HW 5 Solutions

A. Download the apiscores data. A full explanation of this data set is given below in the Optional Problem B. Our goal is to understand what factors influence or affect high-school performance. (Note that this is probably not possibly since this is an observational study. Still we can gain some insight. A less ambigitious and more attainable goal would be to understand how schools with different compositions vary in terms of their API schools. For example, do all schools with a high percentage of teachers with emergency credentials score the same? If not, what accounts for this difference?) One approach to understanding data like this is to ask: what explains the variability in y? For example, it would be nice if all schools got the same score on the API -- and preferably this score would be a good one. But type hist(api) and you'll see that they don't. Why?

I meant this as a rhetorical question -- because it will largely be answered by this data analysis. But it does seem that school performance could be affected by a number of "external" factors above and beyond what is happening at the school. Socio-economic factors can be measured (loosely) by the percent of students who receive free lunches and by the educational background of the parents. Schools with many emergency credentialed teachers might mean that the faculty is less experienced, on average.

a) type (api.table). (I'm assuming you named the data api.table, but if not, type in whatever you did name it.) Notice that you get a large, messy graph. (You might want to make it larger to make any sense of it.) Notice that if you type school[1:2] you'll see that this variable contains the names of the schools. But for some reason R just converted these names to numbers and plotted them. We're not interested in treating school-name as a variable, so first we need to create a new dataframe that does not have the school variable:

api.small <- api.table[, -1]

This "subtracts" the first column (which contains the school variable) and assigns it to api.small.

Now do plot(api.small). The first row contains what are called "marginal plots". These are the same plots you would get if you typed
plot(pct.meals, api99) and then plot(not.high.g, api99), etc. The first one, for example, tells us what the relationship between api scores and the percentage of students at a school who receive free meals looks like, ignoring all of the other factors.

Describe this first row. How do each of these variables relate to the API scores? What relationships look unusual?

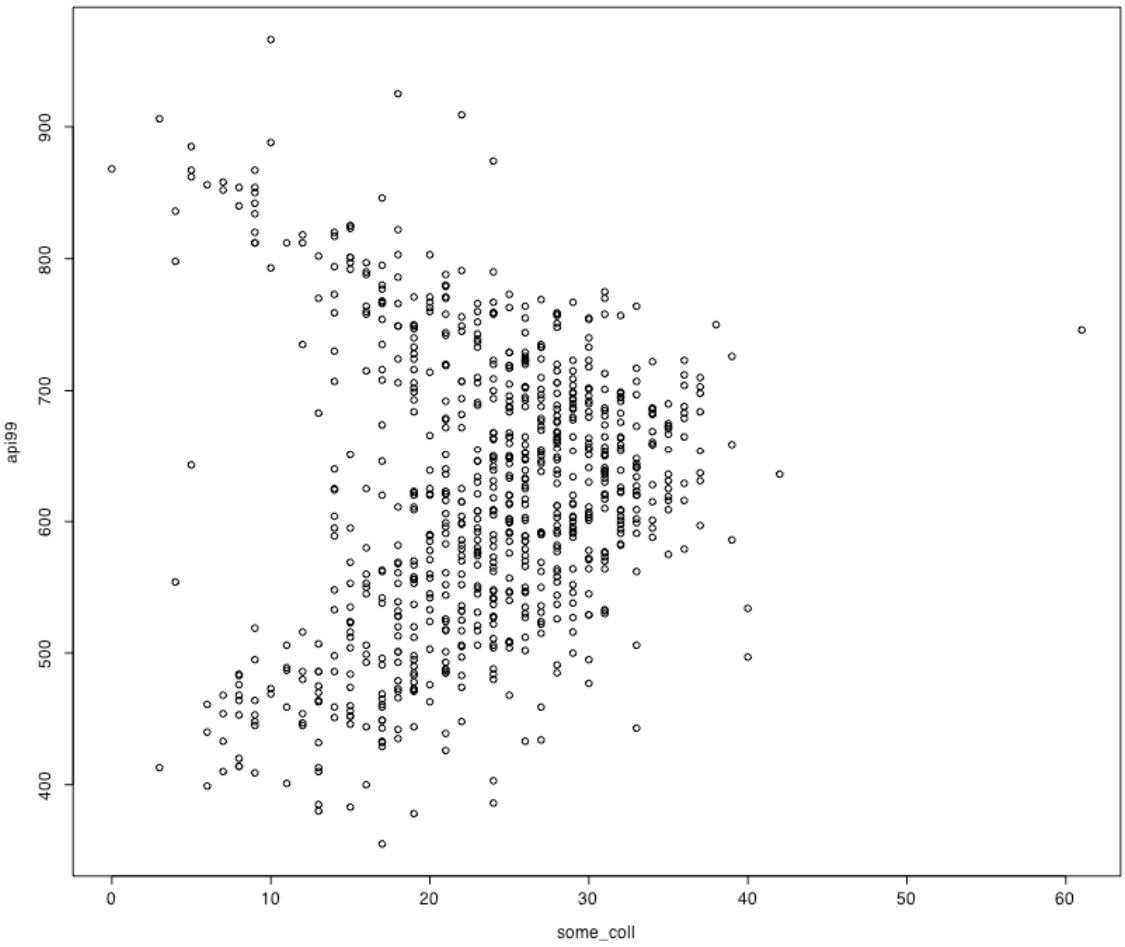
A general caution to everyone: Avoid vague, colloquial language. when possible, I put comments on your assignments to point out where your language is vague. Part of the reason for this class is to give you a vocabulary of statistical terms that have some amount of precision to them.

