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.

Solution

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

```
source("plot_theme.R")
library(graphics)
dat <- UCBAdmissions
## Compute margins of the table
datm <- margin.table(dat, c(1,2))
get_prop <- function(x) { x/sum(x) }
apply(datm, 2, get_prop)
```

```
##           Gender
## Admit           Male    Female
##   Admitted 0.4451877 0.3035422
##   Rejected 0.5548123 0.6964578
```

As we can see from the table above, the admission rate for males is 0.45 vs 0.30 for females.

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.

Solution

```
depnames <- c("A","B","C","D","E","F")
cadmit <- matrix(0, ncol = 2, nrow = length(depnames),
                 dimnames = list(depnames, c("Male", "Female")))
for(j in depnames){
      cadmit[j, ] <- apply(dat[, , j], 2, get_prop)["Admitted", ]
}
print(cadmit)
```

```
##         Male     Female
## A 0.62060606 0.82407407
## B 0.63035714 0.68000000
## C 0.36923077 0.34064081
## D 0.33093525 0.34933333
## E 0.27748691 0.23918575
## F 0.05898123 0.07038123
```

```
## Get names of Depts where admission rate for females is greater than for males
rownames(cadmit)[which(cadmit[, "Male"] < cadmit[, "Female"])]
```

```
## [1] "A" "B" "D" "F"
```

So Deptarments A, B, D, and F show admission rates that are higher for females than for males.

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 $\hat{r}(x, z)$.

Solution

Same computations as in 1b.

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?

Solution

Assuming that `Department` is the *only* confounding variable, means that $Y^a \perp\!\!\!\perp X|Z$. That is, within the department, sex is independent of the admission outcome, (ie as if we would flip a coin to decide admission). We still assume consistency, that is $Y = Y^1 \mathbb{1}\{X = 1\} + Y^0 \mathbb{1}\{X = 0\}$. Therefore, the causal regression function can be computed as

$$\begin{aligned}
\text{Prob of Admission if Female} &\equiv \mathbb{P}(C(X = 1) = 1) \\
&= \mathbb{E}_Z \left\{ \mathbb{P}(C(X = 1)|Z) \right\} \\
&= \mathbb{E}_Z \left\{ \mathbb{P}(C(X = 1) = 1|Z, X = 1) \right\} \\
&= \mathbb{E}_Z \left\{ \mathbb{P}(Y = 1|Z, X = 1) \right\} \\
&= \sum_z \mathbb{P}(Y = 1|Z = z, X = 1)\mathbb{P}(Z = z)
\end{aligned}$$

where the first line is by definition, the second equality follows by the law of iterated expectation, the third equality follows because $Y^a \perp\!\!\!\perp X|Z$, the fourth equality holds by the consistency assumption, the fifth equality is just a restatement of the fourth.

A similar logic allows to identify the probability of admission for males.

We will use the empirical distribution to compute the quantities above, namely the proportions of the total sample that come from each department.
Call this empirical distribution $\hat{f}_Z(\cdot)$. For each x, this is the inner product of the vector $\hat{r}(x, \cdot)$ with the vector $\hat{f}_Z(\cdot)$, hence the the whole $\hat{\theta}(\cdot)$ vector is the product of the $\hat{f}_Z(\cdot, \cdot)$ matrix times the $\hat{f}_Z(\cdot)$ vector:

```
# Marginal distribution of "Department" f-hat(z):

zdist=apply(dat,3,sum)
zdist=zdist/sum(zdist)

# Adjusted effect theta-hat(x):

adjeff=matrix(zdist,1,6)%*%cadmit
adjeff
```

```
##           Male    Female
## [1,] 0.3873186 0.4299554
```

The adjusted effects of sex on admission are 0.3873 for males and 0.4300 for females. These are in the opposite order of the unadjusted effects computed in part (a), so adjusting makes a difference.

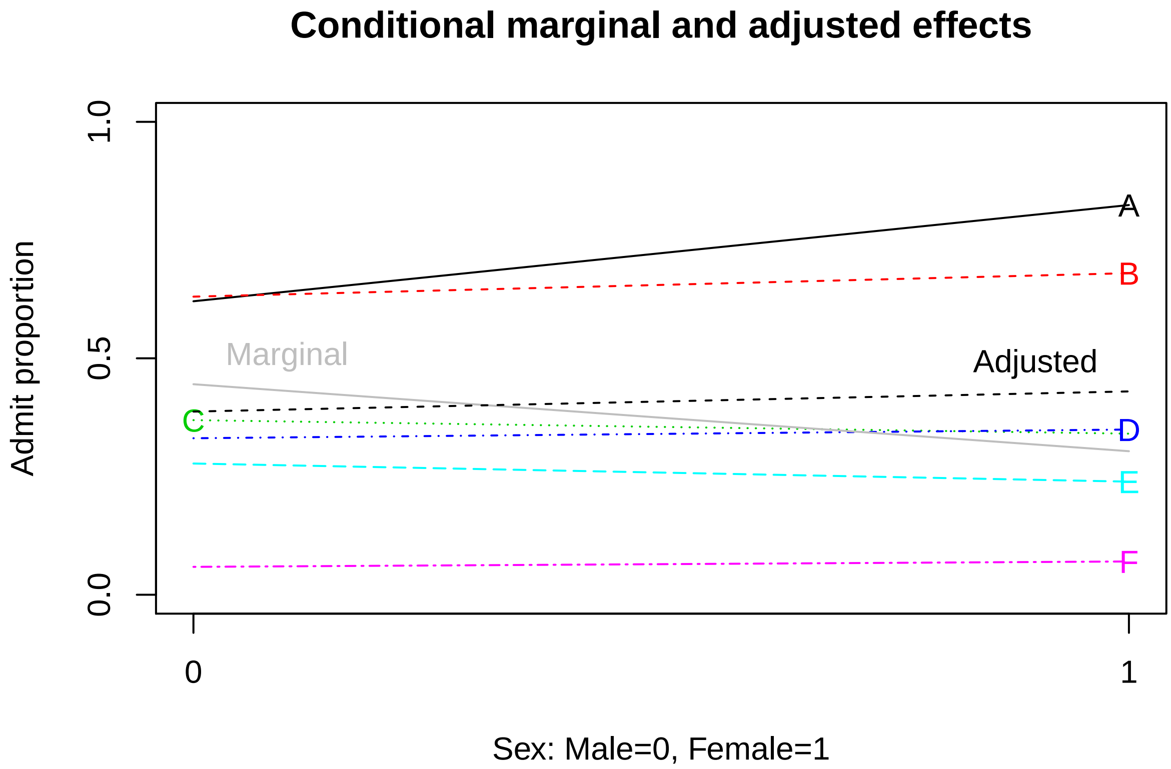
Part e

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).

