Homework 4

Advanced Methods for Data Analysis (36-402)

Due Friday February 18, 2022 *at 3:00 pm*

Solutions – not to be posted online or shared, even after the end of the semester.

You should **always show all your work** and submit both a writeup and *R* code.

- Assignments must be submitted through Gradescope as a PDF. Follow the instructions here: `https://www.cmu.edu/teaching/gradescope/`
- Gradescope will ask you to mark which parts of your submission correspond to each homework problem. This is mandatory; if you do not, grading will be slowed down, and your assignment will be penalized.
- Make sure your work is legible in Gradescope. You may not receive credit for work the TAs cannot read. **Note:** If you submit a PDF with pages much larger than 8.5 × 11", they will be blurry and unreadable in Gradescope.
- For questions involving R code, we strongly recommend using R Markdown. The relevant code should be included with each question, rather than in an appendix. A template Rmd file is provided on Canvas.

1. **The Omitted Variables Effect, Part 2.** First, review question 3 on Homework 3 so you recall the setup to this problem. We'll continue using the same data here, so ensure you have the `UCBAdmissions` data loaded.

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

	Male	Female
Admitted	1198	557
Rejected	1493	1278

(a) 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 "causal" regression function $\theta(x)$). Did it make a difference to adjust for Department?

Solution: First, we set up the data we used in the last homework assignment.

```
library(graphics)
props <- function(x) { x / sum(x) }

dept.admits <- data.frame(dept = LETTERS[1:6],
                          male_admit = numeric(6),
                          female_admit = numeric(6))
for (j in 1:6) {
  dept.admits[j, 2:3] <- apply(UCBAdmissions[,,j], 2, props)[1,]
}
```

If "Department" is the *only* confounding variable, then $C(x) \perp X | Z$. That is, within each department, admissions outcome $C(x)$ is independent of sex X.

The adjusted treatment effect is defined to be $\theta(x) = \mathbb{E}[r(x, Z)]$, where the expectation is over the *marginal* distribution of Z. That means we need an estimate of the marginal distribution of the confounding variable, "Department." We will use the empirical distribution, namely the proportions of the total sample that come from each department. Call this empirical distribution $\hat{f}_Z(\cdot)$. The formula for the adjusted effect is then

$$\hat{\theta}(x) = \sum_{z \in \{A,B,C,D,E,F\}} \hat{r}(x, z) \hat{f}_Z(z).$$

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 $\hat{r}(\cdot, \cdot)$ matrix times the $\hat{f}_Z(\cdot)$ vector:

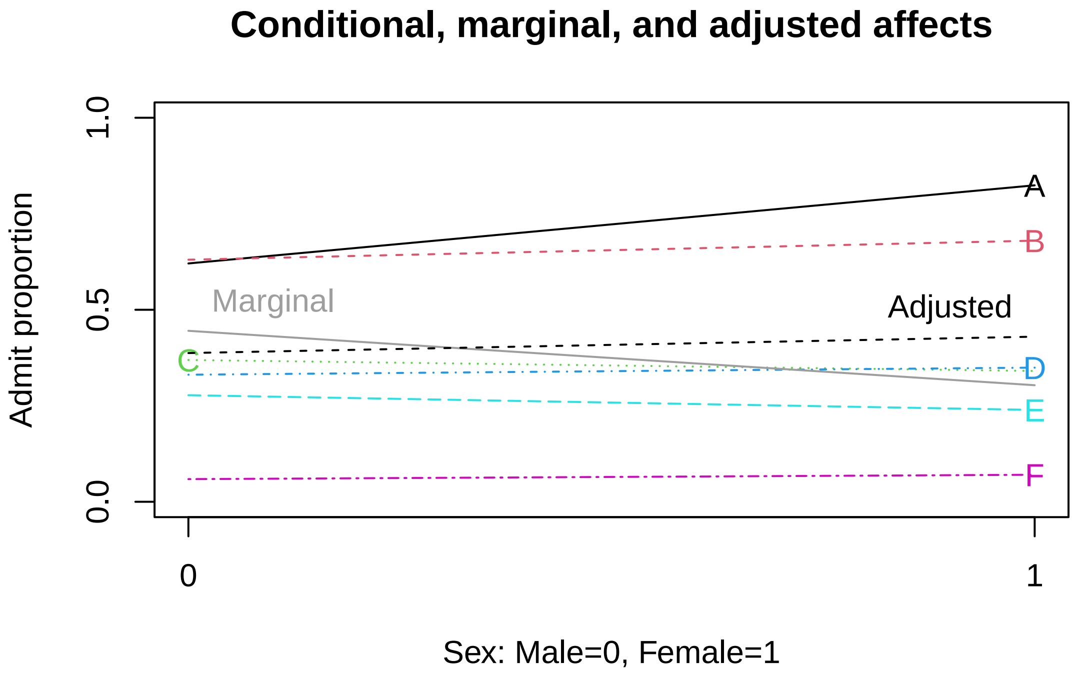
```
zdist <- apply(UCBAdmissions, 3, sum)
zdist <- zdist / sum(zdist)

# Adjusted effect theta-hat(x):

adjeff <- matrix(zdist, nrow=1, ncol=6) %*% as.matrix(dept.admits[, 2:3])
adjeff
##      male_admit female_admit
## [1,]  0.3873186    0.4299554
```

These are in the opposite order of the unadjusted effects computed in Homework 3 question 3 (a), so adjusting makes a difference.

