STAC33

Assignment 7

Due Tuesday March 24 at 11:59pm

To begin:

```
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0    v stringr 1.4.0
## v readr   1.3.1    v forcats 0.4.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Optionally, install and load package `broom`. (You don't need this: you can do the assignment without it.)

1. Work through at least some of Chapter 14 of PASIAS. There are lots of problems there. Problems 14.4–14.9 are good practice for the problem you'll be handing in.

   Hand in the next one.

2. The SAT is a standardized test used in the US as part of the college admissions process. Two of the sections of the test are Math and Verbal. Students receive a score on each. Are the two scores related? The data in `http://ritsokiguess.site/STAC33/sat.csv` are Math and Verbal SAT scores for a number of students. The data file also contains the sex of each student, which we will ignore in this question. Our aim is to predict math SAT score from verbal SAT score.

   (a) (2 marks) Read in and display (some of) the data. How many students are there in the data set?

   Solution: It's a `.csv`, so nothing new here:

```
my_url <- "http://ritsokiguess.site/STAC33/sat.csv"
sat <- read_csv(my_url)

## Parsed with column specification:
## cols(
##   verbal_sat = col_double(),
##   math_sat = col_double(),
##   sex = col_character()
## )

sat

## # A tibble: 162 x 3
##    verbal_sat math_sat sex
##         <dbl>    <dbl> <chr>
##  1        450      450 F
##  2        640      540 F
##  3        590      570 M
##  4        400      400 M
##  5        600      590 M
##  6        610      610 M
##  7        630      610 F
##  8        660      570 M
##  9        660      720 F
## 10        590      640 F
## # ... with 152 more rows
```

Calling the data frame `sat` is safe, since the names of the columns are something different. One point for this much.

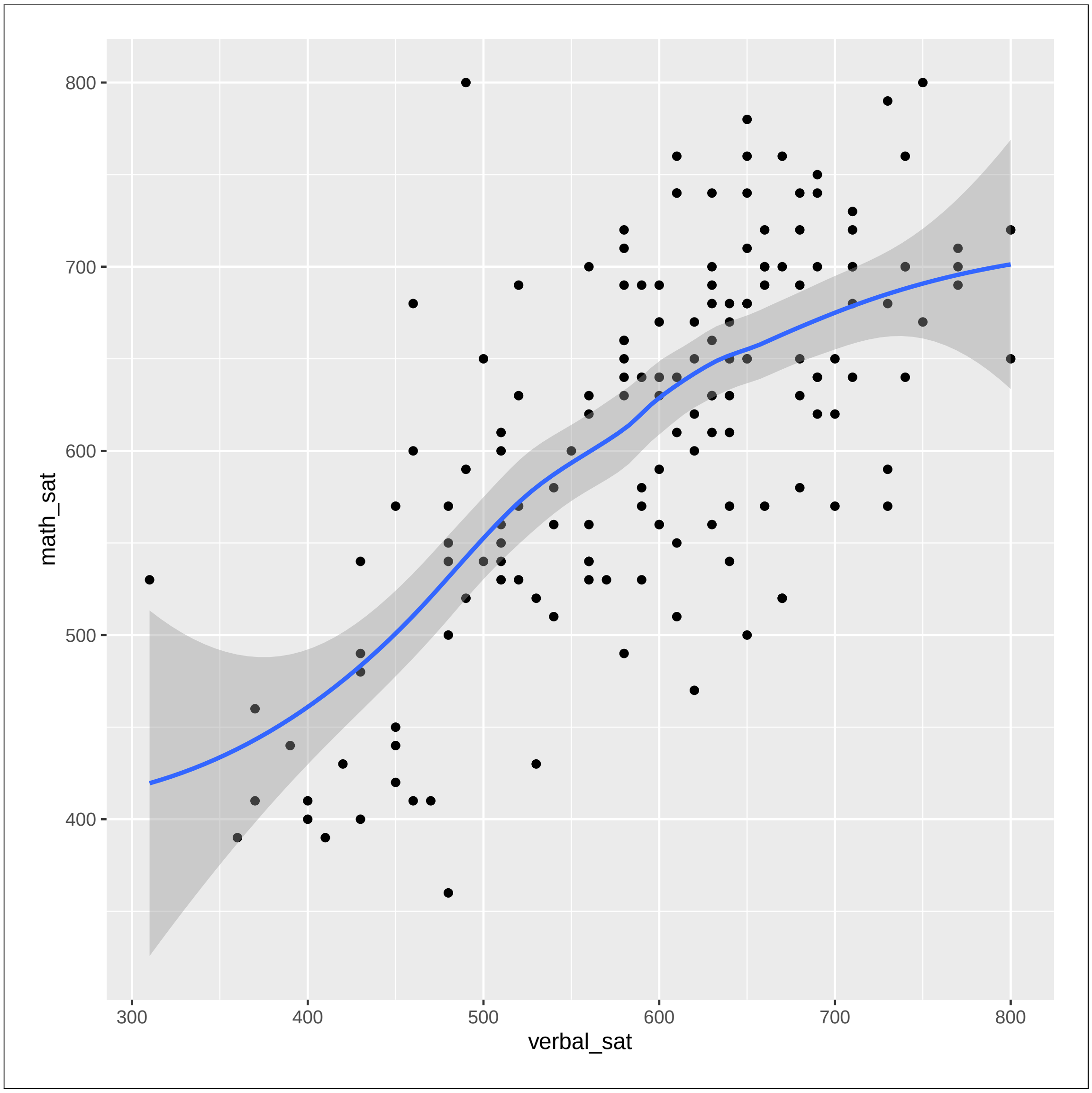
There are 162 students. One point for this. Don't get lazy and forget to say this.

(b) (2 marks) Make a suitable plot of the two SAT scores for each student. Add a smooth trend.

Solution: Both variables are quantitative, so a scatterplot. The math SAT score is what we are trying to predict, so that goes on the y-axis:

```
ggplot(sat, aes(x=verbal_sat, y=math_sat)) + geom_point() + geom_smooth()

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



(c) (3 marks) What do you see in your plot? Explain briefly. Hint: think about (i) form: linear or curved, (ii) direction: up or down or a mixture of both, (iii) strength: strong relationship, moderate, or weak.

Solution: Say something about each of those three things, adding a little explanation:

- form: make a call. I think you can support "approxinately linear", or "curved with some levelling off".
- direction: upward: as the verbal SAT score increases, the math SAT score also tends to increase.
- strength: pick an adjective. I'd call this "moderate", since there is a clear trend but some of the points are away from the line. Make a call.

Extra: SAT scores are always between 200 and 800. Some of any curvature you see might be for scores close to the maximum; if it were possible for scores to go above 800, the relationship might be straighter.

(d) (2 marks) Fit a linear regression predicting math SAT score from verbal SAT score. Display the output.

Solution: This:

```
sat.1 <- lm(math_sat ~ verbal_sat, data=sat)
summary(sat.1)
```

```
##
## Call:
## lm(formula = math_sat ~ verbal_sat, data = sat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -173.590  -47.596    1.158   45.086  259.659
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 209.55417   34.34935   6.101 7.66e-09 ***
## verbal_sat    0.67507    0.05682  11.880  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 71.75 on 160 degrees of freedom
## Multiple R-squared:  0.4687,Adjusted R-squared:  0.4654
## F-statistic: 141.1 on 1 and 160 DF,  p-value: < 2.2e-16
```