Solution

Next, we plot the marginal, conditional and adjusted effects together for visual comparison.

```r
# Start with the conditional effects

plot(c(0,1),c(0,1),pch="",xlab="Sex: Male=0, Female=1",ylab="Admit proportion",
lab=c(1,3,7),main="Conditional marginal and adjusted effects")
wt=c(1,1,0,1,1,1) # where to place the labels
for(j in 1:6){
       lines(c(0,1),cadmit[j,],lty=j,col=j)
       text(wt[j],cadmit[j,1+wt[j]],depnames[j],col=j)
}

# Add the marginal association line

lines(c(0,1),apply(datm, 2, get_prop)["Admitted", ],lty=7,col=8)
text(.1,apply(datm, 2, get_prop)["Admitted", ][1],"Marginal",pos=3,col=8)

# Add the adjusted effect

lines(c(0,1),adjeff,lty=8,col=9)
text(.9,adjeff[2],"Adjusted",pos=3,col=9)
```



Part f

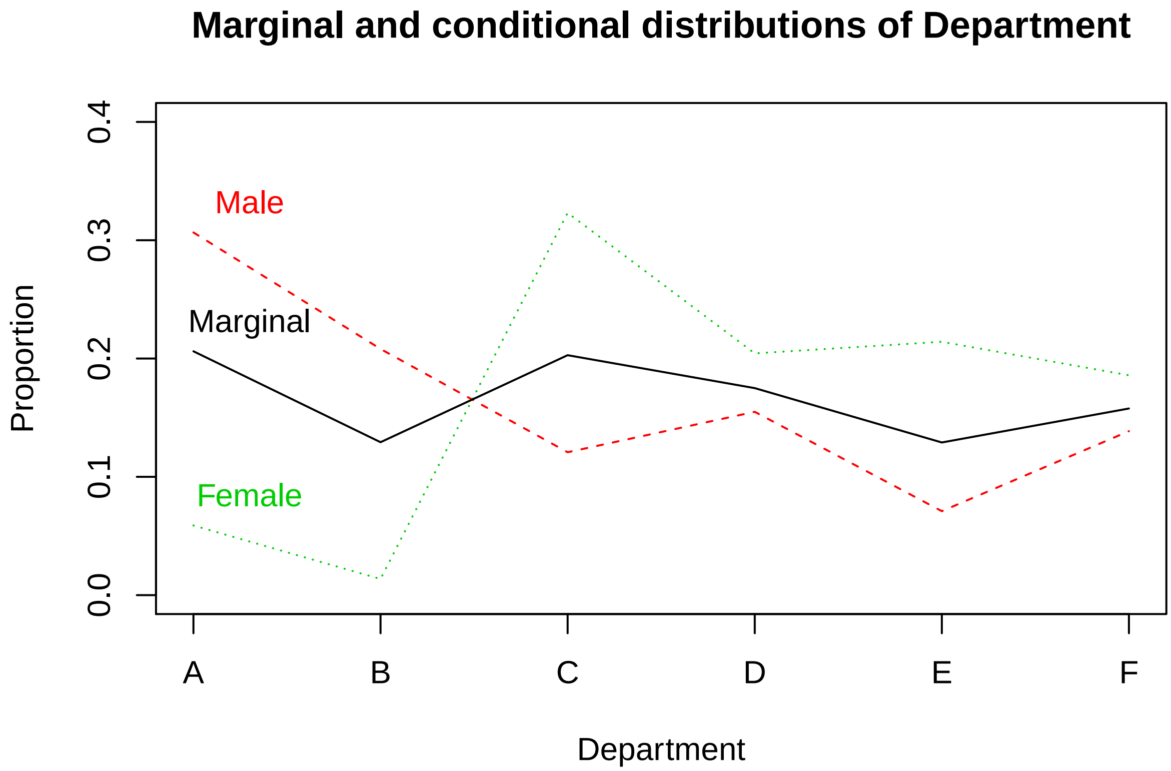
In part (d) you computed an estimate of the adjusted effect of X on Y which is $\mathbb{E}[r(x, Z)]$, where the expected value is with respect to the marginal distribution of Z. According to Remark 16.7 in AOS, the unadjusted effect, which you estimated in part (a), is the regression of Y on X alone, 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).

Solution

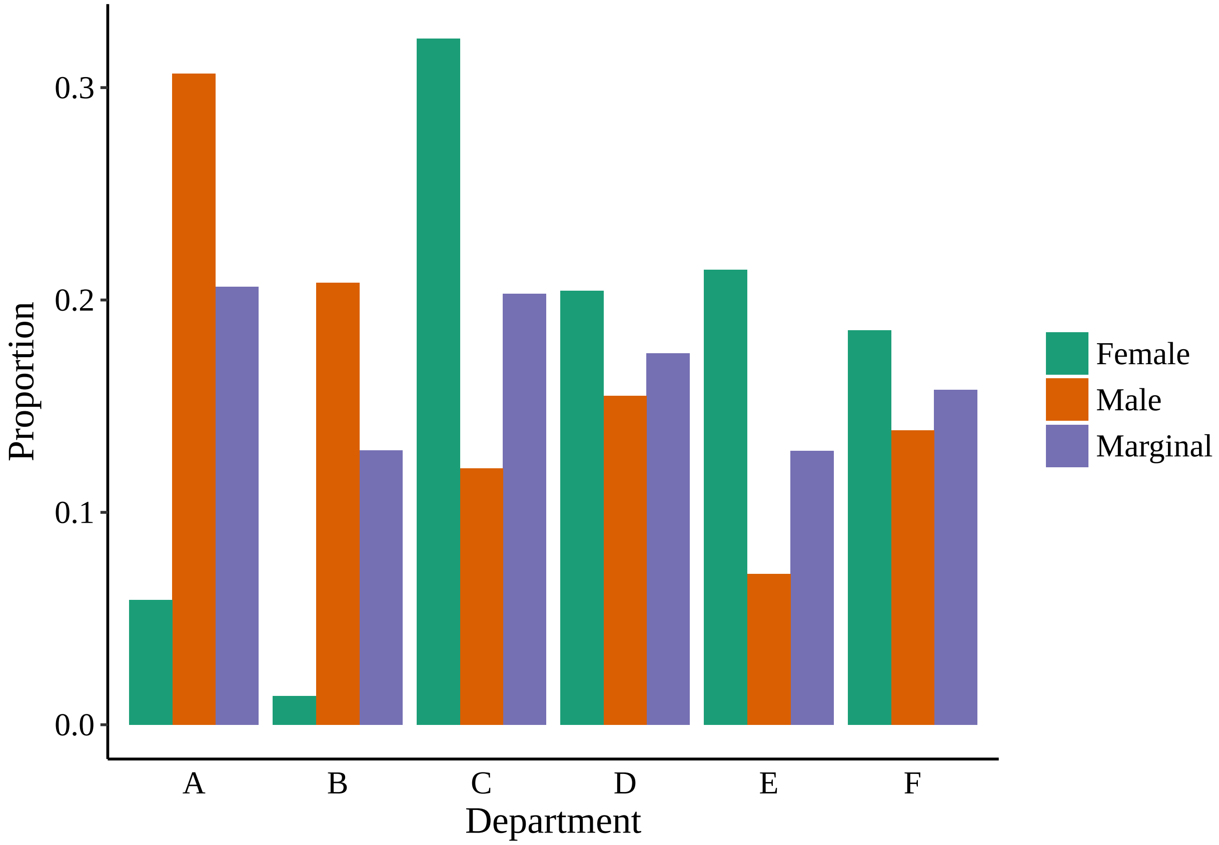
Finally, we compare the conditional distributions of Z given $X = 0$ and $X = 1$ and the marginal distribution of Z. Remark 16.7 in *All of Statistics* gives a formula for the marginal association of X and Y (the unadjusted effect) in terms of the conditional distribution of Z given X: $r(x) = \mathbb{E}[r(x, Z)|X = x]$. But if X and Z are independent (equivalently, if the conditional distribution of Z given $X = x$ is the same as the marginal distribution of Z for all x, the adjusted and unadjusted effects are the same.

```
#  Conditional distribution estimates

zdist0=apply(dat[,1,],2,sum)
zdist0=zdist0/sum(zdist0)
zdist1=apply(dat[,2,],2,sum)
```

```
zdist1=zdist1/sum(zdist1)

zdistgender <- data.frame(prop=c(zdist0, zdist1, zdist),
                          sex=c(rep("Male", length(depnames)),rep("Female", length(depnames)),
                                rep("Marginal", length(depnames))),
                          dept=rep(depnames, 3))

#  Plot the distribution estimates

plot(c(1:6),c(.4,.4,.4,0,0,0),pch="",xlab="Department",
ylab="Proportion",xaxt="n",yaxt="n",
main="Marginal and conditional distributions of Department")
axis(1,1:6,labels=c("A","B","C","D","E","F"))
axis(2,c(0,.1,.2,.3,.4),labels=c("0.0","0.1","0.2","0.3","0.4"))

# Conditional first:

lines(c(1:6),zdist0,lty=2,col=2)
lines(c(1:6),zdist1,lty=3,col=3)
text(1.3,zdist0[1],"Male",pos=3,col=2)
text(1.3,zdist1[1],"Female",pos=3,col=3)

# Then marginal:

lines(c(1:6),zdist,lty=1,col=1)
text(1.3,zdist[1],"Marginal",pos=3,col=1)
```



```
library(RColorBrewer)
colors = brewer.pal(8, "Dark2")

ggplot(zdistgender,aes(x=dept,y=prop,fill=factor(sex))) +
  geom_bar(stat="identity",position="dodge") +
  labs(y="Proportion", fill="", x="Department") +
  scale_fill_manual(values = colors[1:3]) +
  our_theme
```