(b) Draw a plot with all six conditional regression lines computed in Homework 3 question 3 (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 $\mathbb{P}(Y = 1 \mid X = x)$ (computed from Table 1) to the plot. Finally, add the estimated adjusted treatment effect line computed in part (a) of this homework.

Solution:

```r
mt <- margin.table(UCBAdmissions, c(1,2))
mta <- apply(mt, 2, props)[1,]

plot(c(0,1), c(0,1), pch="", lab=c(1, 3, 7),
     xlab="Sex: Male=0, Female=1", ylab="Admit proportion",
     main="Conditional, marginal, and adjusted affects")

wt <- c(1, 1, 0, 1, 1, 1) # where to place the labels

for (j in 1:6) {
  lines(c(0, 1), dept.admits[j, 2:3], lty=j, col=j)
  text(wt[j], dept.admits[j, 2 + wt[j]], LETTERS[j], col=j)
}

# Add the marginal association line
lines(c(0, 1), mta, lty=7, col=8)
text(.1, mta[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)
```



(c) In part (a) 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 *All of Statistics*, the unadjusted effect, which you estimated Homework 3, is the regression of Y on X alone, namely $r(x) = \mathbb{E}[r(x, Z) \mid X = x]$, 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 is different from the estimate of the regression of Y on X.

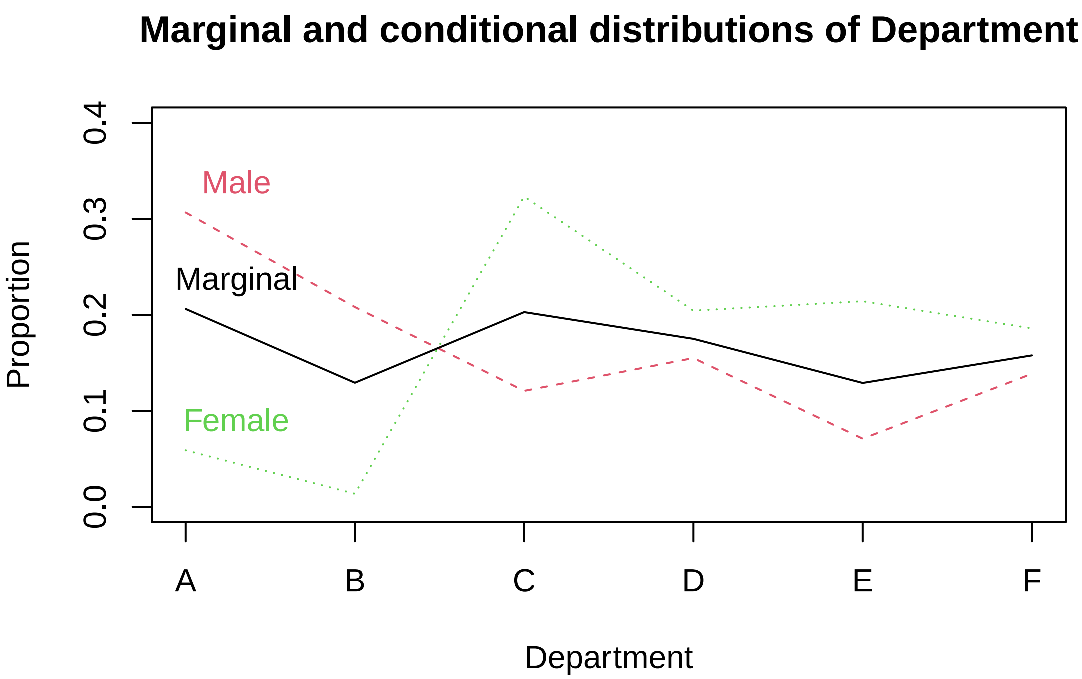
Solution: Finally, we compare the conditional distributions of Z given $X = 0$ and $X = 1$ and the marginal distribution of Z. Remark 16.7 in *All of Statistics* gives a formula for the marginal association of X and Y (the unadjusted effect) in terms of the conditional distribution of Z given X: $r(x) = \mathbb{E}[r(x, Z)|X = x]$. 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.

First let's get conditional distribution estimates:

```r
zdist0 <- apply(UCBAdmissions[,1,], 2, sum)
zdist0 <- zdist0 / sum(zdist0)
zdist1 <- apply(UCBAdmissions[,2,], 2, sum)
zdist1 <- zdist1 / sum(zdist1)
```

Now let's plot the distribution estimates:

```r
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)
```



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. Women 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).

2. **SAT Data Revisited.** 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.”

