Homework 4

Advanced Methods for Data Analysis (36-402)

Due Friday February 15, 2019, at 3:00 PM

See the syllabus for general instructions regarding homework. Note that you should **always show all your work** and submit **exactly two files** (Rmd or R file as *code*, and knitted or scanned/merged pdf or html as *writeup*). Make sure that everything you submit is readable.

Problem 1: The Omitted Variables Effect for Categorical Data

In 1973, the University of California at Berkeley feared that they would be sued for gender bias in their graduate school admissions.
Table 1 shows the numbers of applicants who were admitted and rejected (by the six largest departments,) tabulated by sex.[1]

Table 1: 1973 Berkeley graduate admissions for six largest departments by sex.

	Male	Female
Admitted	1198	557
Rejected	1493	1278

Interest lies in the effect (if any) of sex on admission status. Table 1 might seem to reveal gender bias.

Part a

Show that the proportion of male applicants that were admitted is higher than the proportion of female applicants that were admitted.

Table 2 gives the same data further tabulated according to the six different departments involved in Table 1.

Table 2: 1973 Berkeley graduate admissions by sex and department.

	Department A		Department B		Department C	
	Male	Female	Male	Female	Male	Female
Admitted	512	89	353	17	120	202
Rejected	313	19	207	8	205	391

	Department D		Department E		Department F	
	Male	Female	Male	Female	Male	Female
Admitted	138	131	53	94	22	24
Rejected	279	244	138	299	351	317

Remark: Let Y_j be the binary random variable taking the value 1 if the jth person was admitted and 0 if not. Let X_j be the binary random variable taking the value 1 if the jth person was female and 0 if male. Let Z_j be the categorical random variable taking the values A, B, C, D, E, F that indicate the department.

[1] The data for this problem are available in *R* as the object `UCBAdmissions` if you attach the `graphics` library into your workspace, e.g. `library(graphics)`. The `graphics` library is built into *R*, but not attached by default.

Part b

Show using the data in Table 2 that females are admitted at a higher rate than males by most of the six departments, and say which departments they are. Based on those few departments where males are admitted at a higher rate, say why this might be called a "near example" of Simpson's paradox.

Part c

Compute using the data in Table 2 an estimate of the *conditional regression*

$$r(x, z) \equiv \mathbb{P}(Y = 1 | X = x, Z = z)$$

of Y on X given $Z = z$ for each of the six departments (values of Z.) Call this estimate $\hat{r}(x, z)$.

Part d

Assume that `Department` is the *only* confounding variable. Compute an estimate of the *adjusted treatment effect* of X on Y, (that is, an estimate of the *casual* regression function $\theta(x)$). Did it make a difference to adjust for Department?

Part e

Draw a plot with all six conditional regression lines computed in part (c). That is, on a single set of axes, for each $z = 1, \ldots, 6$, plot the line connecting the point $(0, r(0, z))$ to the point $(1, r(1, z))$. Add the marginal association line that estimates $\mathbb{P}(Y = 1 | X = x)$ (computed from Table~1) to the plot. Finally, add the estimated adjusted treatment effect line computed in part (d).

Part f

In part (d) you computed an estimate of the adjusted effect of X on Y which is $\mathbb{E}[r(x, Z)]$, where the expected value is with respect to the marginal distribution of Z. According to Remark 16.7 in AOS, the unadjusted effect, which you estimated in part (a), is the regression of Y on X alone, namely $r(x) = \mathbb{E}[r(x, Z)|X = x]$, that is, the expected value of $r(x, Z)$ with respect to the conditional distribution of Z given $X = x$. Using the tables of counts, compute estimates of the two conditional distributions of Z (Department) given $X = 0$ (male) and $X = 1$ (female). Plot these on a common set of axes together with the marginal distribution of Z. Based on what you see in the resulting plot, explain why the estimate of the adjusted treatment effect computed in part (d) is different from the estimate of the regression of Y on X, which was computed in part (a).

Part g

First use the law of iterated expectations and definitions to verify that

$$r(x) = \sum_z r(x, z) \mathbb{P}(Z = z | X = x),$$

where as usual we define $r(x) = \mathbb{E}(Y | X = x)$. Then verify with data that the value $\hat{r}(1)$ that you computed in part (a), by ignoring Z altogether, is indeed the same as

$$\sum_z \hat{r}(1, z) \hat{\mathbb{P}}(Z = z | X = 1),$$

using the estimated conditional regression $\hat{r}(1, z)$ in part (c), and the estimated conditional distributon $\hat{\mathbb{P}}(Z = z|X = 1)$ in part (f). In other words, this kind of averaging accomplishes nothing beyond regressing Y directly on X!