Another general note: be aware of what the unit of analysis is. Each point on these plots represents a school, and not a person. So don't say things like, "people's scores go up when they have higher education" or something like that.

The top row of the scatterplot is quite small, and if you have trouble seeing trends, you should look at individual scatterplots. We are trying to understand the relationship between these variables, and if you can't tell from one plot, try another.

The relationship between API scores and schools is, for the most part, a negative (and roughly linear) trend. Schools with a high percent of students receiving free lunch tend to have lower API scores. But note one very interesting feature. There are a large number of schools that give no free lunches, and these schools do not fit this pattern. This is some indication that there might be two distinct populations of schools that need to be analyzed separately.

The other interesting plot is the one comparing API scores to the percent of parents who had some college (but no more). There appear to be two very distinct groups of schools. For one group, the association is negative: schools with a high percentage of parents who had only some college tend to have low API scores. But for another group the trend is positive. Now, I'm not certain these are distinct groups. But it's important to note that there is no mathematical function that can fit this trend.

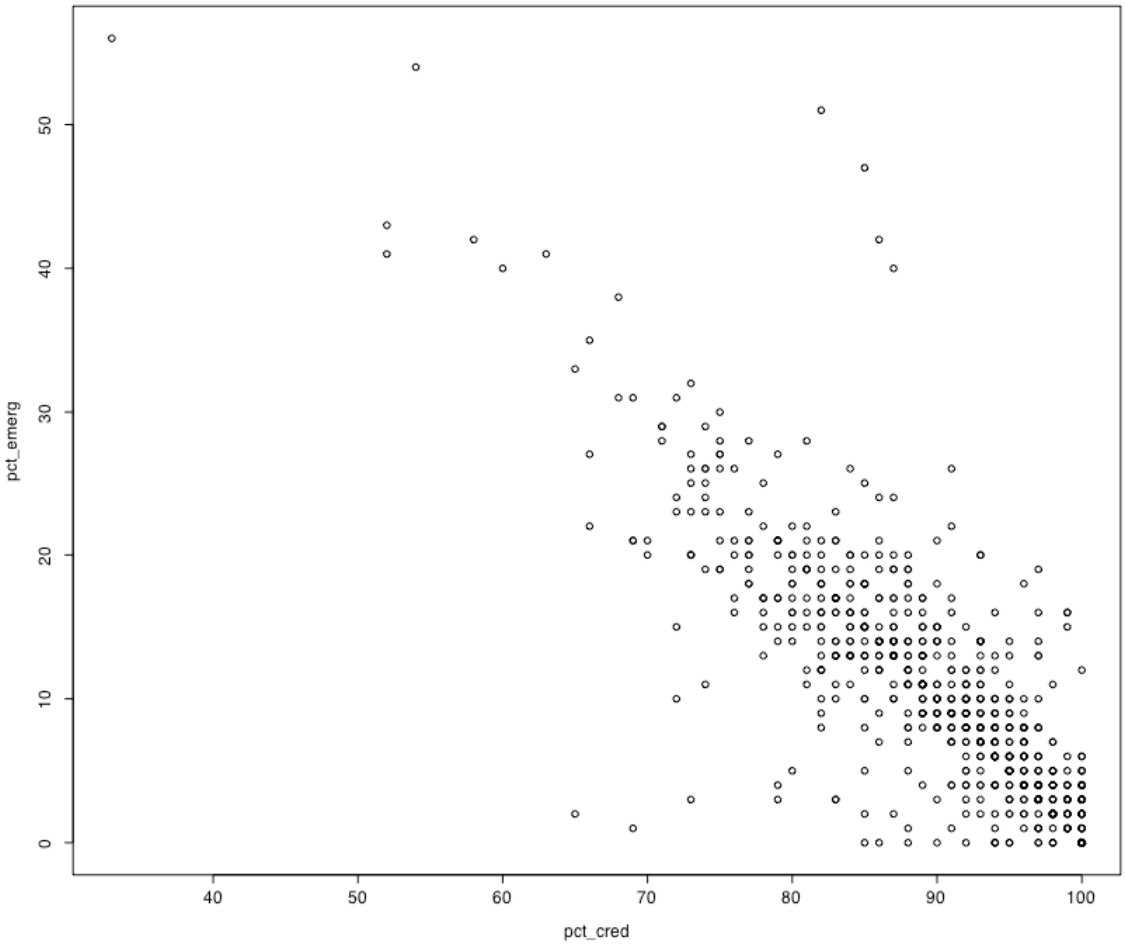


What else do we see? Schools with a high percentage of parents who did not graduate from high school tend to have low API scores. Schools with a high percentage parents who had only a high school degree (but no higher) tend to have lower API scores (but note some potential outliers). Schools with a high percentage of parents receiving college degrees tend to have higher API scores, as do schools with a high percentage of parents with some graduate school (but note a non-linear relationship). Schools whose parents have a high average educational level tend to have higher API scores. Schools with a higher percentage of credentialed teachers tend to have higher API scores (but some large outliers), and the opposite is true for schools with a high percentage of emergency credentialed teachers. (In a perfect world, the percent of emergency credentials plus the percent of other credentials should add to 1, but apparently here they don't.)

