ETC 2420/5242 Lab 6 2017

Di Cook

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.

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

There are ten values for each student for the science score. The explanation for why this is, is long, but long story short, the raw scores that a student earns in the test are not distributed, but rather a large linear model is constructed, and ten predictions are randomly generated for each student from the model. Below is a scatterplot matrix plot of the plausible scores for each student of Australia. You can see that the scores are pretty similar across the variables, because the correlation is high and the scatter is strongly linear.

```
library("GGally")
sci_scores <- pisa_au %>% select(PV1SCIE, PV2SCIE, PV3SCIE, PV4SCIE,
                                 PV5SCIE, PV6SCIE, PV7SCIE, PV8SCIE,
                                 PV9SCIE, PV10SCIE) %>% as.data.frame()
ggscatmat(sci_scores, alpha=0.1)
```

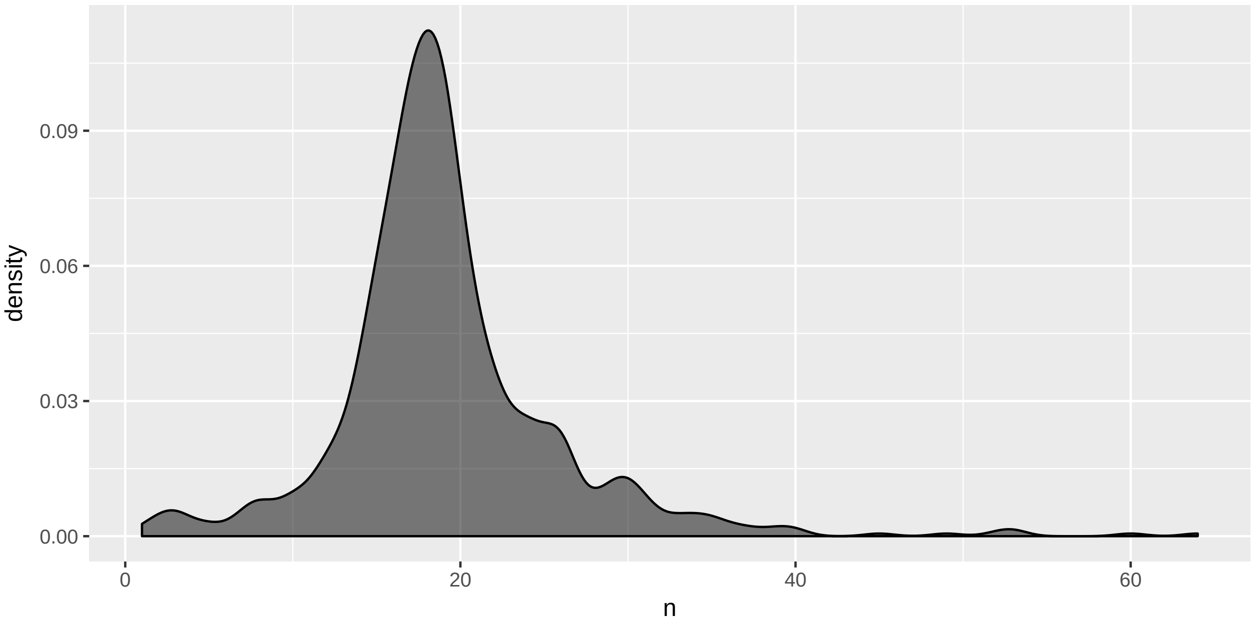
We will create a new variable which is the average of the ten scores for each student.

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

Students are tested at many different schools. How many schools? And what is the distribution of number of students tested at each school?

```
pisa_au %>% group_by(CNTSCHID) %>%
  tally() %>%
  arrange(desc(n)) -> aus_schools
dim(aus_schools)
```

```
# [1] 758   2
ggplot(aus_schools, aes(x=n)) +
  geom_density(fill="black", alpha=0.5)
```



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.

```
pisa_au <- pisa_au %>%
  select(science, ST004D01T, OUTHOURS, ANXTEST, EMOSUPP, PARED, JOYSCIE, WEALTH, ST013Q01TA, ST012Q01TA
```

