ETC 2420/5242 Lab 6 2016

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 `Try PISA 2012 Test 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.

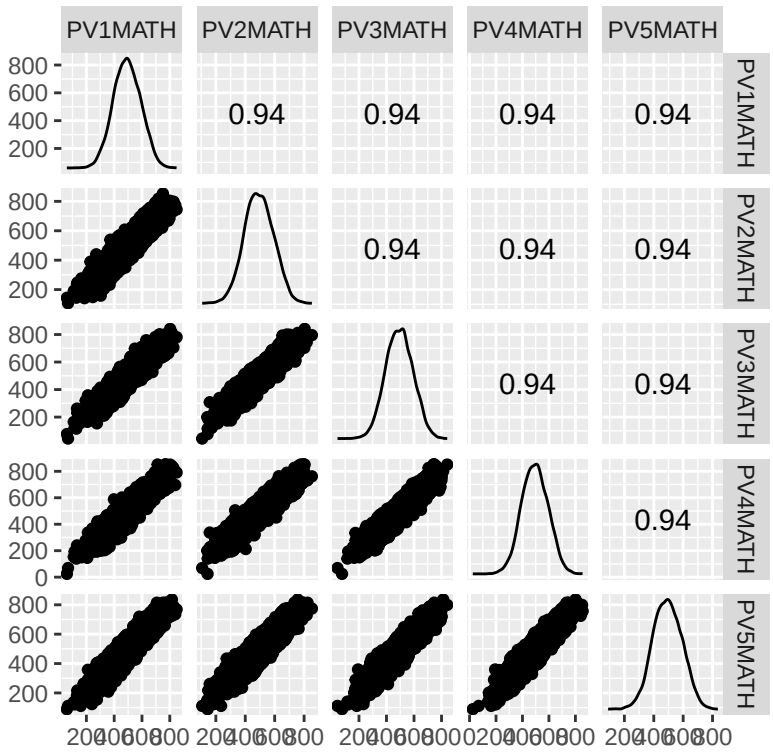
Warmup exercises

- We are going to take a look at the OECD PISA 2012 data used in the first lab.
- The data has standardised test scores for 15 year old students from around the globe.

```
library("tidyr")
library("dplyr")
library("ggplot2")
student2012.sub <- readRDS("../data/student_sub.rds")
australia <- student2012.sub[student2012.sub$CNT=="AUS",]
```

There are five values for each student for the math 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 five predictions are randomly generated for each student from the model. Below is a scatterplot matrix plot of the five plausible scores for each student of Australia. You can see that the scores are pretty similar across the five variables, because the correlation is high and the scatter is strongly linear.

```
library("GGally")
ggscatmat(australia, columns=35:39)
```

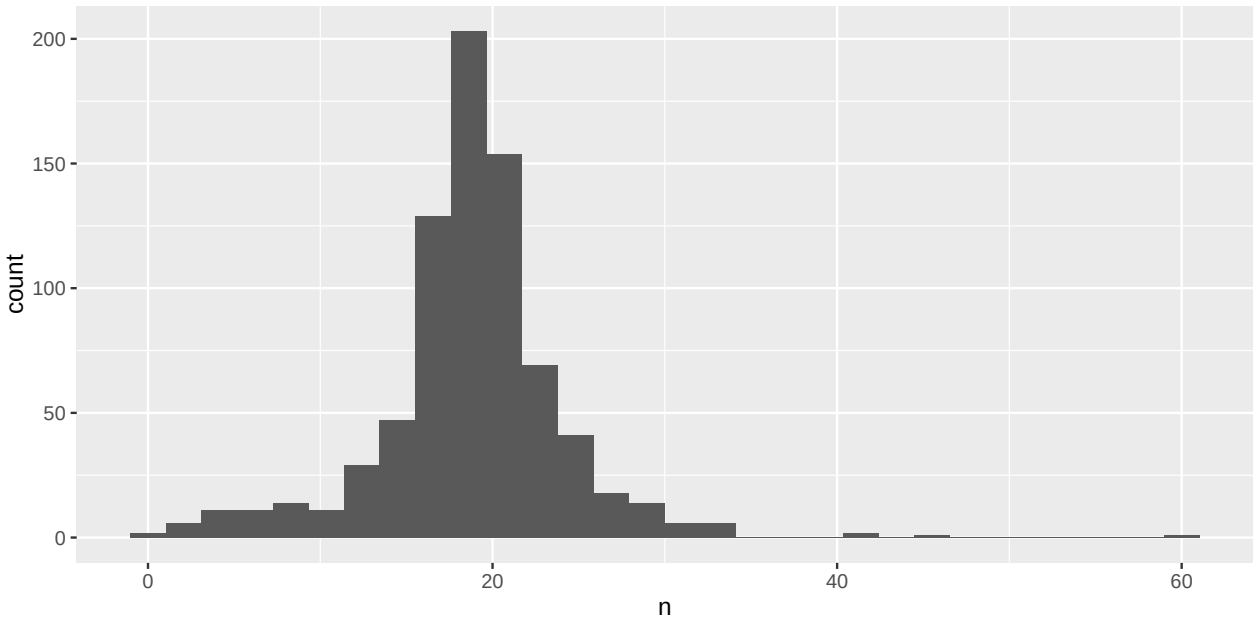


We will create a new variable which is the average of the five scores for each student.

```
australia <- australia %>% mutate(math = (PV1MATH+PV2MATH+PV3MATH+PV4MATH+PV5MATH)/5)
```