It is easy to see that the three distributions are rather different, and this difference accounts for the difference between the marginal and adjusted effects of X on Y. In particular, in the plot in part (c), the departments with the highest admit proportions (A and B) are the departments with the lowest female application rate. Females seem to apply to the departments that had low admission rates for both sexes (C,D,E,F) and avoided the departments that had high admission rates for both sexes (A,B).

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!

Solution

For each department z, the estimate of the conditional regression function is $\hat{r}(x,z)$ equal to the proportion of those students of sex x in the applicant pool of department z who got admitted. These are the same as the proportions computed in part (b).

Similarly, we compute $\hat{\mathbb{P}}(Z = z|X = 1)$ as

```
rzx <- cadmit
depnames <- c("A","B","C","D","E","F")
pz_given_x <- matrix(0, ncol = 2, nrow = length(depnames),
                     dimnames = list(depnames, c("Male", "Female")))
totnum_f <- sum(datm[, "Female"])
totnum_m <- sum(datm[, "Male"])
for(j in depnames){
      pz_given_x[j, "Female"] <- sum(dat[,,j][, "Female"])/totnum_f
      pz_given_x[j, "Male"] <- sum(dat[,,j][, "Male"])/totnum_m
}

print(pz_given_x)
```

```
##         Male     Female
## A 0.30657748 0.05885559
## B 0.20810108 0.01362398
## C 0.12077295 0.32316076
## D 0.15496098 0.20435967
## E 0.07097733 0.21416894
## F 0.13861018 0.18583106
```

```
## Should yield c(1, 1) by Law of Total Probability
apply(pz_given_x, 2, sum)
```

```
##   Male Female
##      1      1
```

```
tau <- sum(pz_given_x[, "Male"]*rzx[, "Male"])
## Confirming we have not done anything strange
abs(apply(datm, 2, get_prop)["Admitted", "Male"] - tau)
```

```
## [1] 0
```

As noticed above, we can compute an estimate $\mathbb{E}(Y|X = 1)$ directly by computing the admission rate among males regardless of the Department. Alternatively, we can compute a weighted sum of the admission rates for males in each dept, where the weights are the estimates of the conditional probabilities of applying to a given dept given that the applicant is male. Formally, by yet another application of the law of iterated expectation, we have

$$\mathbb{E}\left\{Y|X = 1\right\} = \mathbb{E}_{Z|X=1}\left\{E\left\{Y|X = 1, Z\right\}\right\}$$

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.

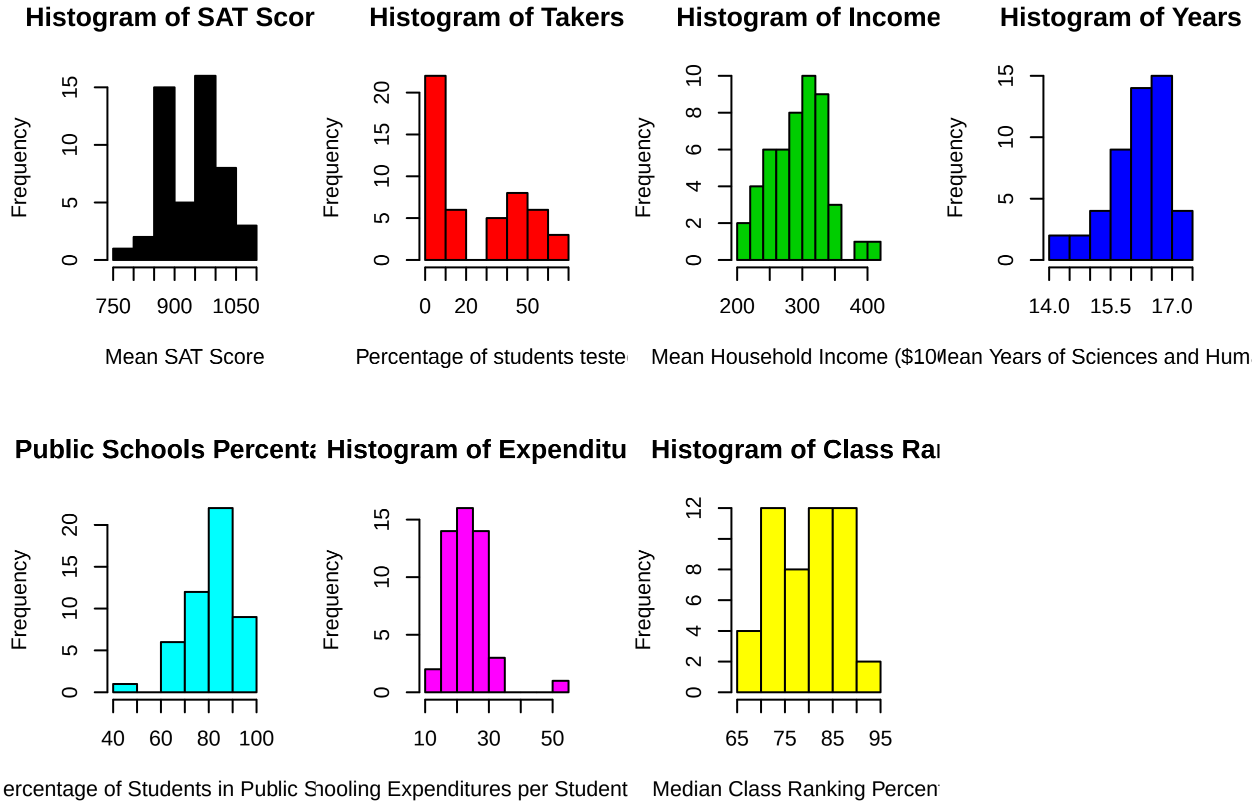
This homework closely mirrors our next Tuesday (February 12) class demo. Some starter code for the EDA has been provided on Canvas.