Make summaries each of the variables, to examine their suitability for modeling.

```
summary(pisa_au)
#     science        ST004D01T        OUTHOURS        ANXTEST
#  Min.   :215.1   Min.   :1.000   Min.   : 0.00   Min.   :-2.5050
#  1st Qu.:423.7   1st Qu.:1.000   1st Qu.: 8.00   1st Qu.:-0.4305
#  Median :502.5   Median :2.000   Median :15.00   Median : 0.1962
#  Mean   :498.7   Mean   :1.507   Mean   :16.84   Mean   : 0.2053
#  3rd Qu.:572.8   3rd Qu.:2.000   3rd Qu.:22.00   3rd Qu.: 0.7186
#  Max.   :802.0   Max.   :2.000   Max.   :70.00   Max.   : 2.5493
#                                  NA's   :4186    NA's   :632
#     EMOSUPP          PARED           JOYSCIE            WEALTH
#  Min.   : NA      Min.   : 3.00   Min.   :-2.1154   Min.   :-6.9778
#  1st Qu.: NA      1st Qu.:12.00   1st Qu.:-0.7825   1st Qu.: 0.0533
#  Median : NA      Median :14.00   Median : 0.0992   Median : 0.6003
#  Mean   :NaN      Mean   :13.35   Mean   : 0.0729   Mean   : 0.5827
#  3rd Qu.: NA      3rd Qu.:15.00   3rd Qu.: 0.5094   3rd Qu.: 1.1207
#  Max.   : NA      Max.   :15.00   Max.   : 2.1635   Max.   : 4.4269
#  NA's   :14530    NA's   :698     NA's   :1952      NA's   :433
#    ST013Q01TA       ST012Q01TA        SENWT
#  Min.   :1.000   Min.   :1.000   Min.   :0.01951
#  1st Qu.:2.000   1st Qu.:3.000   1st Qu.:0.12202
#  Median :3.000   Median :4.000   Median :0.33399
#  Mean   :3.387   Mean   :3.478   Mean   :0.34412
#  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:0.53531
#  Max.   :6.000   Max.   :4.000   Max.   :1.12726
#  NA's   :549     NA's   :562
```

- EMOSUPP is all missing, cannot use this variable.
- OUTHOURS has about one third missing, might be unreliable
- JOYSCIE has 10% missing, might be unreliable

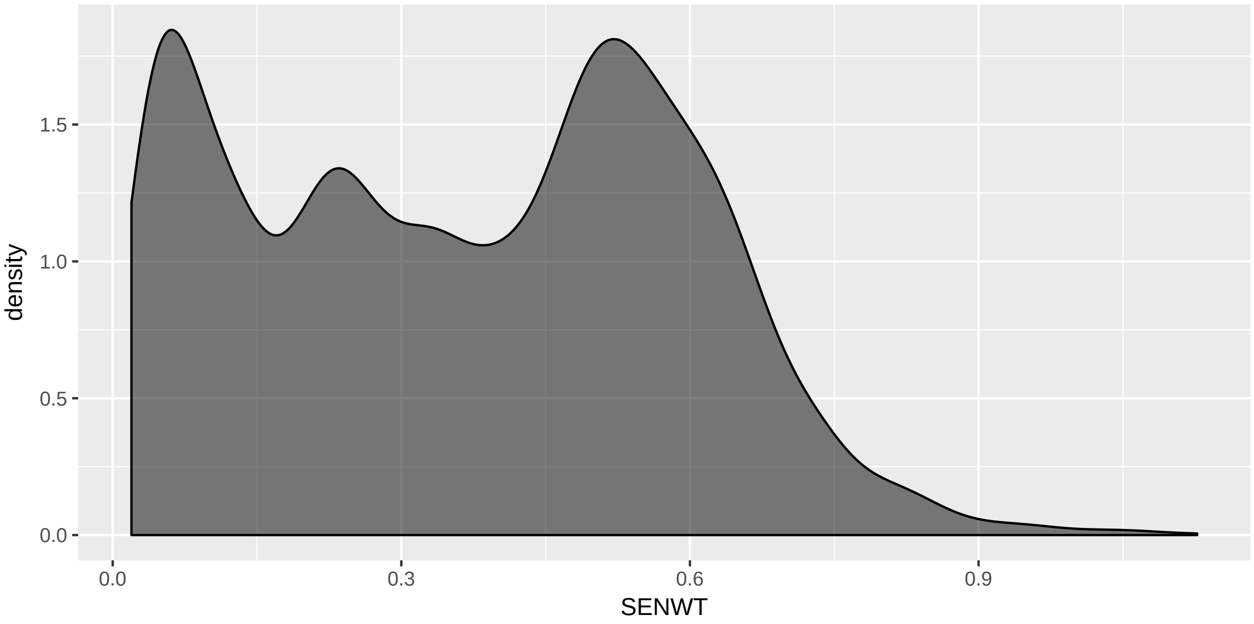
Actions to take:

- Drop EMOSUPP
- Remove any case with missings - and check numbers remaining

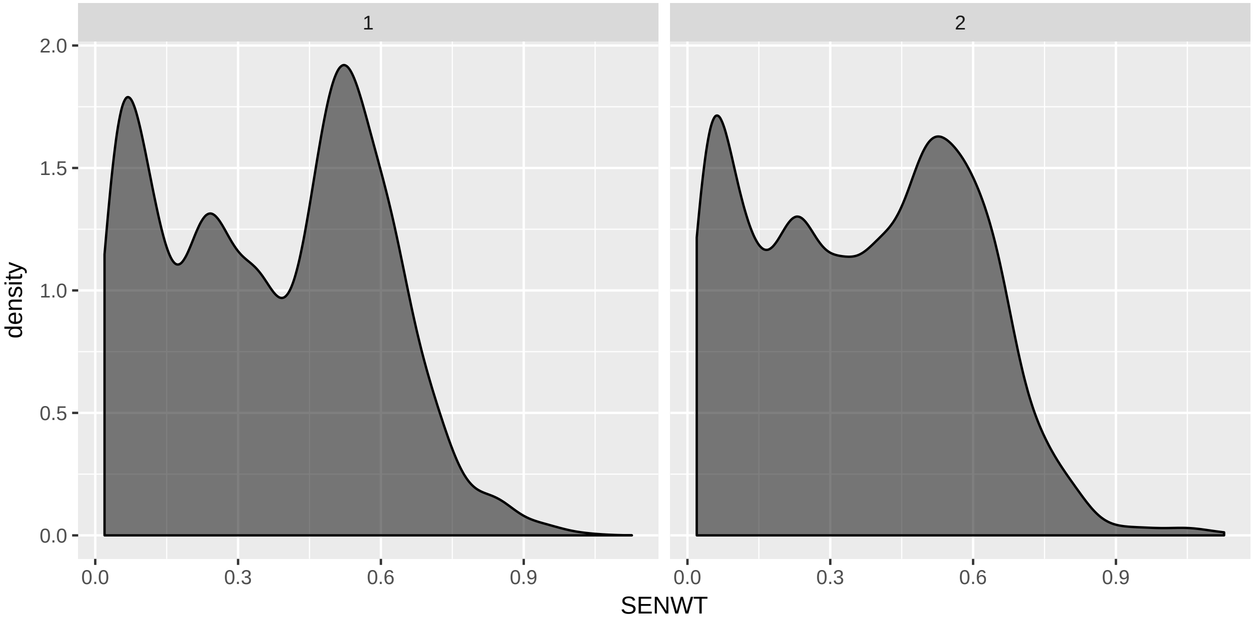
```
pisa_au <- pisa_au %>% select(-EMOSUPP)
aus_nomiss <- pisa_au %>% filter(!is.na(OUTHOURS)) %>%
  filter(!is.na(ANXTEST)) %>% filter(!is.na(PARED)) %>%
  filter(!is.na(JOYSCIE)) %>% filter(!is.na(WEALTH)) %>%
  filter(!is.na(ST013Q01TA)) %>% filter(!is.na(ST012Q01TA))
```

The number of students (observations) drops from 14530 to 9286 about 5000. That's a lot. If we ignored OUTHOURS there would be a loss of much less data, about 2000 records.

The students all have weights associated with them. This is an indication of how many other students they represent in Australia, relative to their socioeconomic and demographic characteristics. Let's look at the distribution of weights



There is a lot of variation in the weights. The weights are bimodal (is the bimodality due to one of the variables in the study that we are using for the model? Its not due to gender!) with a few very large ones. It looks like we will need to take weight into account in the model.