During lecture, we introduced this data set. On Canvas, we’ve provided the code for Demo 2.3 part (a), Exploratory Data Analysis, as *“Demo2_3_starter_SAT.R”*.

As always, include a clear and complete write-up with analysis, discussion, figures and *R* code.

Before starting this problem, go through the demo code for parts (a) and (b) by yourself and make sure you understand the main ideas and concepts.

(a) **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 and median class rank. Let’s explore this thinking further.

To address the research question of how the states rank after accounting for the percentage of takers and median class rank, we define a reduced model that fits the regression line of `sat` on `takers` and `rank`. 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` and `rank`? 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. The `rank` function does this, and we specify how to break the ties:

```r
satdata <- read.table("CASE1201.ASC", header = TRUE)

satdata$satrank <- rank(-satdata$sat, ties.method="min")
head(satdata)
```

```
##          state  sat takers income years public expend rank satrank
## 1         Iowa 1088      3    326 16.79   87.8  25.60 89.7       1
## 2 SouthDakota 1075      2    264 16.07   86.2  19.95 90.6       2
## 3 NorthDakota 1068      3    317 16.57   88.3  20.62 89.8       3
## 4      Kansas 1045      5    338 16.30   83.9  27.14 86.3       4
## 5    Nebraska 1045      5    293 17.25   83.6  21.05 88.5       4
## 6     Montana 1033      8    263 15.91   93.7  29.48 86.4       6
```

It looks like the states were already ranked by SAT scores. (This is true if you view the entire data frame as well.)

Next, we will run the requested model and rank the states by their residuals.

```
smallm <- lm(sat ~ takers + rank, data = satdata)
summary(smallm)
##
## Call:
## lm(formula = sat ~ takers + rank, data = satdata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -98.48 -22.31   5.46  21.40  53.89
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 412.8554   194.2227   2.126   0.0388 *
## takers       -0.8170     0.6584  -1.241   0.2208
## rank          6.9574     2.2229   3.130   0.0030 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.83 on 47 degrees of freedom
## Multiple R-squared:  0.7814,Adjusted R-squared:  0.7721
## F-statistic:    84 on 2 and 47 DF,  p-value: 3.032e-16
```

To rank by residuals, we run

```
satdata$newrank <- rank(-residuals(smallm), ties.method="min")
ordering <- order(residuals(smallm), decreasing = TRUE)
head(satdata[ordering, c("state", "sat", "satrank", "newrank")])
##            state  sat satrank newrank
## 35   Connecticut  896      35       1
## 1           Iowa 1088       1       2
## 28  NewHampshire  925      28       3
## 41 Massachusetts  888      40       4
## 36       NewYork  896      35       5
## 7      Minnesota 1028       7       6
```

We can also find the states whose ranks changed the most:

```
changerank <- satdata$satrank - satdata$newrank

satdata[which.max(changerank), ]
##            state sat takers income years public expend rank satrank newrank
## 41 Massachusetts 888     65    246 16.79   80.7  31.74 69.9      40       4
satdata[which.min(changerank), ]
##       state sat takers income years public expend rank satrank newrank
## 12 Arkansas 999      4    295 15.49   86.4  15.71 89.2      12      43
```

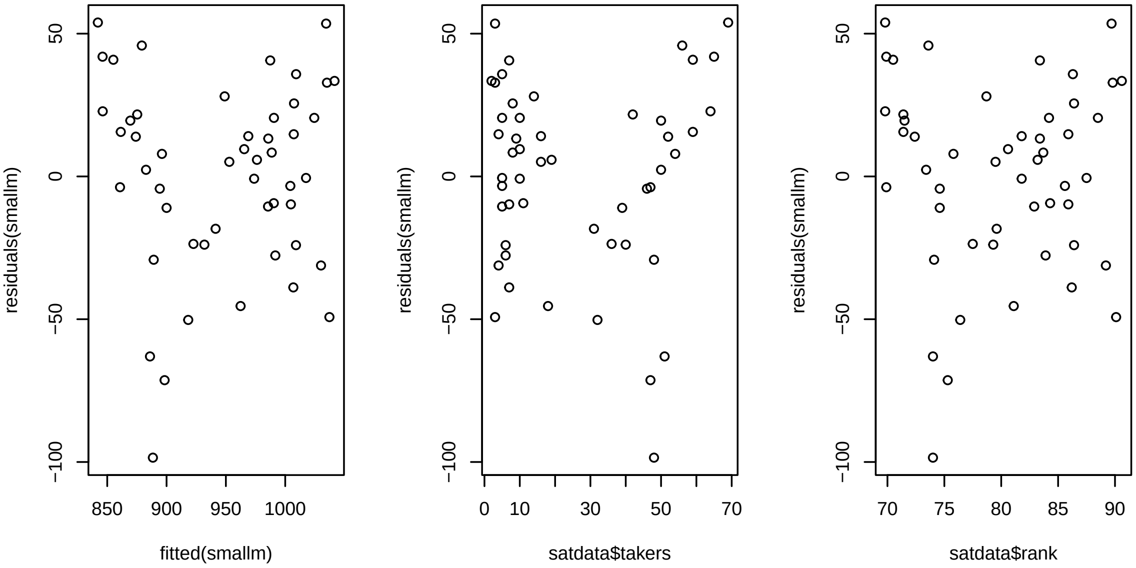
We see that Massachusetts jumped 37 spots while Arkansas fell 31 spots. Looking at Massachusetts first, we see that the value of `takers` = 65, with a negative coefficient in the model, is almost the maximum, and the value of `rank` = 69.9, with a positive coefficient, is very low. Both of these lead to Massachusetts having 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 a very high `rank` = 89.2 and a very low `takers` = 4, which gives it nearly the largest fitted value, while its actual `sat` = 999 is around the 80th percentile giving it a low residual.