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.

Solution

```r
satdata = read.table("CASE1201.ASC", header = TRUE)
attach(satdata)
par(mfrow = c(2, 4))
hist(satdata$sat, main = "Histogram of SAT Scores", xlab = "Mean SAT Score", col = 1)
hist(satdata$takers, main = "Histogram of Takers",
xlab = "Percentage of students tested", col = 2)
hist(satdata$income, main = "Histogram of Income",
xlab = "Mean Household Income ($100s)", col = 3)
hist(satdata$years, main = "Histogram of Years",
xlab = "Mean Years of Sciences and Humanities", col = 4)
hist(satdata$public, main = "Public Schools Percentage",
xlab = "Percentage of Students in Public Schools", col = 5)
hist(satdata$expend, main = "Histogram of Expenditures",
xlab = "Schooling Expenditures per Student ($100s)", col = 6)
hist(satdata$rank, main = "Histogram of Class Rank",
xlab = "Median Class Ranking Percentile", col = 7)
```



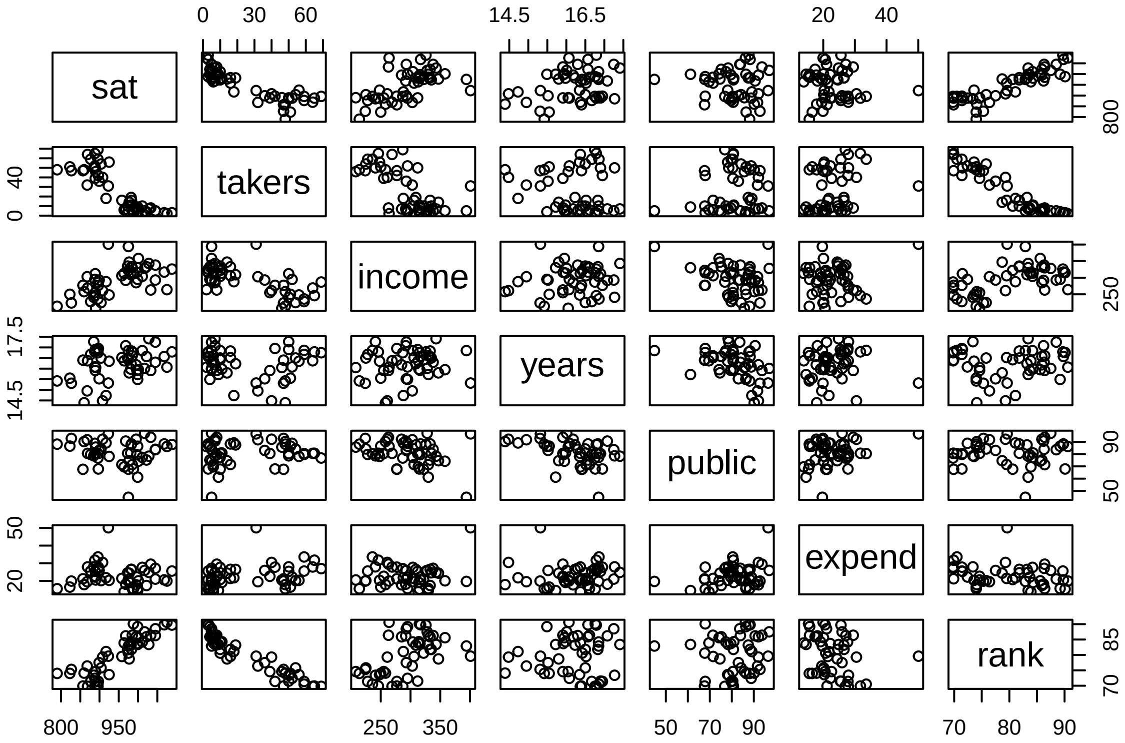
Exploratory data analysis allows us to look at the variables contained in the data set before beginning any formal analysis. First we examine the variables individually through histograms (Fig. 8). Here we can see the general range of the data, shape (skewed, gapped, symmetric, etc.), as well as any other trends. For example, we note that one state has almost double the amount of secondary schooling expenditures of all the other states. We may be interested in determining which state this is, and can do so in one line of code:

```
satdata[which(expend == max(expend)), ]
```

```
##     state sat takers income years public expend rank
## 29 Alaska 923     31    401 15.32   96.5   50.1 79.6
```

Next we look the variables together.

```
par(mfrow = c(1, 1))
plot(satdata[,-1])
```



The scatterplot matrix shows the relationships between the variables at a glance. Generally we are looking for trends here. Doed the value of one variable tend to affect the value of another? If so, is that relationship linear? These types of questions help us think of what type of interaction and higher order terms we might want to include in the regression model.

The scatterplot matrix shows clear relationships between sat, takers, and rank. Interestingly, we can also note Alaska's features, since we know it's the state with the very high 'expend' value. We can see that Alaska has a rather average sat score despite its very high levels of spending. For now we will leave Alaska in the data set, but a more complete analysis would seek to remove outliers and high influence points. In fact, this data-set contains two rather obvious outliers. One feature visible in both the scatterplot and the histogram is the gap in the distribution of takers. When there is such a distinct gap in a variable's distribution, sometimes it is a good idea to consider a transformation from a continuous variable to an indicator variable. Since subtle trends are often difficult to spot in scatterplot matrices, sometimes a correlation matrix can be useful, as seen above. Correlation matrices usually print 8-10 significant digits, so the use of the 'round' command tends to make the output more easily readible. We note that both the income and the years variables have moderately strong positive correlations with the response variable (sat). The respective correlations of 0.58 and 0.33 indicate that higher levels of income and years of education in sciences and humanities are generally associated with higher mean sat scores. However, this does not imply causation, and each of these trends may be nullified or even reversed when accounting for the other variables in the data set! A variable such a 'years' may be of particular interest to researches. Although neither science nor humanities are directly tested on the SAT, researchers may be interested in whether an increase in the number of years of such classes is associated with a significant increase in SAT score. This may help them make recommendation to schools as to how to plan their cirricula.