b) Later we'll see that relationships between our predictor variables can be problematic. Do you see any evidence that the predictor variables are related? Give an example of two related predictor variables and describe that relationship in the context of the data. (In other words, don't just say "x has a linear relationship with y",

but instead explain what these means in terms of schools/parents/teachers/ etc.)

Many of the predictor variables are associated with each other. But it seems that the percent of teachers with emergency credentials should be related to the percent with regular credentials:



In fact we see that schools with a high percentage of credentialed teachers tend to have a low percentage of emergency credentials. Although there is one outlier, it also fits this trend.

c) You'll notice that the relationships between the API scores and the various predictors are not terribly linear. Nonetheless, we'll plow ahead and fit a linear model. There are two ways to do this:

The easy way:

fullfit <- lm(api99 ~ . , data=api.small)

The "." on the right hand side means "all of the variables except api99 that are in the dataframe api.small"

The hard way:
attach(api.small)
fullfit <- lm(api99 ~ pct.meals + not.high.g + high.grad+some.coll+coll.grad+grad.schl+avg.ed+pct.cred+pct.emerg)

One thing to be concerned about: if in problem A you had typed "attach(api.table)", before attaching a new table (api.small) you should first detach the old one. So type "detach(api.table)" and then you can type "attach(api.small)". The reason for this is that both tables have the same variable names, and this will lead to confusion later.

Write down the equation of the fitted model. (Type "summary(fullfit)")

Predicted API score = 896 - .02 (pct_meals) - 4.01 (not_high_g) - 2.59 (high+grad) + .82 (some_coll)+ 3.37*coll_grad + 6.56 * grad_schl

d) Note that the column of p-values ("Pr(>|t|)") has only one value less than 5%. Strictly speaking, this means that all of the slopes are not different from 0! This can't be correct! The problem is that the violation of the assumption of linearity (we assumed all relationships were linear, and they're not) has severe repercussions. (There are other problems, too, but we'll get to those much later.) But for now, assume that everything is fine and the relationships are linear. What does the model tell us about the role that emergency credentials play in school performance? (Be as detailed as you can.)

