# ETC 2420/5242 Lab 6 2017

*SOLUTION*

*Week 6*

## Purpose

This lab is to practice fitting and diagnosing multiple linear regression models.

## Reading

- The web site OECD PISA has a lot of information about the data. Click on the `TTest your skills online by answering some PISA 2016 science questions` and do some of the questions that students had to answer. How many did you get right out of how many attempted? _________
- Read the material on fitting multiple regression models in Statistics online textbook, Diez, Barr, Cetinkaya-Rundel.
- Read the code in the lecture notes on diagnostics for linear models from Week 5.

## Warmup exercises

- We are going to take a look at the OECD PISA 2015 data focusing on Australia.

```r
library("tidyverse")
library("forcats")
load("pisa_au.rda")
```

```r
pisa_au <- pisa_au %>%
  mutate(science = (PV1SCIE+PV2SCIE+PV3SCIE+PV4SCIE+
                    PV5SCIE+PV6SCIE+PV7SCIE+PV8SCIE+
                    PV9SCIE+PV10SCIE)/10)
```

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

| Variable name | Description |
|---|---|
| ST004D01T | Gender |
| OUTHOURS | Out-of-School Study Time |
| ANXTEST | Personality: Test Anxiety (WLE) |
| EMOSUPP | Parental emotional support (WLE) |
| PARED | Index highest parental education in years of schooling |
| JOYSCIE | Enjoyment of science (WLE) |
| WEALTH | Family wealth (WLE) |
| ST013Q01TA | How many books are there in your home? 1 0-10 books, 2 11-25 books, 3 26-100 books, 4 101-200 books, |
| ST012Q01TA | How many in your home: Televisions; 1 None, 2 One, 3 Two, 4 Three or more |
| SENWT | Weight |

Subset the data to contain just these variables.

```r
pisa_au <- pisa_au %>%
  select(science, ST004D01T, OUTHOURS, ANXTEST, EMOSUPP, PARED, JOYSCIE, WEALTH, ST013Q01TA, ST012Q01TA
pisa_au <- pisa_au %>% select(-EMOSUPP)
aus_nomiss <- pisa_au %>% filter(!is.na(OUTHOURS)) %>%
  filter(!is.na(ANXTEST)) %>% filter(!is.na(PARED)) %>%
  filter(!is.na(JOYSCIE)) %>% filter(!is.na(WEALTH)) %>%
  filter(!is.na(ST013Q01TA)) %>% filter(!is.na(ST012Q01TA))
```

Model building will be done using:

- Response: `science` (standardised)
- Explanatory variables: all the remaining variables

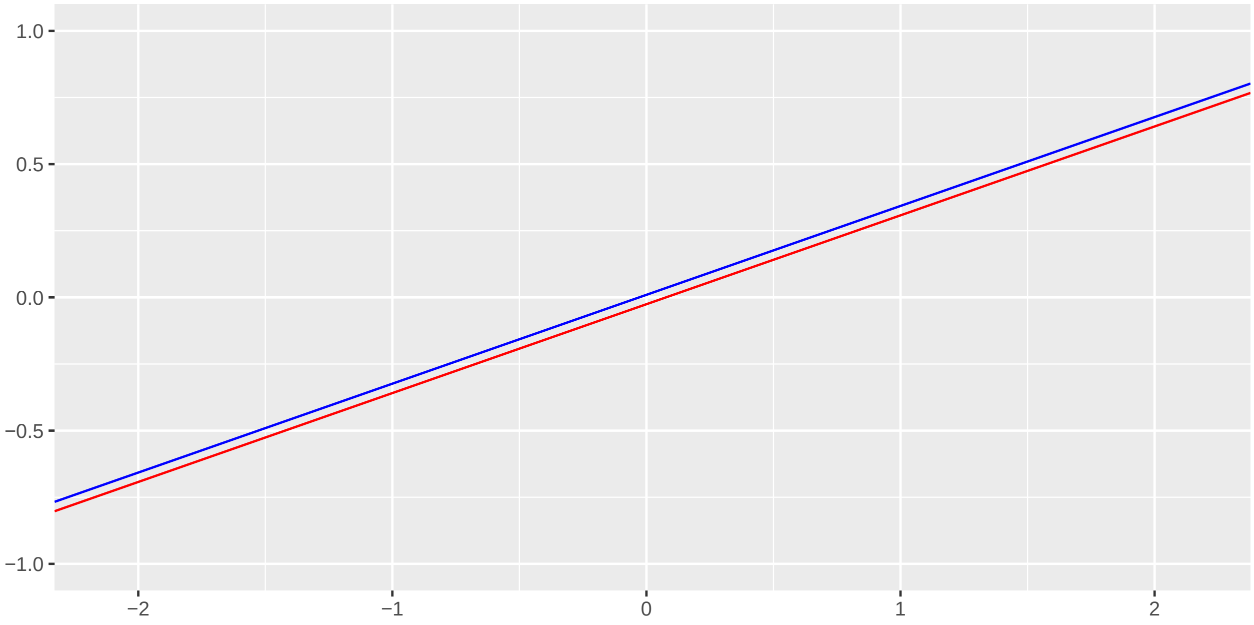
Some variables need to be treated as categorical variables, so it is best if they are forced to be factors before modeling:

```
aus_nomiss <- aus_nomiss %>%
  mutate(ST004D01T=factor(ST004D01T, levels=c(1,2), labels=c("f", "m")))
aus_nomiss <- aus_nomiss %>%
  mutate(science_std = (science-mean(science))/sd(science))
```