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.

Solution

First, we are asked to rank by SAT scores:

```
satorder=sort.list(sat,decreasing = T)
firsttable=data.frame(state=state,satrank=satorder)
# firsttable[satorder,]
```

It looks like the states were already ranked by SAT scores. Next, we will run the requested model and rank the states by their residuals.

```
smallm=lm(sat~takers+rank+expend)
summary(smallm)
```

```
##
## Call:
## lm(formula = sat ~ takers + rank + expend)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -78.30 -11.22   4.41  20.89  57.12
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 370.0643   180.6067   2.049  0.04619 *
## takers       -1.0075     0.6137  -1.642  0.10746
## rank          6.9108     2.0604   3.354  0.00160 **
## expend        2.2429     0.7601   2.951  0.00497 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.35 on 46 degrees of freedom
## Multiple R-squared:  0.8162, Adjusted R-squared:  0.8042
## F-statistic: 68.09 on 3 and 46 DF,  p-value: < 2.2e-16
```

```
rrank=sort.list(smallm$residuals,decreasing = TRUE)
nrank=NULL
nrank[rrank]=1:50
```

```
newtable=data.frame(state=state,newrank=nrank,oldrank=1:50)
newtable[rrank,]
```

```
##             state newrank oldrank
## 28  NewHampshire       1      28
## 35   Connecticut       2      35
## 1           Iowa       3       1
## 2    SouthDakota       4       2
## 3    NorthDakota       5       3
## 7      Minnesota       6       7
## 41 Massachusetts       7      41
## 13     Tennessee       8      13
## 8           Utah       9       8
## 36       NewYork      10      36
## 4         Kansas      11       4
## 21      Illinois      12      21
## 5       Nebraska      13       5
## 40      Virginia      14      40
## 32       Vermont      15      32
## 44     NewJersey      16      44
## 39      Maryland      17      39
## 43   RhodeIsland      18      43
## 34      Delaware      19      34
## 23      Missouri      20      23
## 6        Montana      21       6
## 10     Wisconsin      22      10
## 14     NewMexico      23      14
## 27          Ohio      24      27
## 37         Maine      25      37
## 18      Colorado      26      18
## 47        Hawaii      27      47
## 11      Oklahoma      28      11
## 15         Idaho      29      15
## 19    Washington      30      19
## 20       Arizona      31      20
## 42  Pennsylvania      32      42
## 22     Louisiana      33      22
## 24      Michigan      34      24
## 38       Florida      35      38
## 26       Alabama      36      26
## 9        Wyoming      37       9
## 17      Kentucky      38      17
## 46       Indiana      39      46
## 12      Arkansas      40      12
## 33    California      41      33
## 25  WestVirginia      42      25
## 16   Mississippi      43      16
## 31        Oregon      44      31
## 45         Texas      45      45
## 49       Georgia      46      49
## 30        Nevada      47      30
## 48 NorthCarolina      48      48
## 50 SouthCarolina      49      50
## 29        Alaska      50      29
```

```
changerank=newtable$oldrank-newtable$newrank
sum(abs(changerank)>10)
```

```
## [1] 27
```

There are some large jumps like NewHampshire that used to be 28 and is now 1, or Wyoming that was 9 and is now 37. Overall, 27 states moved up or down more than 10 spots in the ranking. We can identify the biggest movers using the variable `changerank` that we created.

```
# Find the state that went up the most
satdata[which(changerank==max(changerank)),]
```

```
##            state sat takers income years public expend rank
## 41 Massachusetts 888     65    246 16.79   80.7  31.74 69.9
```

```
# Find the state that went down the most
satdata[which(changerank==min(changerank)),]
```

```
##       state  sat takers income years public expend rank
## 9   Wyoming 1017      5    328 16.01   97.0  25.96 87.5
## 12 Arkansas  999      4    295 15.49   86.4  15.71 89.2
```

We see that Massachusetts jumped 34 spots while Arkansas and Wyoming fell 28 spots. Looking at Massachusetts first, we see that it has nearly the smallest fitted value. The actual `sat`$=$ 888 is low, but not near the smallest, so the residual is quite large. Arkansas, on the other hand, has nearly the largest fitted value, while its actual `sat`$=$ 999 is around the 80th percentile giving it a low residual.

Check the lecture notes for a more detailed discussion of why ranking by residuals is like changing the baseline for comparison from a horizontal line $Y = 0$ to a “inclined” line (regression line), which tells us “what should be expected” from a given state.

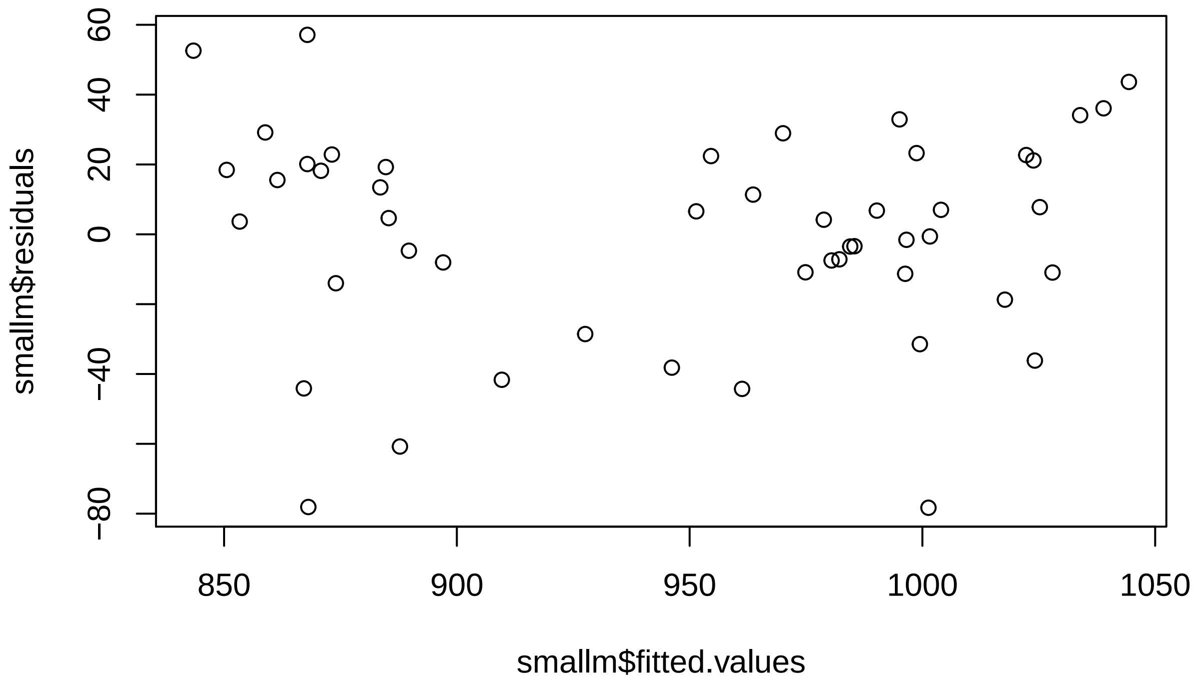
c. Check the Residuals of the Reduced Model

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?