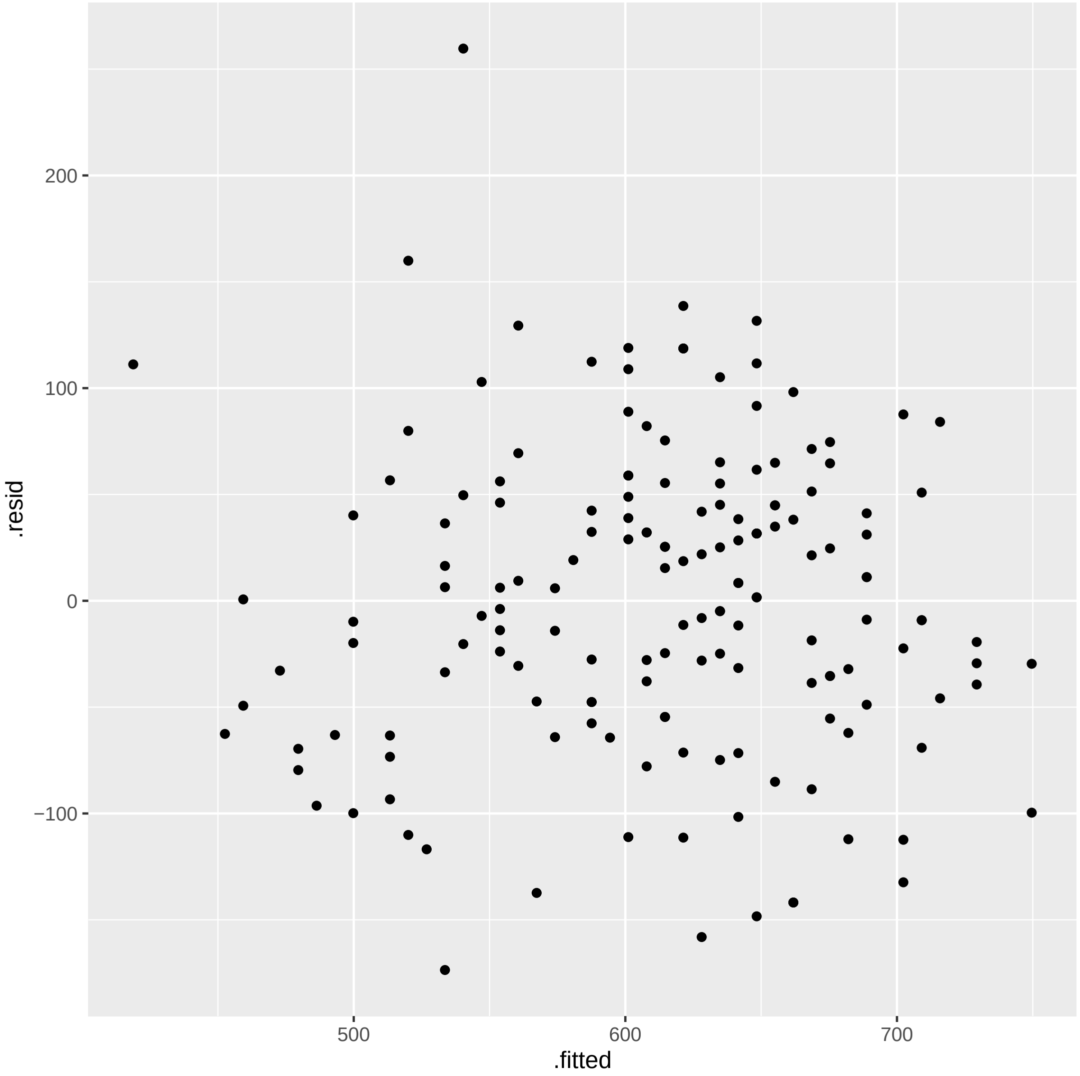
(e) (2 marks) Is there a relationship between SAT verbal and math scores for all students (of whom the students in this data set are a sample)? Explain briefly.

Solution: This is what the hypothesis test for the slope is assessing. The P-value is (very) small, so there definitely is a relationship. (The relationship observed here is definitely more than chance.)

(f) (3 marks) Obtain a plot of residuals against fitted values. What do you conclude from it? Explain briefly.