Test the model fitting, by fitting a model for science against gender, and joy of science.

```
aus_glm_test <- glm(science_std~ST004D01T+JOYSCIE,
                    data=aus_nomiss, weights=SENWT)
#summary(aus_glm_test)
```

Sketch what this model looks like.

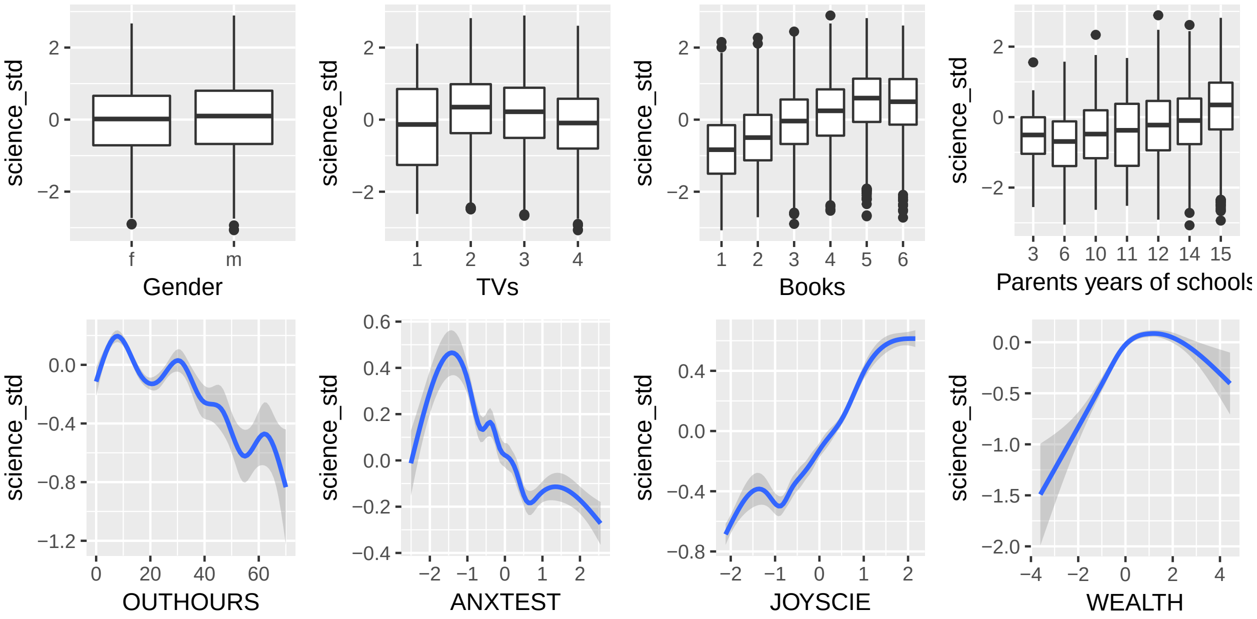


## Question 1 (5pts)

- Make plots of the response variable `science_std` against each of the possible explanatory variables.

```
A good selection of plots are side-by-side boxplots for the categorical (or discrete) predictors, and a
```

- Which variables look like they should be most important for predicting the response?



Most of the variables have a small relationship with science scores, so will likely be important for the model. Some have a nonlinear association, e.g. books, TVs, PARED, ANXTEST, WEALTH, which means that the linear model will be inadequate. It might be best to exclude OUTHOURS from the model, because there is insufficient data at the large values, as indicated in these plots by the wide standard error bands. It doesn't make a lot of sense for scores to go down with more study, and I suspect this is due to the low support at higher values.