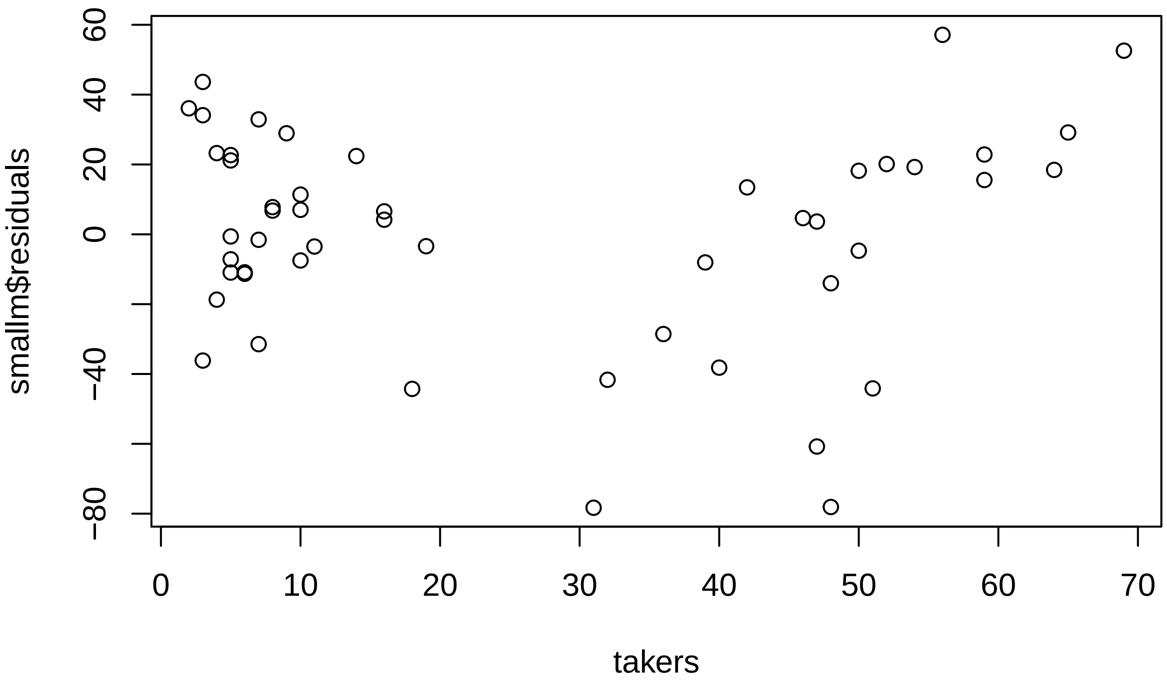
Solution

We begin with the requested residual plots. One should also plot residuals against other variables as mentioned in part (d).

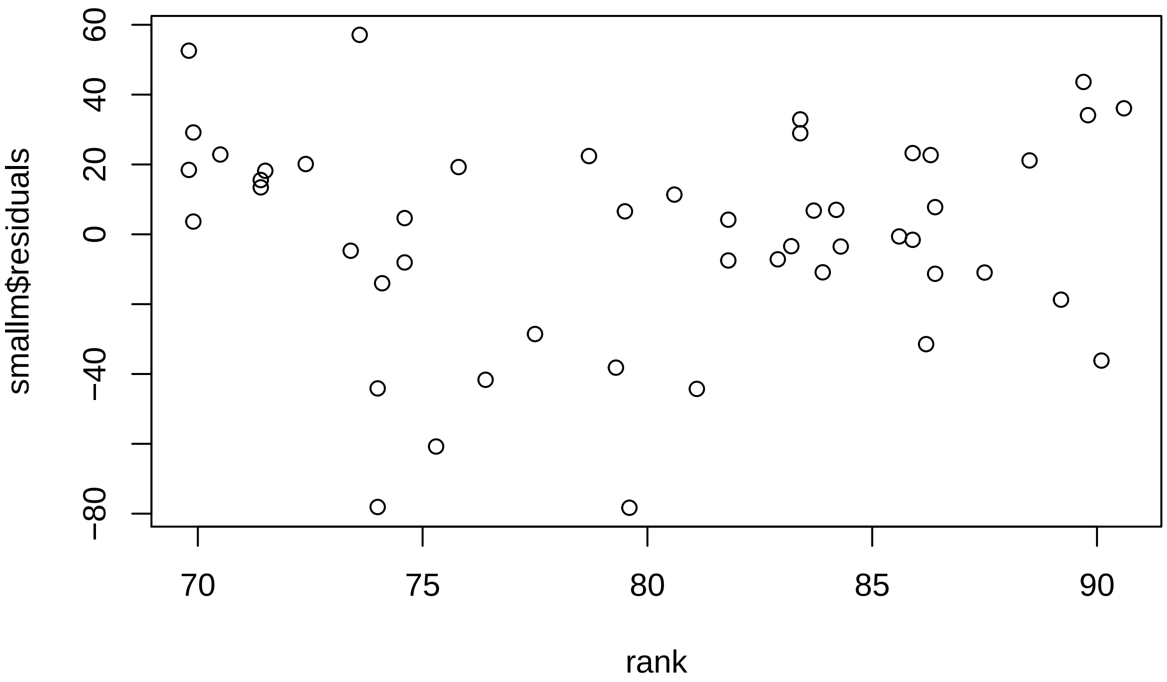
```
plot(smallm$fitted.values, smallm$residuals)
```



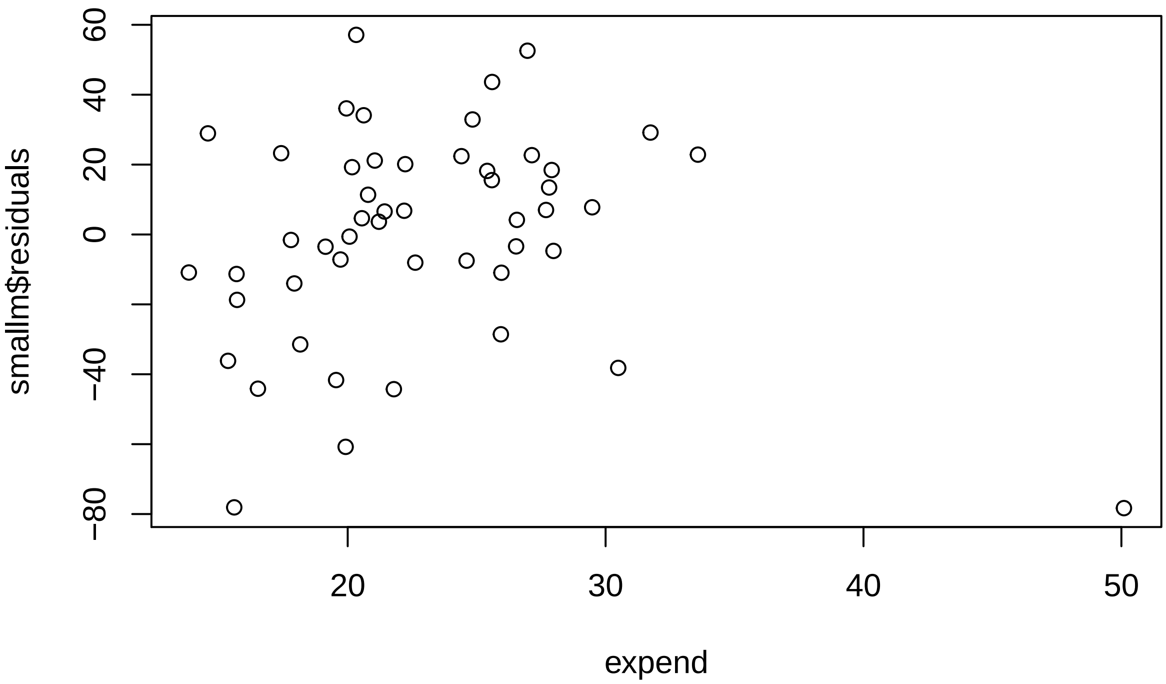
```
plot(takers, smallm$residuals)
```



```
plot(rank,smallm$residuals)
```



```
plot(expend,smallm$residuals)
```