(b) **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` and `rank`. Do you see any patterns in the residual plots?

Solution: We begin with the requested residual plots.

```
par(mfrow=c(1,3))
plot(fitted(smallm), residuals(smallm))
plot(satdata$takers, residuals(smallm))
plot(satdata$rank, residuals(smallm))
```

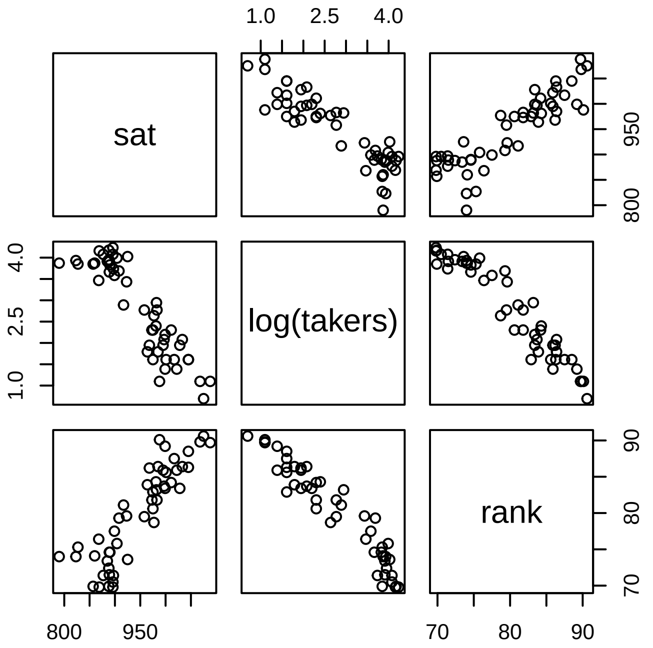


The plot against fitted values has the look of a U-shape suggesting a nonlinearity in the relationship between `sat` and at least one of the predictors. A similar shape appears in the both the plot against `takers` and the plot against `rank` suggesting that either or both might have a nonlinear relationship to `sat`. We will talk about how to deal with the apparent nonlinearity in part (f).

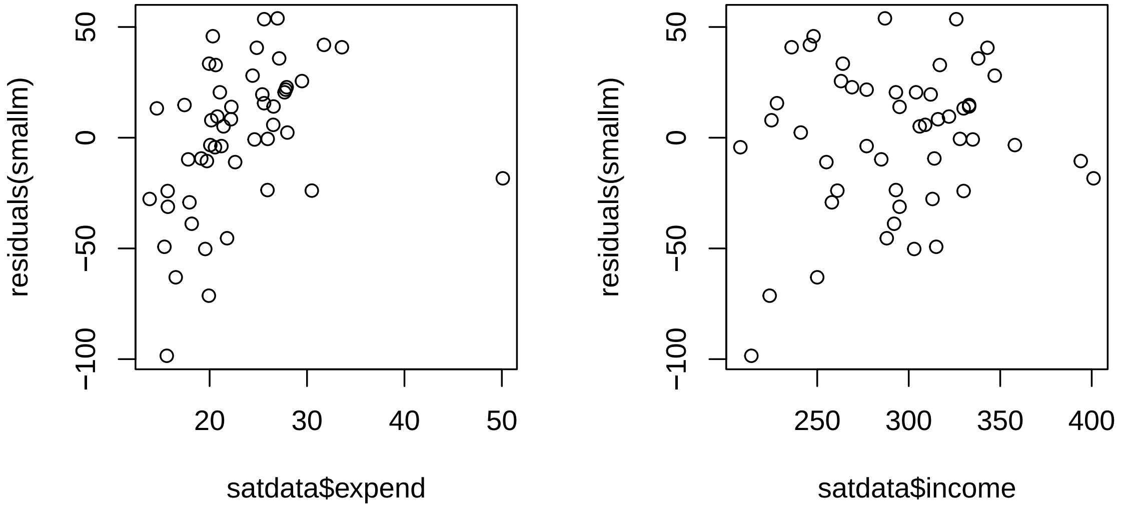
(c) **Was There a Better Reduced Model? (Part 1)** Does it appear that any transformations of variables would improve the relationships between the variables in the reduced model? How could you use the model you have developed to explore whether state expenditure on education (`expend`) is related to SAT scores, after controlling for the confounders? 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 (d), we begin by transforming `takers` which had a curved plot against `sat` in the original exploratory data analysis.

```
pairs(~ sat + log(takers) + rank, data=satdata)
```



```
par(mfrow=c(1,2))
plot(satdata$expend, residuals(smallm))
plot(satdata$income, residuals(smallm))
```