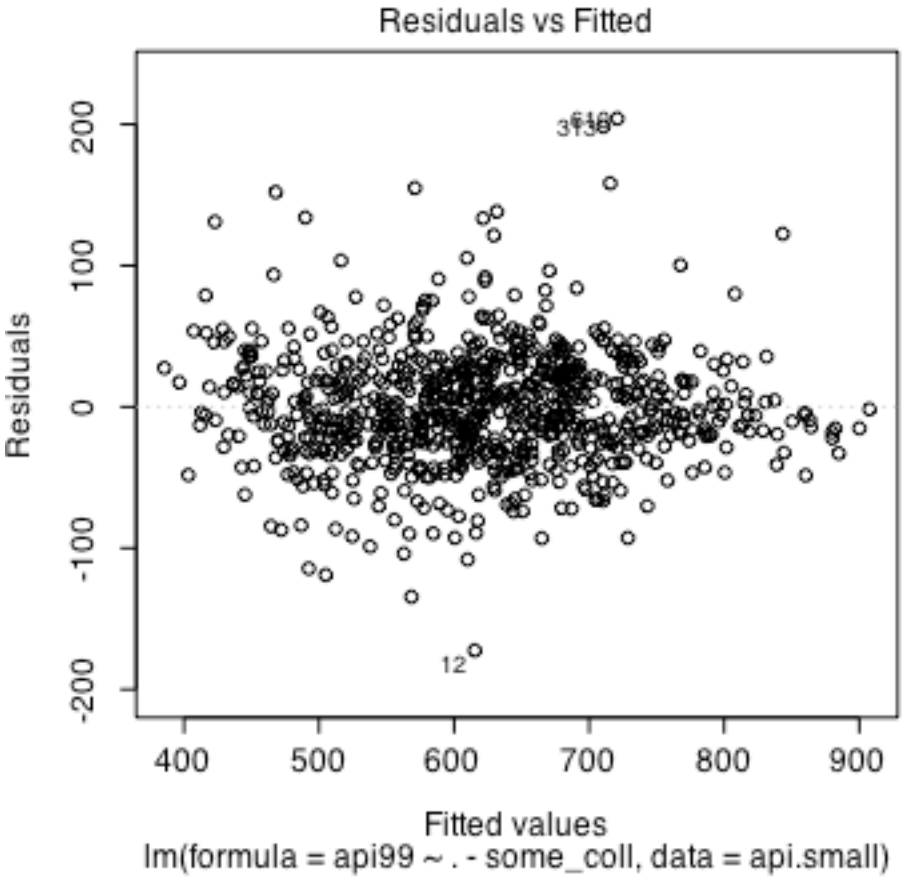
pct_emerg is the only statistically significant variable (although grad_schl -- the % of parents who went to graduate school -- is marginally significant.) The interpretation of this coefficient is that, all other factors being the same, schools with a larger percentage of teachers with emergency credentials tend to have lower API scores. In fact, each

additional percent increase of emergency credentials is associated with an average decline of 1.5 API points.

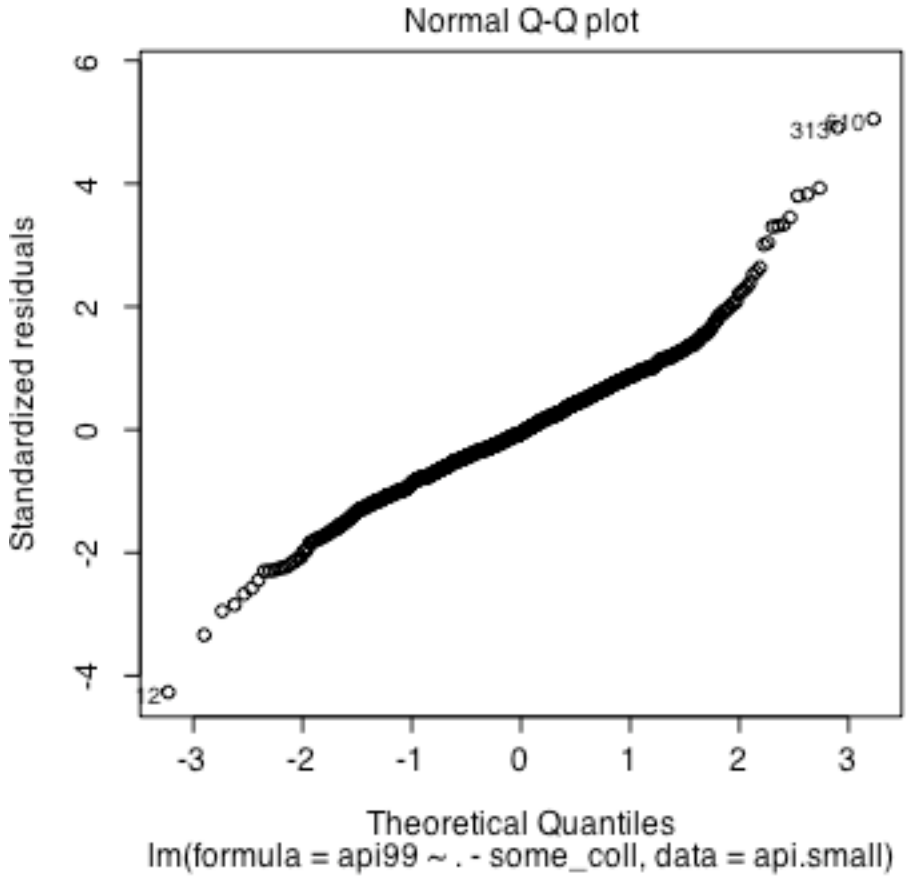
e) The least significant preditor is some_coll. Refit the model without it. How does this change the summary?