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)
#
# Call:
# glm(formula = science_std ~ ST004D01T + JOYSCIE, data = aus_nomiss,
#     weights = SENWT)
#
# Deviance Residuals:
#      Min        1Q    Median        3Q       Max
# -2.44167  -0.32606  -0.00879   0.32087   2.10647
#
# Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
# (Intercept) -0.025605   0.013102  -1.954   0.0507 .
# ST004D01Tm   0.035240   0.018813   1.873   0.0611 .
# JOYSCIE      0.333510   0.008377  39.815   <2e-16 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# (Dispersion parameter for gaussian family taken to be 0.2996778)
#
#     Null deviance: 3264.3  on 9285  degrees of freedom
# Residual deviance: 2781.9  on 9283  degrees of freedom
# AIC: 27324
#
# Number of Fisher Scoring iterations: 2
```

Sketch what this model looks like.

Question 1

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

Question 2

- Fit the weighted multiple regression model to all the explanatory variables.
- Summarise the coefficients for the model fit.
- Not all variables are significant in the model. What variables can be dropped? Re-fit the model with this subset.

Question 3

- Compute the leverage and influence statistics.
- What value would be considered to be the cutoff for considering a case to have high leverage?

- How many cases have high influence?

Question 4

- Plot the observed vs fitted values. How good is the model for predicting science score? (Is it weak, moderate or strong?)
- Plot residuals vs fitted. What do you learn about the model fit by looking at this plot?
- Make a histogram of residuals, and a qqplot (normal probability plot). Do these look like a sample from a normal model?

Question 5

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

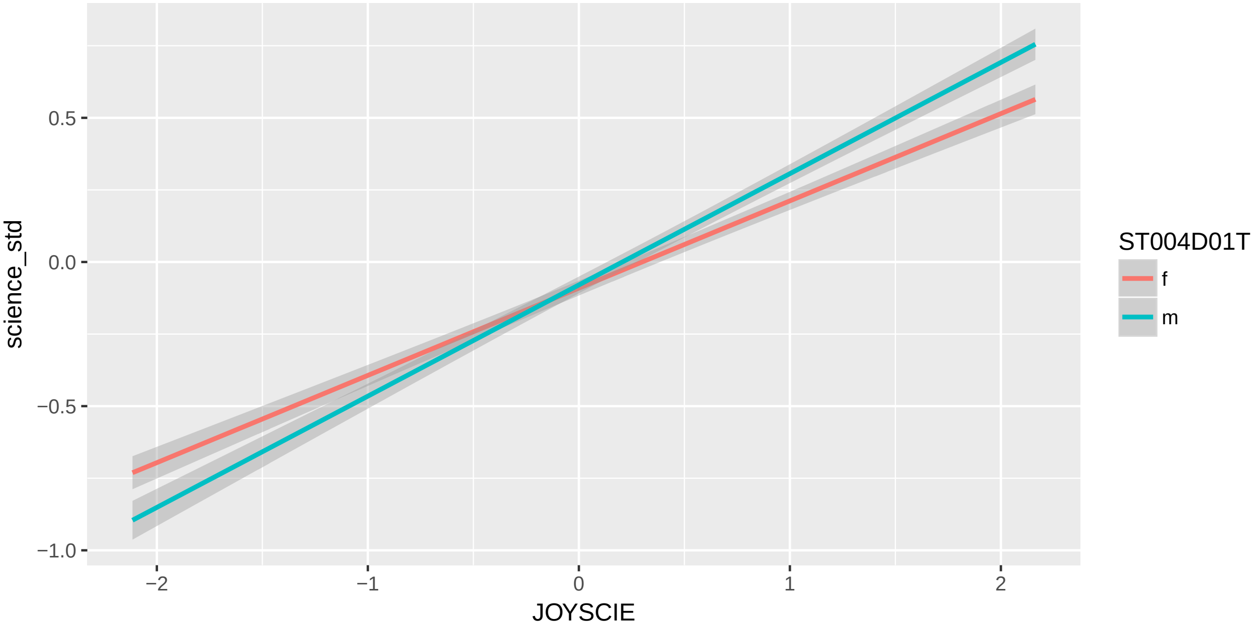
Question 6

Interpret the model, by answering these questions:

- For boys how much do science scores increase or decrease on average, in relative to girls, keeping everything else fixed?
- For each additional hour spent studying how much does science score increase on average, keeping everything else fixed?
- For a household with more than 100 books does the average science score change, keeping all else fixed?

Question 7

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.



Question 8

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

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.