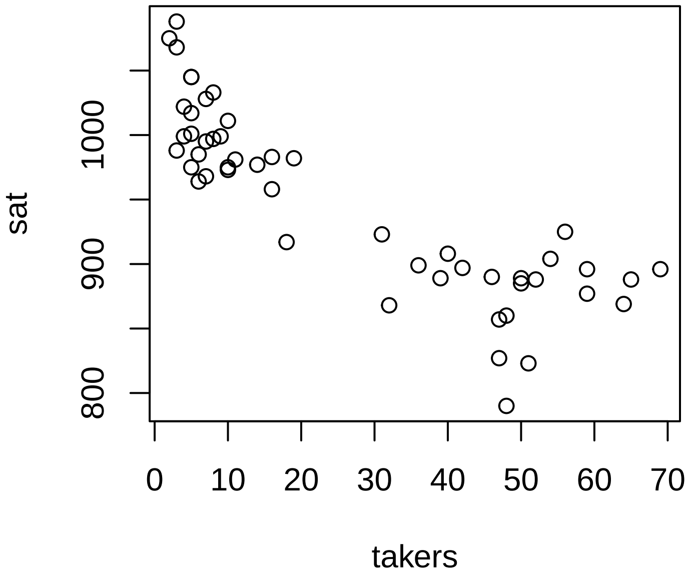
Except for the one state with the largest expenditures, the residuals seem to increase with `expend`. This suggests that expenditures probably help to predict SAT scores. There is a slightly weaker trend in the plot against `income`. Indeed, if you had fit models that included `expend` and/or `income`, you would see that `expend` adds more predictive power than does `income`, and `expend` alone is about as good as the two together. We would build simply add `expend` to the model we developed in the previous parts to see its relationship with SAT scores.

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`.

(d) **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(sat ~ takers, data=satdata)
```



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 (which you did not have to do):

```
satdata$lowtakers <- I(satdata$takers <= 20)
fit2 <- lm(sat ~ takers*lowtakers + rank, data = satdata)
summary(fit2)
##
## Call:
## lm(formula = sat ~ takers * lowtakers + rank, data = satdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -90.226 -10.981   6.687  18.281  50.327
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          468.2303   191.9676   2.439 0.018731 *
## takers                 0.9792     0.8131   1.204 0.234776
## lowtakersTRUE        134.6081    35.8972   3.750 0.000503 ***
## rank                   4.9324     2.2368   2.205 0.032599 *
## takers:lowtakersTRUE  -3.5124     1.5596  -2.252 0.029243 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.07 on 45 degrees of freedom
## Multiple R-squared:  0.8346,Adjusted R-squared:  0.8199
## F-statistic: 56.76 on 4 and 45 DF,  p-value: < 2.2e-16
```

3. **More Conditional Expectation Practice.** Suppose X, Y, and Z have the following distribution:

$$Z \sim \text{Bernoulli}\left(\frac{1}{2}\right)$$

$$X \mid Z = z \sim \text{Bernoulli}\left(\frac{e^{4z-2}}{1+e^{4z-2}}\right)$$

$$Y \mid Z = z, X = x \sim \text{Bernoulli}\left(\frac{e^{2(z+x)-2}}{1+e^{2(z+x)-2}}\right)$$

This could, for example, model a binary outcome variable Y given a binary treatment X and confounding variable Z.

(a) Find an expression for $\mathbb{P}(Y = y \mid X = x)$, in terms of $\mathbb{P}(Z)$, $\mathbb{P}(X \mid Z = z)$, and $\mathbb{P}(Y \mid Z = z, X = z)$.

Use the expression to calculate $\mathbb{P}(Y = 1 \mid X = 1)$. You may do this by hand or by using R to do the arithmetic.

Solution: By the law of total probability,

$$\mathbb{P}(Y = y \mid X = x) = \sum_{z=0}^{1} \mathbb{P}(Y = y \mid X = x, Z = z)\mathbb{P}(Z = z \mid X = x).$$

We can easily calculate $\mathbb{P}(Y = y \mid X = x, Z = z)$, since we are told this is Bernoulli. To get $\mathbb{P}(Z = z \mid X = x)$, we must use Bayes' rule, since we only know $\mathbb{P}(X = x \mid Z = z)$. According to Bayes' rule,

$$\begin{aligned}
\mathbb{P}(Z = z \mid X = x) &= \frac{\mathbb{P}(X = x \mid Z = z)\mathbb{P}(Z = z)}{\mathbb{P}(X = x)} \\
&= \frac{\mathbb{P}(X = x \mid Z = z)\mathbb{P}(Z = z)}{\mathbb{P}(X = x \mid Z = 0)\mathbb{P}(Z = 0) + \mathbb{P}(X = x \mid Z = 1)\mathbb{P}(Z = 1)}.
\end{aligned}$$

(We've used the law of total probability again in the denominator.) By plugging this into the previous expression, we have our desired result.

To calculate the number, we can use R. Here we'll write simple functions that calculate $\mathbb{P}(Z = z \mid X = x)$ and $\mathbb{P}(Y = y \mid X = x)$. Note that `dbinom` is the binomial probability mass function, and with `size=1`, the binomial is the same as the Bernoulli.