## Question 2 (4pts)

- Fit the weighted multiple regression model to all the explanatory variables.
- (2pts) Summarise the coefficients for the model fit.

All the variables make a significant positive contribution to science score, except for ANXTEST and ST012Q01TA (TVs).

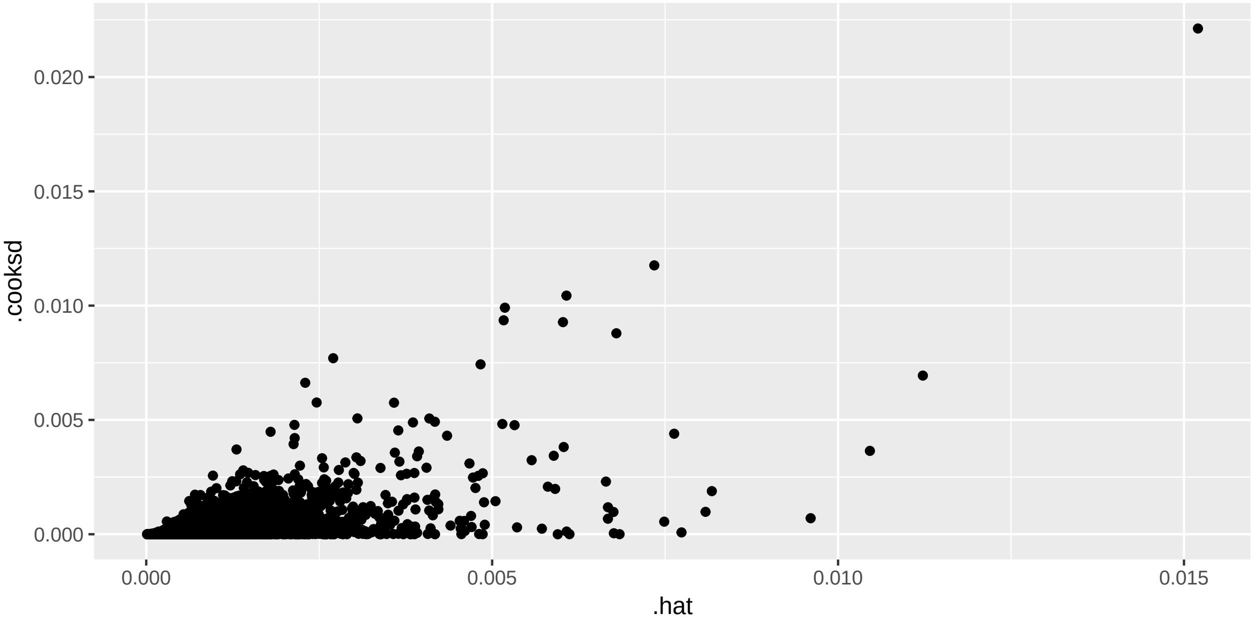
- (2pts) Not all variables are significant in the model. What variables can be dropped? Re-fit the model with this subset.

ST004D01T (Gender) is not significant in this model.

```
#
# Call:
# glm(formula = science_std ~ ANXTEST + PARED + JOYSCIE + WEALTH +
#     ST013Q01TA + ST012Q01TA, data = aus_nomiss, weights = SENWT)
#
# Deviance Residuals:
#      Min        1Q    Median        3Q       Max
# -2.25312  -0.29317   0.00247   0.29050   2.08263
#
# Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
# (Intercept) -1.122324   0.084554 -13.273  < 2e-16 ***
# ANXTEST     -0.092009   0.009092 -10.120  < 2e-16 ***
# PARED        0.068310   0.004894  13.958  < 2e-16 ***
# JOYSCIE      0.269074   0.007842  34.314  < 2e-16 ***
# WEALTH       0.043069   0.012028   3.581 0.000345 ***
# ST013Q01TA   0.193621   0.006611  29.288  < 2e-16 ***
# ST012Q01TA  -0.146459   0.013758 -10.646  < 2e-16 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# (Dispersion parameter for gaussian family taken to be 0.2527558)
#
#     Null deviance: 3264.3  on 9285  degrees of freedom
# Residual deviance: 2345.3  on 9279  degrees of freedom
# AIC: 25747
#
# Number of Fisher Scoring iterations: 2
```

## Question 3 (3pts)

- Compute the leverage and influence statistics.