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

```
australia %>% group_by(SCHOOLID) %>%
  tally() %>%
  arrange(desc(n)) -> aus_schools
dim(aus_schools)
# [1] 775   2
ggplot(aus_schools, aes(x=n)) + geom_histogram()
```



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

We need to replace the codes for missing, with NAs, so that R will handle these as missing values.

```
australia <- australia %>%
  select(math, ST04Q01, ST06Q01, ST57Q01, ST15Q01, ST19Q01, ST26Q01, ST26Q02, ST26Q04, ST26Q06, ST27Q02
australia$ST06Q01[australia$ST06Q01 > 9990] <- NA
australia$ST57Q01[australia$ST57Q02 > 9990] <- NA
australia$ST15Q01[australia$ST15Q01 > 6] <- NA
australia$ST19Q01[australia$ST19Q01 > 6] <- NA
australia$ST26Q01[australia$ST26Q01 > 6] <- NA
australia$ST26Q02[australia$ST26Q02 > 6] <- NA
```

```
australia$ST26Q04[australia$ST26Q04 > 6] <- NA
australia$ST26Q06[australia$ST26Q06 > 6] <- NA
australia$ST27Q02[australia$ST27Q02 > 6] <- NA
australia$ST28Q01[australia$ST28Q01 > 6] <- NA
```

Tabulate each of the variables, to examine the support that we have on each, and whether it is enough to be included in the model.

```
australia %>% group_by(ST04Q01) %>% tally()
# # A tibble: 2 x 2
#   ST04Q01     n
#     <dbl> <int>
# 1       1  7075
# 2       2  7406
australia %>% group_by(ST06Q01) %>% tally()
# # A tibble: 5 x 2
#   ST06Q01     n
#     <dbl> <int>
# 1       4  1634
# 2       5  7760
# 3       6  3517
# 4       7   427
# 5      NA  1143
australia %>% group_by(ST57Q01) %>% tally()
# # A tibble: 34 x 2
#    ST57Q01     n
#      <dbl> <int>
# 1        0   835
# 2        1  1185
# 3        2  1279
# 4        3   835
# 5        4   663
# 6        5   801
# 7        6   524
# 8        7   334
# 9        8   352
# 10       9   145
# # ... with 24 more rows
australia %>% group_by(ST15Q01) %>% tally()
# # A tibble: 5 x 2
#   ST15Q01     n
#     <dbl> <int>
# 1       1  7069
# 2       2  3249
# 3       3   611
# 4       4  2907
# 5      NA   645
australia %>% group_by(ST19Q01) %>% tally()
# # A tibble: 5 x 2
#   ST19Q01     n
#     <dbl> <int>
# 1       1 11080
# 2       2   870
```

```
# 3       3   447
# 4       4   956
# 5      NA  1128
australia %>% group_by(ST26Q01) %>% tally()
# # A tibble: 3 x 2
#   ST26Q01     n
#     <dbl> <int>
# 1       1 12528
# 2       2  1451
# 3      NA   502
australia %>% group_by(ST26Q02) %>% tally()
# # A tibble: 3 x 2
#   ST26Q02     n
#     <dbl> <int>
# 1       1 13086
# 2       2  1050
# 3      NA   345
australia %>% group_by(ST26Q04) %>% tally()
# # A tibble: 3 x 2
#   ST26Q04     n
#     <dbl> <int>
# 1       1 13631
# 2       2   459
# 3      NA   391
australia %>% group_by(ST26Q06) %>% tally()
# # A tibble: 3 x 2
#   ST26Q06     n
#     <dbl> <int>
# 1       1 13477
# 2       2   589
# 3      NA   415
australia %>% group_by(ST27Q02) %>% tally()
# # A tibble: 5 x 2
#   ST27Q02     n
#     <dbl> <int>
# 1       1    81
# 2       2  1134
# 3       3  3693
# 4       4  9172
# 5      NA   401
australia %>% group_by(ST28Q01) %>% tally()
# # A tibble: 7 x 2
#   ST28Q01     n
#     <dbl> <int>
# 1       1  1587
# 2       2  1834
# 3       3  4234
# 4       4  2786
# 5       5  2394
# 6       6  1260
# 7      NA   386
```