```r
# P(X | Z = z)
px_given_z <- function(x, z) {
  return(dbinom(x, 1, prob = exp(4 * z - 2) / (1 + exp(4 * z - 2))))
}

# P(Z | X = x)
pz_given_x <- function(z, x) {
  numerator <- px_given_z(x, z) * 1/2

  denominator <- px_given_z(x, 0) * 1/2 + px_given_z(x, 1) * 1/2

  return(numerator / denominator)
}

# P(Y | X = x)
py_given_x <- function(y, x) {
  result <- 0.0
  for (z in 0:1) {
    term <- dbinom(y, 1,
                   prob = exp(2 * (x+ z) - 2) / (1 + exp(2 * (x + z) - 2))) *
      pz_given_x(z, x)

    result <- result + term
  }

  return(result)
}

py_given_x(1, 1)
## [1] 0.835405
```

(b) Write an R function to simulate a sample $(X_1, Y_1, Z_1), \dots, (X_n, Y_n, Z_n)$ from this model. Your code should first draw Z, use the simulated z to simulate a value x from $X \mid Z = z$, and finally simulate a value y from $Y \mid X = x, Z = z$, each using the appropriate distribution. Write the code to do this simulation 10,000 times and place the results in a data frame. This will give you a sample from the joint distribution of X, Y, Z of size $n = 10,000$.

Using your data frame, estimate $\mathbb{P}(Y = 1 \mid X = 1)$ empirically. Compare this to your result from (a).

Solution: Our simulation code is:

```
n <- 10000

z <- rbinom(n, 1, 1/2)
x <- rbinom(n, 1, exp(4 * z - 2) / (1 + exp(4 * z - 2)))
y <- rbinom(n, 1, exp(2 * (z + x) - 2) / (1 + exp(2 * (z + x) - 2)))

cases <- x == 1
sum(y[cases] == 1) / length(y[cases])
## [1] 0.8418642
```

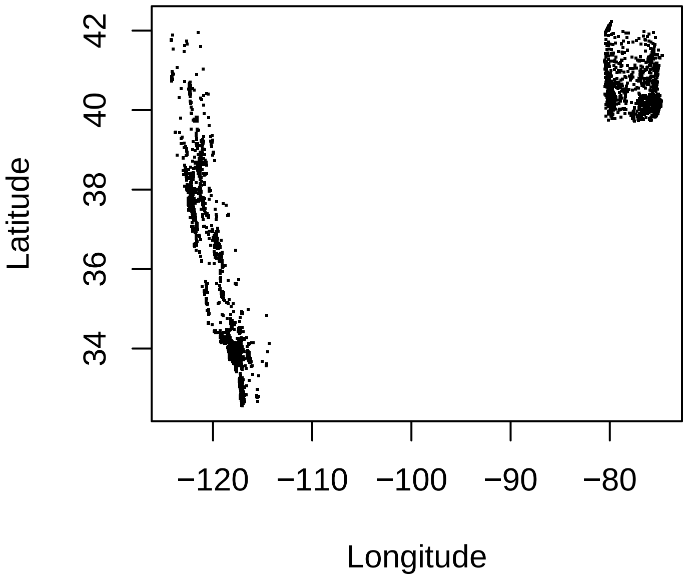
4. **Housing Data Revisited.** Return to the housing data with which you worked in Homework 2. In this problem, we will take into account the locations of the census tracts using some combination of the two location variables `Latitude` and `Longitude`.

   For parts (a) through (d) below, use *only* the training data set. Also, draw all scatter plots using the graphical parameter `pch="."` in order to cut down on the overlap of points.

   (a) Draw the scatter plot of `Latitude` and `Longitude`. From the scatter plot, identify at least two places, locations in (`Latitude`, `Longitude`) space, where census tracts cluster.

Solution:

```
housetrain <- read.csv("housetrain.csv", header = TRUE)
housetest <- read.csv("housetest.csv", header = TRUE)
plot(Latitude ~ Longitude, pch=".", data=housetrain)
```

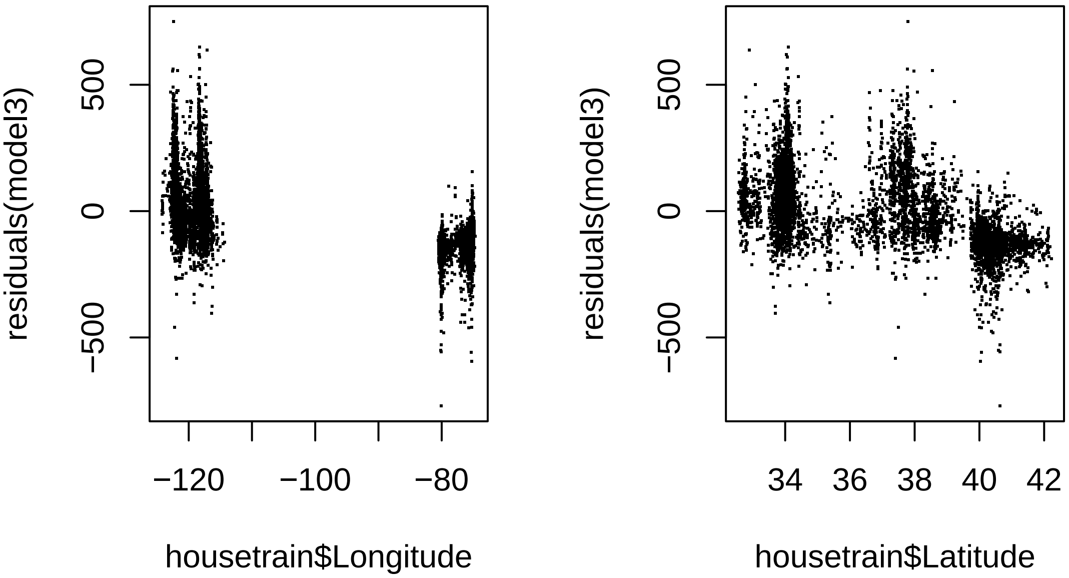


The census tracts are in Pennsylvania (those with `Longitude` > 80) and in California (those with `Longitude` < −115.) There are two clusters in Pennsylvania that correspond to Pittsburgh (`Longitude` close to −80 and `Latitude` between 40 and 41) and Philadelphia (highest `Longitude` > −75.5 and lowest `Latitude`.) There are also a number of clusters in California corresponding to the San Francisco, Los Angeles, and San Diego areas.

(b) Whether you did this or not in Homework #2, plot the residuals from Model 3 (the one with both `Mean_household_income` and `Median_household_income` as predictors) against each of the location variables. Are there patterns that suggest that location might be useful in predicting house values? Explain.

Solution: We begin by refitting Model 3 and then plotting the residuals against the location variables.

```
model3 <- lm(Median_house_value ~ Mean_household_income +
             Median_household_income, data = housetrain)

par(mfrow=c(1, 2))
plot(housetrain$Longitude, residuals(model3), pch=".")
plot(housetrain$Latitude, residuals(model3), pch=".")
```



The `Longitude` plot shows two striking features. First, house prices in California are higher than in Pennsylvanis (OK, not so striking) and second that prices seem more spread out in California than in Pennsylvania. These same features show up in the `Latitude` plot, where the left and central clusters are Southern and Northern California, and the right cluster is mostly Pennsylvania.

(c) The most straightforward way to take location into account is to include the two location variables in a linear model. Define Model 4 to be the linear regression of `Median_house_value` on the four predictors consisting of the two in Model 3 and the two location variables. Fit this model and plot its residuals against the four predictors and its fitted values. Between the residual plots and the summary of the regression, explain why it appears that one of Model 3 or Model 4 is better than the other.

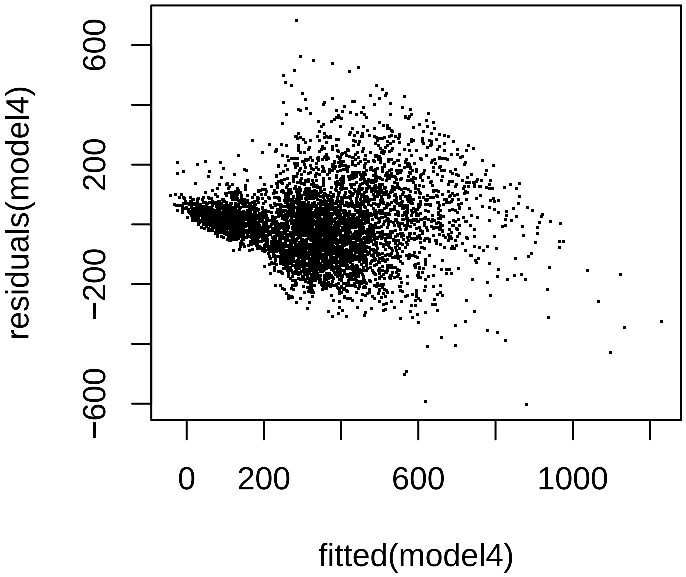
Solution:

```
model4 <- lm(Median_house_value ~ Mean_household_income +
             Median_household_income + Latitude + Longitude,
             data = housetrain)
summary(model4)
```

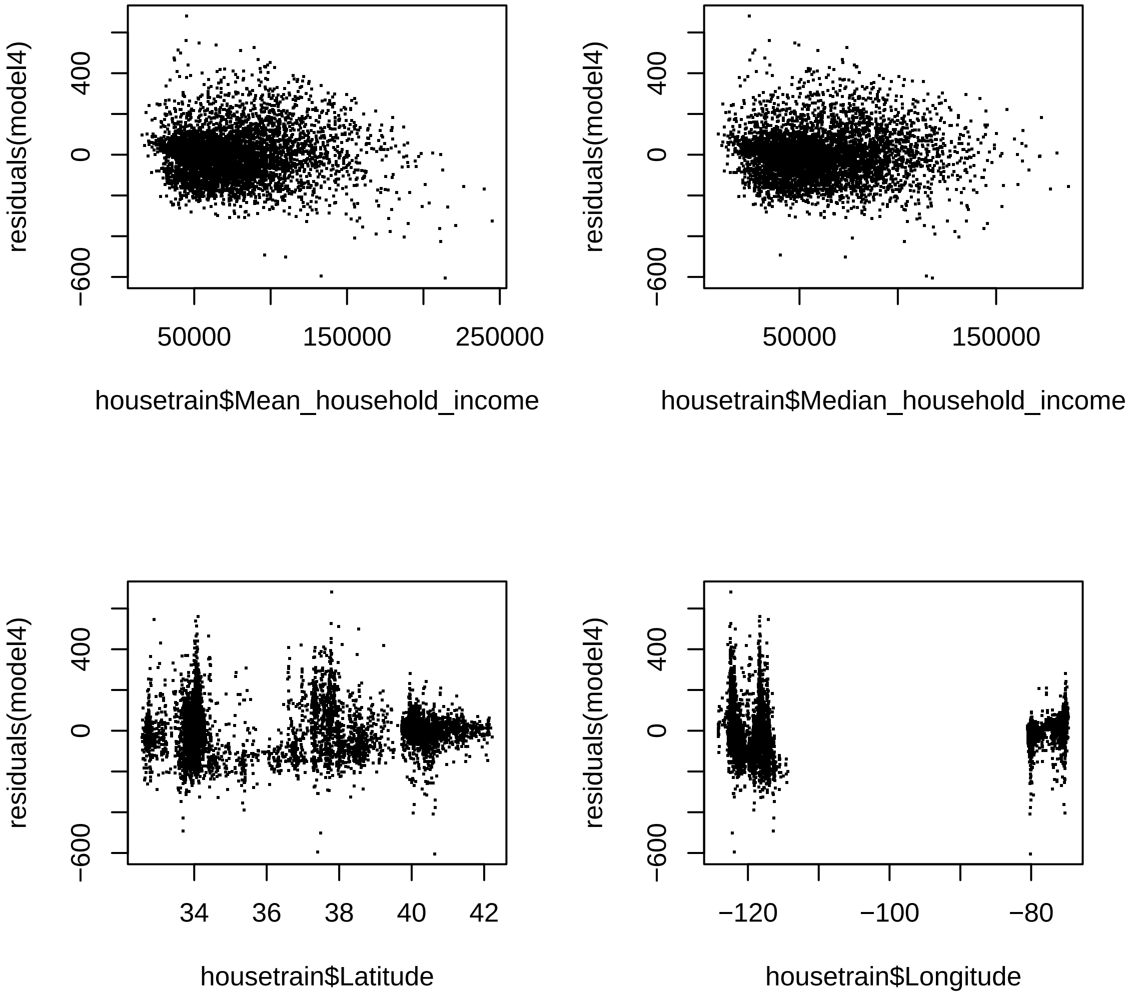
```
##
## Call:
## lm(formula = Median_house_value ~ Mean_household_income + Median_household_income +
##     Latitude + Longitude, data = housetrain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -603.90  -75.39   -2.33   59.12  681.32
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              7.945e+01  4.138e+01   1.920   0.0549 .
## Mean_household_income    5.558e-03  1.659e-04  33.498   <2e-16 ***
## Median_household_income -1.761e-03  1.981e-04  -8.889   <2e-16 ***
## Latitude                -1.145e+01  8.282e-01 -13.825   <2e-16 ***
## Longitude               -3.535e+00  1.250e-01 -28.283   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 119.2 on 5298 degrees of freedom
## Multiple R-squared:  0.6798,Adjusted R-squared:  0.6796
## F-statistic:  2812 on 4 and 5298 DF,  p-value: < 2.2e-16
```

The R^2 is a lot bigger from Model 4 compared to Model 3.

```
plot(fitted(model4), residuals(model4), pch=".")
```



```
par(mfrow=c(2,2))
plot(housetrain$Mean_household_income, residuals(model4), pch=".")
plot(housetrain$Median_household_income, residuals(model4), pch=".")
plot(housetrain$Latitude, residuals(model4), pch=".")
plot(housetrain$Longitude, residuals(model4), pch=".")
```



The residuals of Model 4 in the various location clusters are centered at near 0, unlike the residuals of Model 3.

To be continued. We'll resume this problem next week with some cross-validation, so be sure to save your code and data.