Solution: This:

```
ggplot(sat.1, aes(x=.fitted, y=.resid)) + geom_point()
```

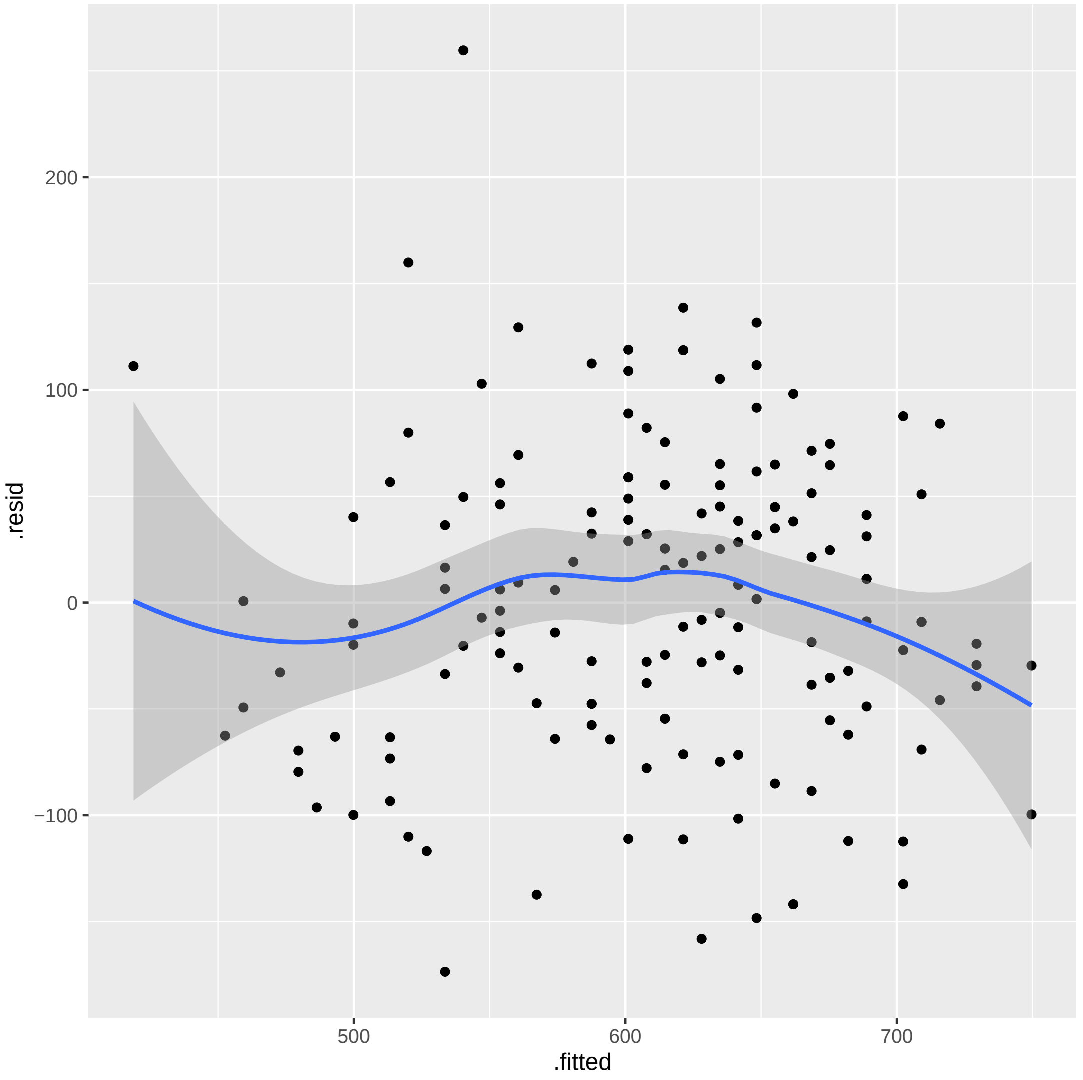


I see a big mess of randomness. The *only* concern I have is that point right at the top with a large positive residual, which looks like an outlier. Remember that we do not automatically remove outliers, because they might contain information; the thing to do with this one would be to identify it and see if there is anything that would make it different from the other observations. If there isn't, we are stuck with it.

I wouldn't add a smooth trend, because it tends to over-emphasize any patterns in the plot:

```
ggplot(sat.1, aes(x=.fitted, y=.resid)) + geom_point() + geom_smooth()

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



That suggests a downward-opening curve, which is really not supported by the points (that are mostly a long way from the trend).

Make a call about any pattern you see, with an extra reminder to *not look too hard*, because if you do, your brain will supply some pattern that is probably no more than randomness. You're looking for something you can describe fairly compactly, like a curved pattern. I don't see anything like that here.

Extra: I didn't want to make this assignment any longer, but what I myself would be interested in doing now is to add `sex` to the model, and see if there is actually any difference between males and females:

```
sat.2 <- lm(math_sat ~ verbal_sat + sex, data=sat)
summary(sat.2)
```

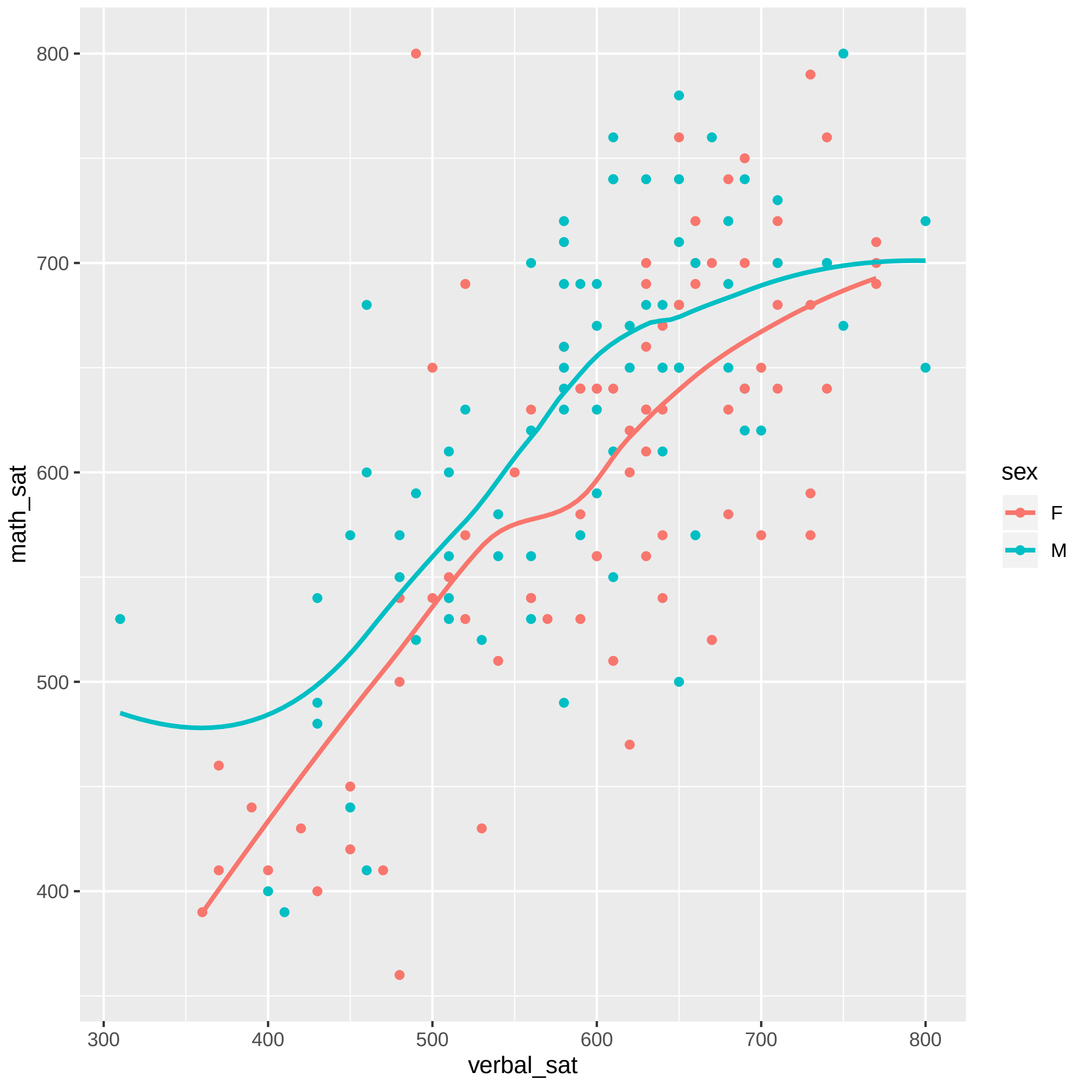
```
##
## Call:
## lm(formula = math_sat ~ verbal_sat + sex, data = sat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -167.786  -43.444   -2.023   44.512  279.214
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 184.58164   34.06782   5.418 2.19e-07 ***
## verbal_sat    0.68613    0.05513  12.446  < 2e-16 ***
## sexM         37.21856   10.93993   3.402 0.000846 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 69.49 on 159 degrees of freedom
## Multiple R-squared:  0.5047,Adjusted R-squared:  0.4985
## F-statistic: 81.02 on 2 and 159 DF,  p-value: < 2.2e-16
```

There is actually a significant difference between the sexes, in addition to the effect of verbal SAT score. What kind of difference? One of the sexes, namely `F`, is the baseline, and the Estimate for `sexM` says how males compare to the baseline. This one says that if you had two people with the same verbal SAT score, one of whom was male and the other female, the male would be expected to have a 37-point *higher* math SAT score.

How does that look on a scatterplot? If we add `sex` as a `colour` and ask for a smooth trend, we'll get a separate smooth trend for each sex:

```
ggplot(sat, aes(x=verbal_sat, y=math_sat, colour=sex)) + geom_point() +
    geom_smooth(se=F)
```

```
## ‘geom_smooth()‘ using method = ’loess’ and formula ’y ~ x’
```



The trend says that male math SAT scores are consistently higher on average than female ones, for the same verbal SAT score. But this is "on average": you can see that the data points are all over the place, and if you took an individual male and female with the same verbal SAT score, it's pretty much a coin toss which of them would have the higher math SAT score.

3. Work through (at least some of) the remaining problems in Chapter 14 of PASIAS. 14.1–14.4 are multiple regressions; 14.10 has a categorical variable in it, and the remaining problems are a mixture.