Problem 2: SAT Scores Data

In 1982, average SAT scores were published with breakdowns of state-by-state performance in the United States. The average SAT scores varied considerably by state, with mean scores falling between 790 (South Carolina) to 1088 (Iowa).

Two researchers examined compositional and demographic variables to examine to what extent these characteristics were tied to SAT scores. The variables in the data set were:

1. `state`: state name
2. `sat`: mean SAT score (verbal and quantitative combined)
3. `takers`: percentage of total eligible students (high school seniors) in the state who took the exam
4. `income`: median income of families of test takers, in hundreds of dollars
5. `years`: average number of years that test takers had in social sciences, natural sciences, and humanities (combined)
6. `public`: percentage of test takers who attended public schools
7. `expend`: state expenditure on secondary schools, in hundreds of dollars per student
8. `rank`: median percentile of ranking of test takers within their secondary school classes. Possible values range from 0-99, with 99th percentile students being the highest achieving.

Notice that the states with high average SATs had low percentages of takers. One reason is that these are mostly midwestern states that administer other tests to students bound for college in-state. Only their best students planning to attend college out of state take the SAT exams. As the percentage of takers increases for other states, so does the likelihood that the takers include lower-qualified students.

This homework closely mirrors our next Tuesday (February 12) class demo. Some starter code for the EDA has been provided on Canvas.

a. Exploratory Data Analysis

Conduct an EDA on the SAT dataset. For instance, get a sense of the marginal distribution of each of the variables and the pairwise correlations. Identify if any observation appears to be unusual. Basically, spend some time investigating the data. Write a two-paragraph summary of what you see supported by some plots and / or summary tables.

b. Using Residuals to Create Better Rankings for SAT Data

First rank the states based on raw SAT scores. This approach however doesn't seem reasonable: Some state universities require the SAT and some require a competing exam (the ACT). States with a high proportion of takers probably have *in state* requirements for the SAT. In states without this requirement, only the more elite students will take the SAT, causing a bias. We would like to rank the states by SAT scores, corrected for percent taking the exam, the median class rank and expenditure for secondary school (expenditure). Let's explore this thinking further.

To address the research question of how the states rank after accounting for the percentage of takers, median class rank and expenditure, we define a reduced model that fits the regression line of `sat` on `takers`, `rank`

and `expend`. Instead of ranking by actual SAT score, we can then rank the schools by how far they fall above or below the fitted regression line value. A residual is defined as the difference between the observed value and the predicted value.

Sort the states by residual value and display the old ranking next to each state name. What do you see? Do the rankings shift once we control for the variables `takers`, `rank` and `expend`?
Interpret and discuss your results. Find the state that rose the most in the rankings and the state that fell the most. For those two states, explain why their ranks changed so much.

c. Check the Residuals of the Reduced Model

One of the assumptions of the basic regression model is that the magnitude of residuals is relatively constant at all levels of the response. It is important to check that this assumption is upheld here. Hence, plot the residuals of the reduced model versus fitted values, `takers`, `rank` and `expend`. Do you see any patterns in the residual plots?

d. Was there a Better Reduced Model? (Part 1)

Does it appear that any transformations of variables would improve the relationships between the variables in the reduced model? Would it make sense to account for such things as expenditures or income and what information would that then give you? Propose and discuss alternative "candidate" reduced models. (You don't need to fit any of the models you propose.)

e. Was there a Better Reduced Model? (Part 2)

Look carefully at the relationship between `sat` and `takers`. First, notice that there are no states with `takers` between 20 and 30. Describe how the relationship between `takers` and `sat` appears markedly different in those states with `takers` below 20 from the relationship in those states with `takers` above 30. Describe a model that would accommodate such a difference in relationships. (You don't need to fit the model you describe.)

f. Fitting a model with all the variables

Now, try to fit a model with all the variables. Is an increase in the number of years of classes associated with a significant increase in SAT score, after taking into account all the other variables? Think about some ways to approach this question. What is the null hypothesis? How would you assess its plausibility?

When would this association (if any) be useful to guide policy? That is, when would this association be actual causation? Briefly elaborate on potential confounders.