- Gender is clearly ok
- Most children start at age 5 or 6 in Australia, there is less data for 4 and 7 year olds, but probably ok

- Students in Australia don't spend much time on homework. Two hours is the mode, and then numbers dwindle. It will be necessary to combine groups, maybe 9 or more hours into one group. Most data is missing on this variable, though - 42.5%. So it might be better to just not use it for modeling.
- Mothers full-time, part-time and at-home are large groups, number of mothers looking for work is low, but might be ok
- Fathers are mostly in full-time work, numbers in all other categories are relatively low.
- Most Australian 15 year olds have a desk, their own room, computer and internet! Numbers without these are low, but maybe still enough for modeling.
- Most households have three or more TVs! Households without a TV are very few, and may not be enough to be able to build a model.
- Most households have between 26-100 books. Most of the categories have large numbers so all should be ok to use.

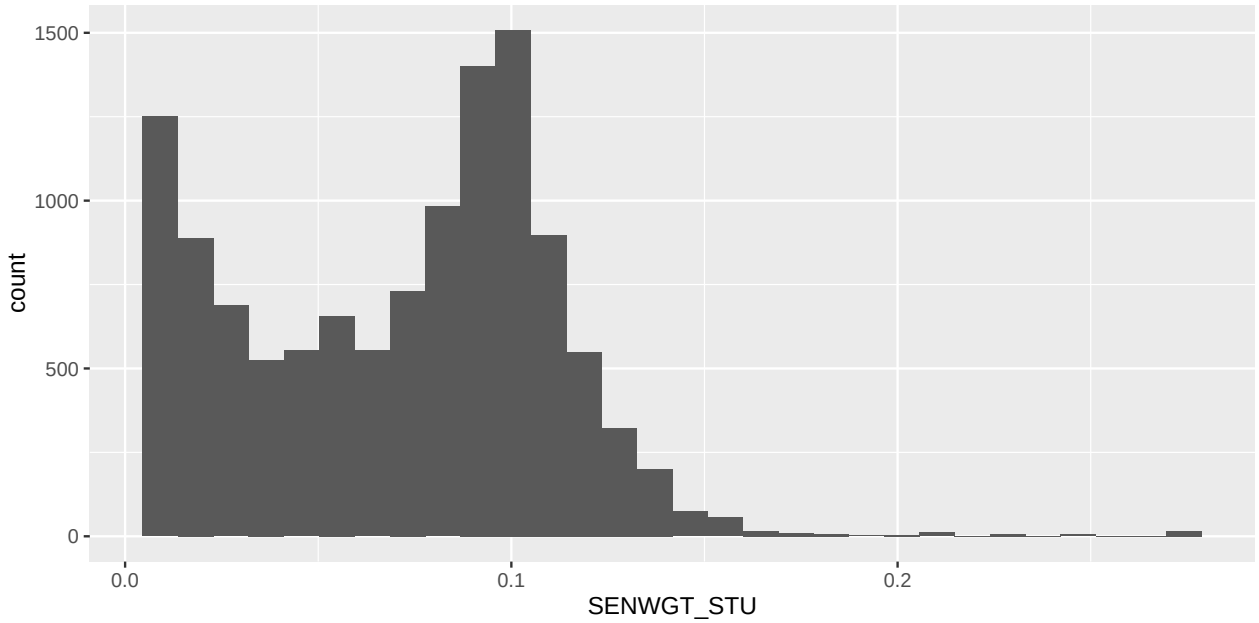
Actions to take:

- Drop ST57Q01
- Remove any case with missings - hopefully this is not very much

```
australia <- australia %>% select(-ST57Q01)
aus_nomiss <- australia %>% filter(!is.na(ST04Q01)) %>%
  filter(!is.na(ST06Q01)) %>% filter(!is.na(ST15Q01)) %>%
  filter(!is.na(ST19Q01)) %>% filter(!is.na(ST26Q01)) %>%
  filter(!is.na(ST26Q02)) %>% filter(!is.na(ST26Q04)) %>%
  filter(!is.na(ST26Q06)) %>% filter(!is.na(ST27Q02)) %>%
  filter(!is.na(ST28Q01))
```

The number of students (observations) drops from 14481 to 11918 about 2500. That's a lot, but not a huge amount, about 17%, so its ok to start building a model from.