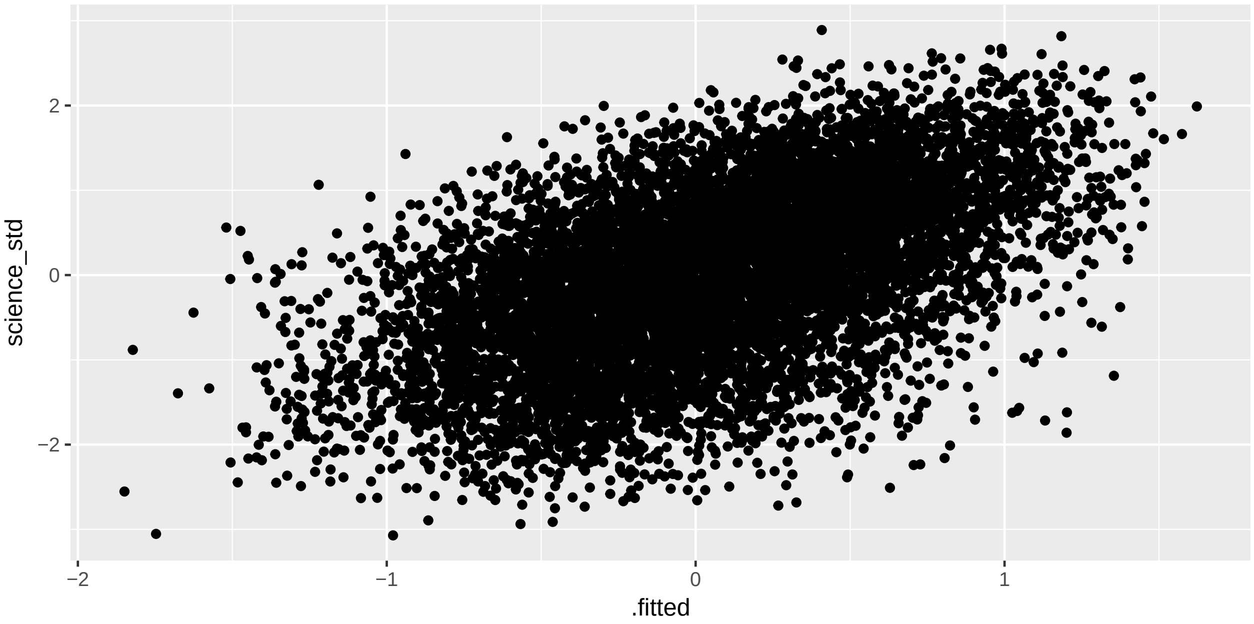
- (1pt) What value would be considered to be the cutoff for considering a case to have high leverage? `2*p/n=2*6/12118=0.00099`
- (2pts) How many cases have high influence?

2388

```
A lot of cases would be considered to have high leverage. No cases would be considered to
have high influence, the CooksD values are all very small.
```

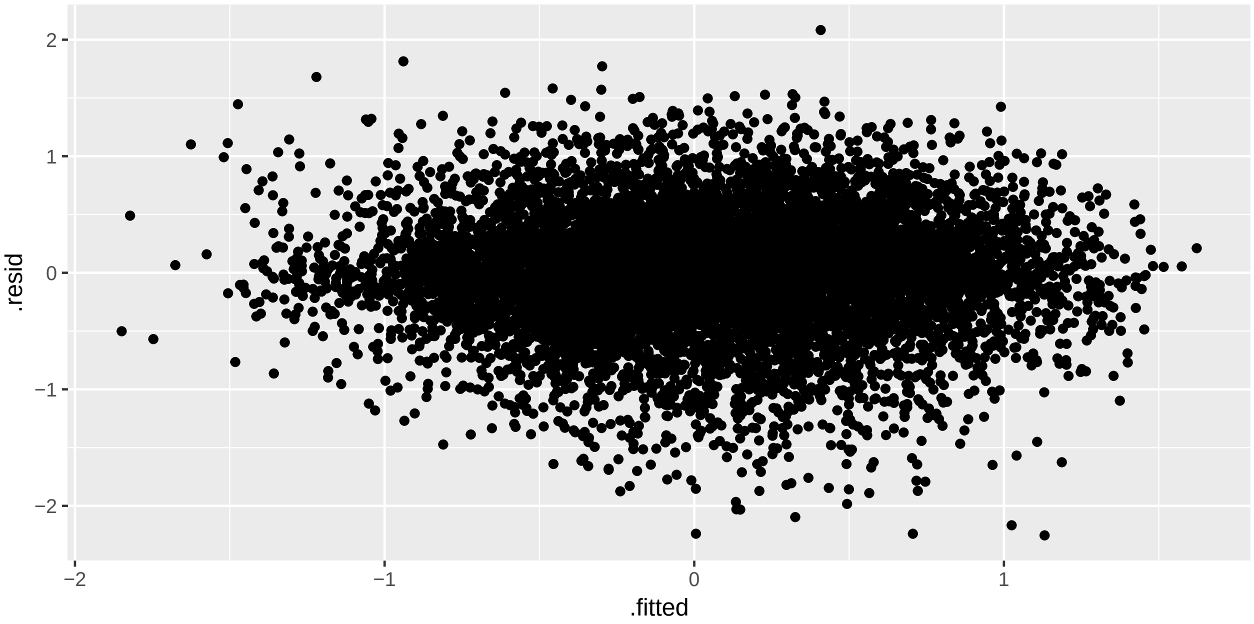
## Question 4 (6 pts)