The plot against fitted values has the look of a U-shape suggesting a nonlinearity in the relationship between `sat` and at least one of the predictors. A similar shape appears in the both the plot against `takers` and the plot against `rank` suggesting that either or both might have a nonlinear relationship to `sat`. We will talk about how to deal with the apparent nonlinearity in part (d). Finally, the plot of `expend` vs `residuals` show that `Alaska` might be an outlier, futher analysis would be need to address this point.

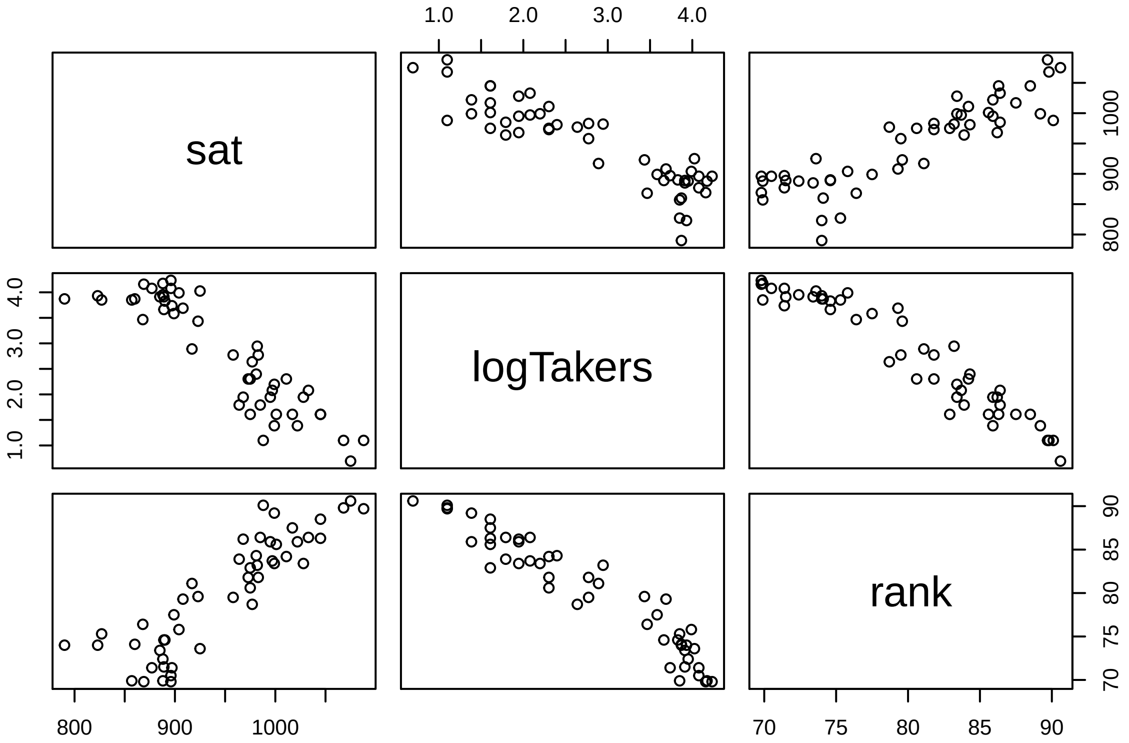
d. Was there a Better Reduced Model? (Part 1)

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 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.)

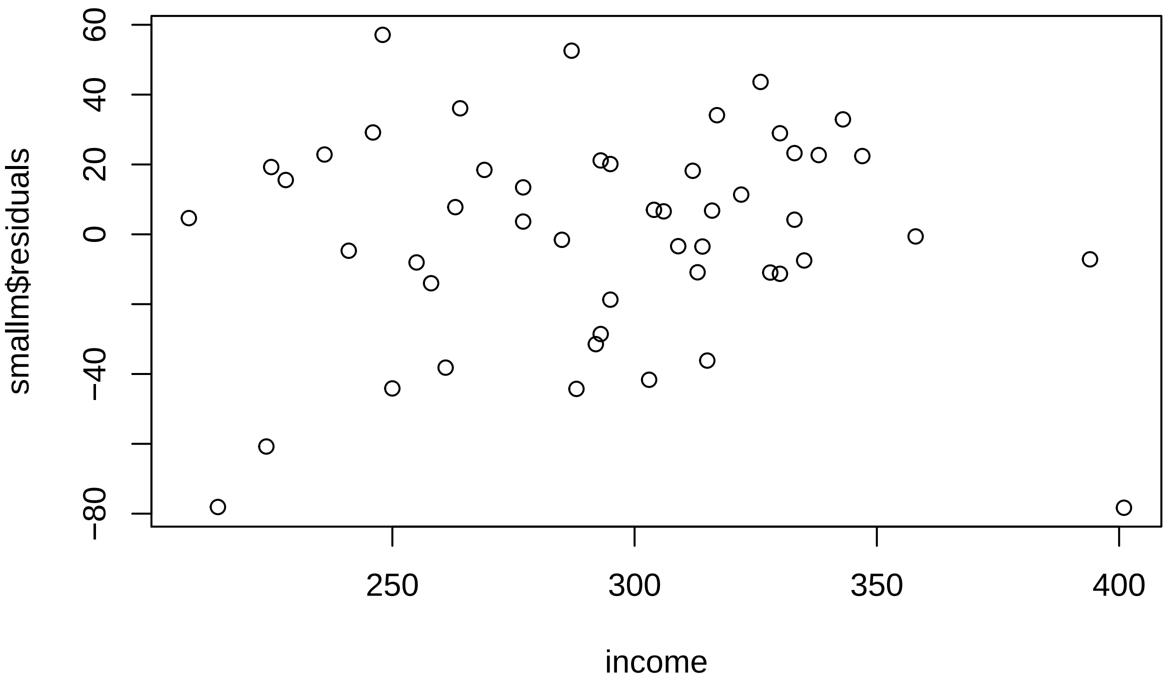
Solution

With regard to the U shapes in the plots in part (c), we begin by transforming `takers` which had a curved plot against `sat` in the original exploratory data analysis.

```
logTakers=log(takers)
pairs(cbind(sat,logTakers,rank))
```



```
plot(income, smallm$residuals)
```



There is a trend in the plot of residuals against `income`.