Although none of the coefficients changed dramatically, their statistical significance did. The model now suggests that the percent of parents who graduated high school (but did not go to college), the percent who went to grad school, and the percent of emergency teachers, are all useful predictors.

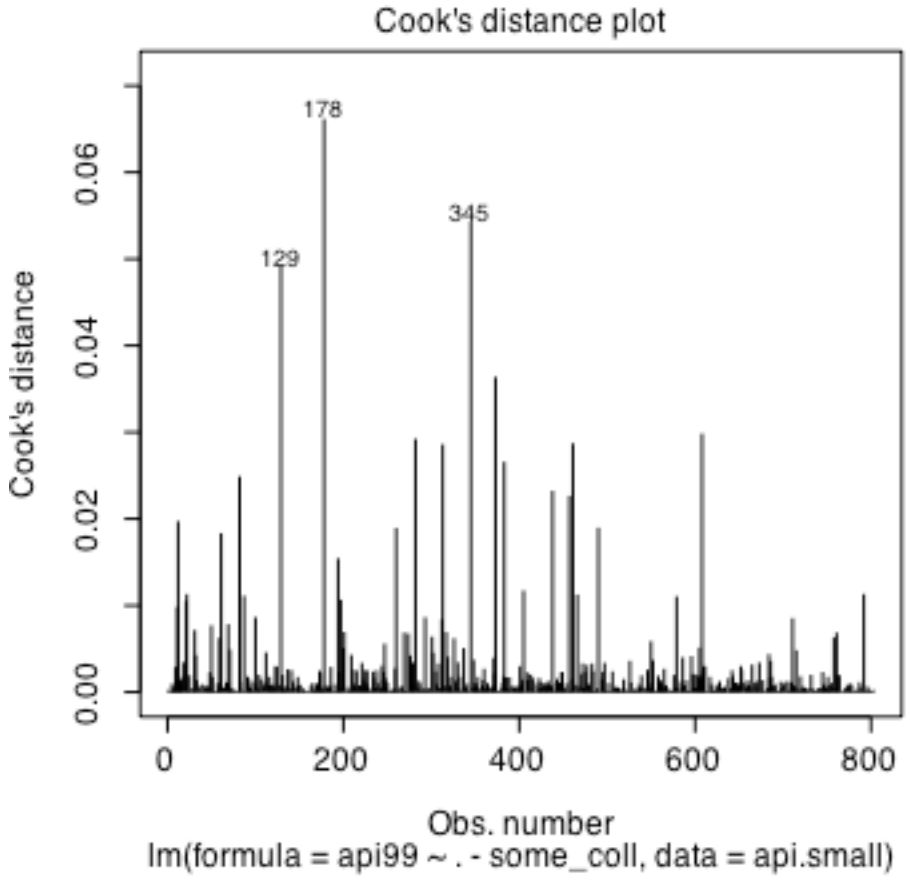
f) Check the diagnostic plots of this new model. Describe. (If the new model was named fitwithoutcoll.lm type plot(fitwithotucoll.lm))



This plot suggests the linearity assumption is valid -- there are no systematic trends in the residuals.



The residuals are fairly normally distributed, except in the tails.



There are no observations that appear to be particularly more influential than the others, although 178, 345, and 129 are the most influentila.

We should be cautious though. Although it appears the model fits relatively well, we know for a fact that the relation between the response and several of the predictors is non linear. This means that if we were to fix that, perhaps these predictors would become signifcant in the full model and we would achieve a better fitting model with greater predictive accuracy.