- (2pts) Plot the observed vs fitted values. How good is the model for predicting science score? (Is it weak, moderate or strong?)



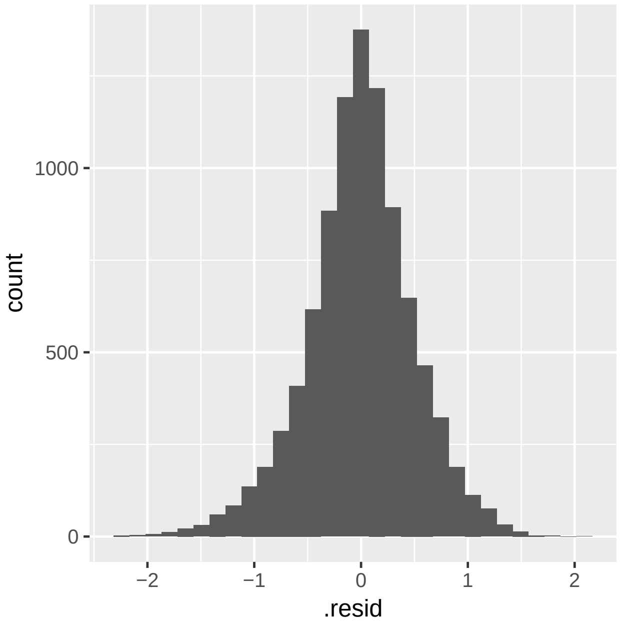
```
The model is a moderate to weak explanation of science scores. There is a lot of variation
in the scores that is clearly not explained by the predictors.
```

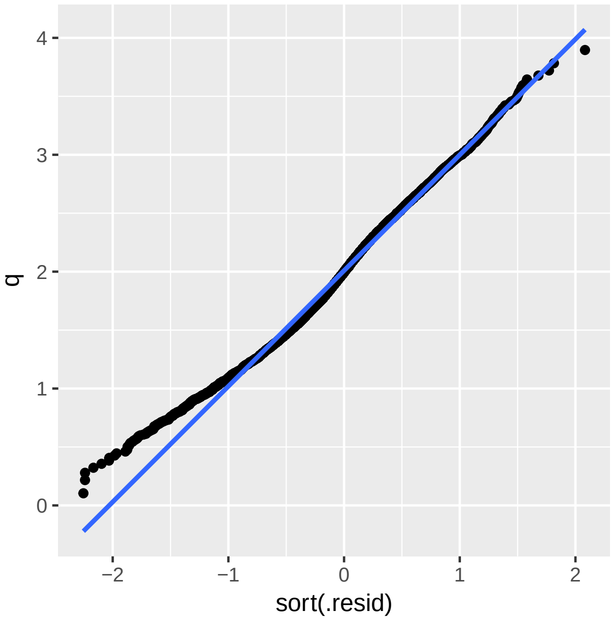
- (2pts) Plot residuals vs fitted. What do you learn about the model fit by looking at this plot?



```
There is very small amount of heteroskedasticity. The variation in residuals at smaller
fitted values is slightly smaller than at larger values, and the residuals tend to be
positive for low fitted values.
```

- (2pts) 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 tail is too long at the low values, as indicted in the histogram and the normal probability plot. Small residuals are higher than expected.