We see that log(`takers`) has a slightly straighter relationship with `sat` than does `takers`, and is very highly correlated with `rank`. Indeed, if you had fit the linear regression of `sat` on log(`takers`) alone, you would see that it predicts about as well as the regression on both log(`takers`) and `rank`.

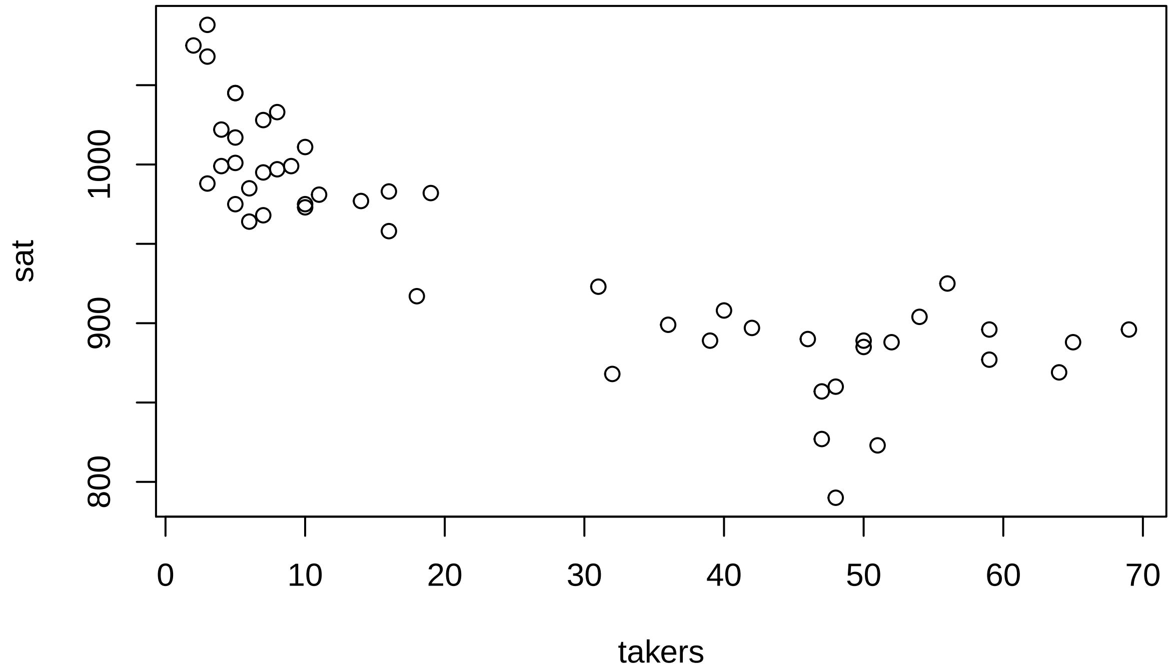
e. Was there a Better Reduced Model? (Part 2)

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.)

Solution

We plot `sat` against `takers` so that we can examine the relationship closely.

```
plot(takers,sat)
```



The gap between 20 and 30 on the `takers` scale brings our attention to the two groups: low-takers (`takers`$<$ 20) and high-takers (`takers`$>$ 30). In the low-takers group there is a definite decrease in `sat` as `takers` increases. In the high-takers group, there is virtually *no* relationship between `sat` and `takers`. The correlation in the high-takers group is almost 0. If we wanted to fit a model to predict `sat` from `takers`, we would need to introduce a categorical variables to distinguish the two groups (such as `hightakers`=`takers`$>$ 30 and `lowtakers`=`takers`$<$ 20) so that we could fit separate regressions to the two sets of states. Presumably the slope would be 0 in the `hightakers` set and negative in the other set. This is confirmed after fitting another regression model:

```
satdata$lowtakers <- I(satdata$takers <= 20)
fit2 <- lm(sat ~ expend + takers*lowtakers + rank, data = satdata)
summary(fit2)
```

```
##
## Call:
## lm(formula = sat ~ expend + takers * lowtakers + rank, data = satdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.191 -17.019   0.997  16.102  53.199
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           470.0403   155.1474   3.030  0.00409 **
## expend                  3.0407     0.6094   4.989 9.99e-06 ***
## takers                  0.9724     0.6571   1.480  0.14607
## lowtakersTRUE         169.9182    29.8626   5.690 9.62e-07 ***
## rank                    3.8790     1.8201   2.131  0.03869 *
## takers:lowtakersTRUE  -5.0557     1.2979  -3.895  0.00033 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.3 on 44 degrees of freedom
## Multiple R-squared:  0.8943, Adjusted R-squared:  0.8823
## F-statistic: 74.49 on 5 and 44 DF,  p-value: < 2.2e-16
```

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.

Solution

We can think of the null hypothesis as the coefficient for `years` being 0. To assess its plausibility, we could compare the variance explained by the models with and without `years` and see whether including `years` leads to a significant increase in the variance "explained". More on this in the next weeks!

```
fitall <- lm(sat ~ . -state, data = satdata)
summary(fitall)
```

```
##
## Call:
## lm(formula = sat ~ . - state, data = satdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.630 -13.138   2.792  14.351  44.155
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -89.945149 200.903674  -0.448 0.656668
## takers          0.679258   0.819402   0.829 0.411809
## income         -0.001946   0.144735  -0.013 0.989335
## years          17.586434   6.358229   2.766 0.008400 **
## public         -0.459184   0.550043  -0.835 0.408546
## expend          2.397267   0.807275   2.970 0.004912 **
## rank            8.598127   2.002671   4.293 0.000102 ***
## lowtakersTRUE  53.698338  22.562135   2.380 0.021928 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.02 on 42 degrees of freedom
## Multiple R-squared:  0.8931, Adjusted R-squared:  0.8753
## F-statistic: 50.13 on 7 and 42 DF,  p-value: < 2.2e-16
```

Association would be causation in this case if

- After conditioning on the observed covariates, the number of years is as good as randomized with respect to the potential outcomes for the SAT scores. That is, we can conceptualize as many potential outcomes as number of years: $C(X = x)$ is the SAT score had `years` been set to x. This is the reasoning we discussed in class and in AOS.

- The linear model is correct. If the assumptions are not met, then it is not clear if the $\hat{\beta}$ is actually capturing some treatment effect or if it's just garbage.

Recall that a confouder is a variable that affects both the treatment and the outcome. Another confounder not included in this study could be the students' family income: if wealthier students tend to spend more years in science, social and humanities and are also able to get extra private lessons on SAT outside school, which does not get measured in the study, then this could artifically inflate the treatment effect. Therefore, in this scenario, a policy aiming to simply increase the number of years of science, social and humanities would not necessarily lead to better SAT scores.