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?) 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: `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:

```
aus_nomiss$ST04Q01 <- factor(aus_nomiss$ST04Q01)
aus_nomiss$ST15Q01 <- factor(aus_nomiss$ST15Q01)
aus_nomiss$ST15Q01 <- factor(aus_nomiss$ST15Q01)
aus_nomiss$ST19Q01 <- factor(aus_nomiss$ST19Q01)
aus_nomiss$ST26Q01 <- factor(aus_nomiss$ST26Q01)
aus_nomiss$ST26Q02 <- factor(aus_nomiss$ST26Q02)
aus_nomiss$ST26Q04 <- factor(aus_nomiss$ST26Q04)
aus_nomiss$ST26Q06 <- factor(aus_nomiss$ST26Q06)
aus_nomiss <- aus_nomiss %>% mutate(math_std = (math-mean(math))/sd(math))
aus_nomiss$ST06Q01 <- aus_nomiss$ST06Q01 - 4
```

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

```
aus_glm_test <- glm(math_std~ST04Q01+ST26Q04+ST28Q01, data=aus_nomiss, weights=SENWGT_STU)
summary(aus_glm_test)
#
# Call:
# glm(formula = math_std ~ ST04Q01 + ST26Q04 + ST28Q01, data = aus_nomiss,
#     weights = SENWGT_STU)
#
# Deviance Residuals:
#      Min        1Q    Median        3Q       Max
# -1.17644  -0.14944  -0.01805   0.13373   0.99936
#
# Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
# (Intercept) -0.855925   0.025023 -34.206   <2e-16 ***
# ST04Q012     0.193580   0.016809  11.516   <2e-16 ***
# ST26Q042    -0.607698   0.064350  -9.444   <2e-16 ***
# ST28Q01      0.241709   0.006062  39.871   <2e-16 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# (Dispersion parameter for gaussian family taken to be 0.05947803)
#
#     Null deviance: 819.22  on 11917  degrees of freedom
# Residual deviance: 708.62  on 11914  degrees of freedom
# AIC: 34896
#
# Number of Fisher Scoring iterations: 2
```

Sketch what this model looks like.

Question 1

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

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 math 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:

- 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?
- ....

Question 7

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

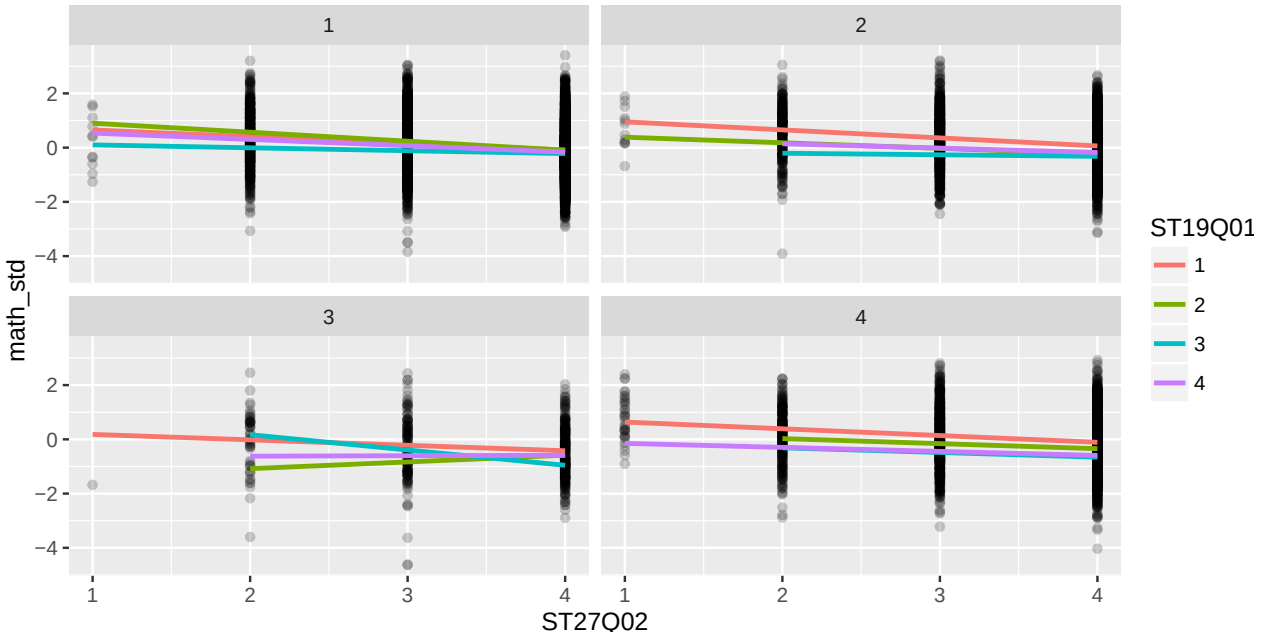
Question 8

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.
- Everything as before except for more than 3 TVs at home.
- Everything else as before except male student, and mum working part-time.

Question 9

This plot shows just a few of the variables with linear models fit separately in each level of the categorical variables: `math_std` 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.



TURN IN

- Your `.Rmd` file
- Your Word (or pdf) 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
- DUE: Wednesday after the lab, by 7am, loaded into moodle

Resources

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