## Question 5 (2pts)

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

```
#    ANXTEST      PARED    JOYSCIE     WEALTH ST013Q01TA ST012Q01TA 
#   1.010217   1.110384   1.045175   1.383982   1.127823   1.321527 
```

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

## Question 6 (6pts)

Interpret the model, by answering these questions:

- (2pts) For boys how much do science scores increase or decrease on average, in relative to girls, keeping everything else fixed?

Nothing! Because gender was dropped from the model.

- For each additional hour spent studying how much does science score increase on average, keeping everything else fixed?

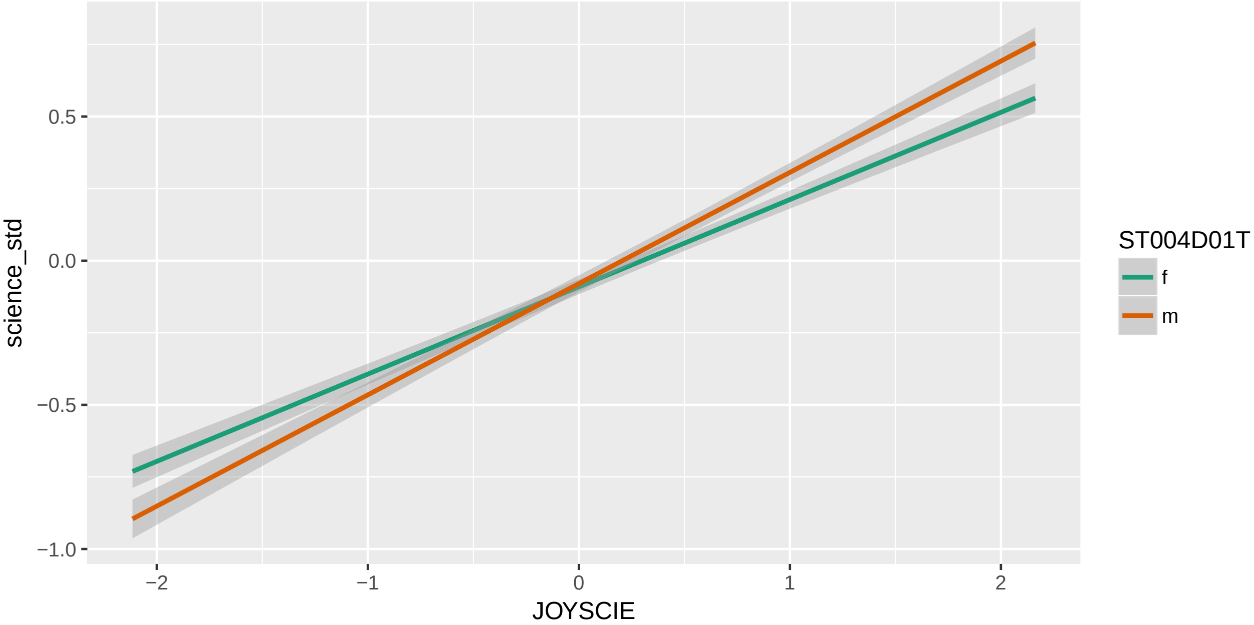
Nothing! Because OUTHOURS was dropped from the model.

- (2pts) For a household with more than 500 books does the average science score change, over a household with 0-10 books, keeping all else fixed?

This is 5 steps up in the X direction, 0.193621*5=0.968105. So an increase in about 1 standard unit of science scores is achieved.

## Question 7 (3pts)

This plot shows `science_std` plotted against JOYSCIE separately by ST004D01T (gender). Is there evidence that an interaction term should be fitted to the model? Explain.



```
Yes, this would suggest that for boys, as enjoyment of science increase, the average
score increases also. The increase is not as steep for girls.
```

## Question 8 (6pts)

Find the best model for science scores, which could include interaction terms, and justify your choice.

```
I expect various answers here. The end result should have model deviance lower than the
one for the main effects model already fitted, 2345.3. I am curious to learn how low this
can get, and what groups arrive at. Regardless of what is fitted, it will remain a fairly
weak model, probably at best 30% of variation in scores explained. The model diagnostics
should be reported for the final model, CooksD, VIF, residual plots. (I am calculating
this % by differencing null and residual deviation, and dividing by null deviance.)
```

## TURN IN

- Your `.Rmd` file
- Your `.html` file that results from knitting the Rmd.
- Make sure your group members are listed as authors, one person per group will turn in the report

## Resources

- Statistics online textbook, Diez, Barr, Cetinkaya-Rundel.