- Not having a desk, computer of internet decreased the average by 0.286384, 0.370127, 0.408397 respectively
- Each additional TV in the household, up to 3 or more, corresponded to an average decrease in math score of 0.217575
- More books in the household corresponded to an increase in average math score, 0.212939 for each category of book numbers. We used this variable as a continuous variable, which is debatable. It is forcing a linear relationship to a nonlinear coding. We could have forced it to be fit as a categorical variable, and then interpretation may have been a little simpler.

Question 7

Using analysis of variance determine how much additional explanatory power including books in the model produces.

The gain is 8.552512 percent.

Question 8

Predict the average math score for these demographic groups

- Female student, started school at 4, mum and dad working full-time, has a desk, computer, own room, and internet, no TV at home and between 0-10 books at home.

```
0.168205
```

- Everything as before except for more than 3 TVs at home.

```
0.168205-3*0.217575=-0.48452
```

- Everything else as before except male student, and mum working part-time.

```
0.168205+0.219131+0.089772=0.477108
```

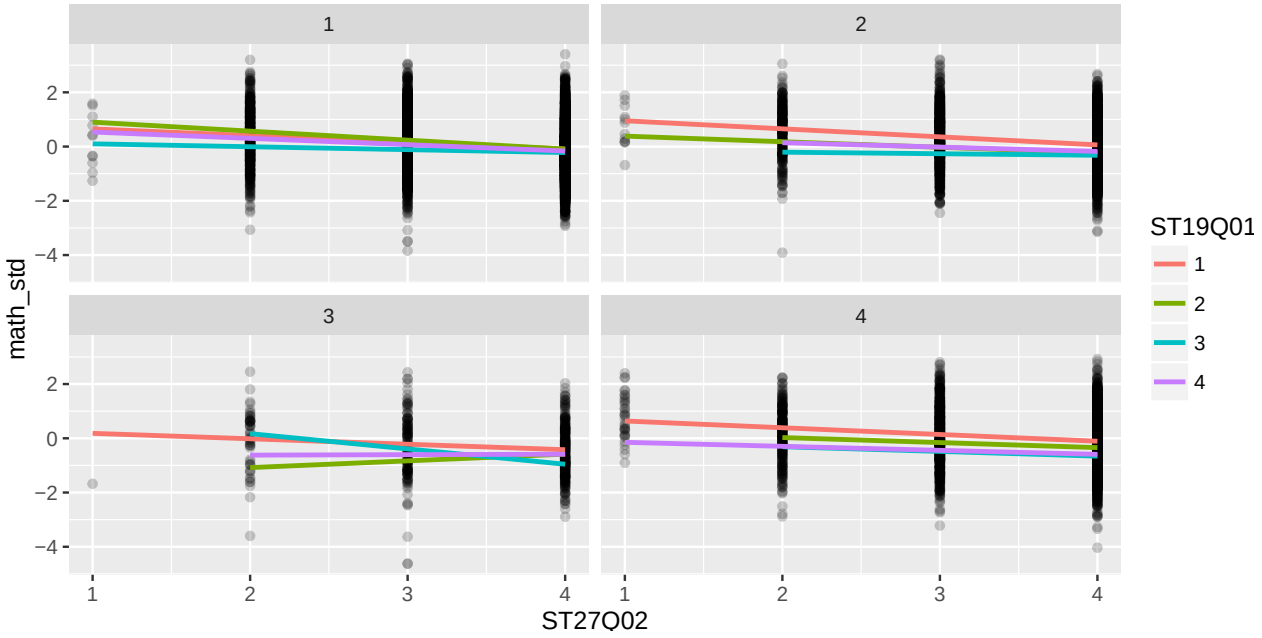
```
Compute the average and standard deviation of the math scores, to convert the standardised
scores to raw numbers.
```

- `0.168205*92.19892+505.6558=521.16`
- `-0.48452*92.19892+505.6558=460.98`
- `0.477108*92.19892+505.6558=549.64`

```
#          m        s
# 1 505.6558 92.19892
```

Question 9

This plot shows just a few of the variables with linear models fit separately in each level of the categorical variables: math score is plotted against number of TVs in the household (ST27Q02), separately for fathers job status (ST19Q01), and coloured by mothers job status (ST15Q01). Is there evidence that an interaction term should be fitted to the model? Explain.




```
There is a small amount of evidence to suggest a interaction terms are needed. The lines
are close to parallel, and mostly in the same order from top to bottom. Some slight
crossing of the lines for the mother and father looking for work can be seen. It pretty
weak, so the model will likely not be hugely improved by adding the complexity of interaction
terms.
```