Homework 3 Solutions

A. (Review of hypothesis testing). It is a well-known fact (that has been empirically verified) that a flipped coin lands heads approximately 50% of the time. Did you know, however, that a spun coin lands on its head less often than 50% of the time? The reasoning is that if you spin a coin, the extra weight on the head side pulls the coin so that it is slightly more likely to land with the tails-side up.

a) Suppose you spin a coin 100 times. How many times would it have to land Heads before you would believe my claim about spinning coins? Why?

This problem was intended to be a review of some fundamental concepts surround hypothesis testing (although I know that for some of you the concept of "power" might be new.) The idea was to strip it of its "t-test" trappings and put it in a context that has more intuitive appeal -- flipping a coin. You weren't asked to deal with this formally (although some of you did, which is good), but if you were, you might begin by stating the null and alternative hypotheses:
H0: $p = .5$
Ha: $p < .5$
where p = probability of the coin landing heads.

In other words, someone (me) is claiming that the probability of getting heads is less than 50% and you are reacting in an appropriately skeptical manner.

Now you have to decide how skeptical. If the null hypothesis is true, we expect to get about 50 heads, give or take. If the alternative is true, however, we'll get fewer than 50 heads (although if the truth is that p is below but close to 50%, we might still get more than 50 heads on one particular trial -- but on average, over repeated trials, we should get fewer than 50% heads.) If you're very skeptical, you might require a very low number of heads before you're willing to change your mind, maybe $x = 30$. If you're not so skeptical, maybe you'll admit defeat if about 40 heads come up. Some of you were too generous and said you would reject the null hypothesis if $x = 49$. As we're about to see, if you do this, you'll

live in a world in which about half of the time you think that coins are biased when they really aren't.

b) Suppose in (a) you had decided that you would conclude that the probability of getting heads when spinning a coin was NOT .5 if you spun it 100 times and observed x heads or fewer. (x should be whatever value you used to answer (a)). Let p = true probability of getting heads when spinning a coin. Assume p = .5. Use pbinom to find the probability that you will mistakenly conclude that p is less than .5. This figure is called the significance level and is famously represented by the lower-case greek letter alpha. The significance level is the probability of making the mistake of rejecting the null hypothesis (that p = .5) when in fact the null hypothesis is true.

The significance level is the probability of rejecting the null hypothesis when it's true. This is also called a Type I error, although in my opinion it could have been better named. Let's suppose our rule is to reject the null hypothesis (and conclude that in truth, spun coins have p < .5) if we get 40 or fewer heads. Then P(reject null hypothesis assuming p = .5) = P(number of heads <= 40 assuming p = .5) = pbinom(40, 100, .50) = 0.0284 or about 2.8%. This is pretty low. It means that if spun coins really land on heads with probability .5, you'll very rarely look foolish by declaring that they're actually biased and land heads less with probability less than .5.

If your cuttoff was x = 30, then pbinom(30, 100,.50) = 0.000039, which is very skeptical indeed. If your cutoff was x = 49, then pbinom(49,100,.5) = .46, which means that 46% of the time you'll believe that "fair" coins are really unfair.

c) Now assume that in truth, p = .45. What's the probability, using the value x you chose in part (a), that you will (correctly) conclude that p is not .5 if you carry out the experiment? This value is called the "power": the probability of correctly rejecting the null hypothesis when, in fact, it is false and shoudl be rejected. You want the power to be as high as possible.

The "power" is the probability of doing a good thing: the probability that you'll reject the null hypothesis when it is, in fact, false. That is Power = P(say that p < .5 when in fact p is less than .5). For our cutoff rule, Power = P(number of heads < 40 when p is less than .5). The problem with this is that this probability depends on exactly which value p really is, and so the problem tells you to assume p = .45.

Power = P(# of heads < 40 given that p = .45) =
pbinom(40,100,.45) = 0.18.

This is disappointingly low. It means that if the truth is 45%, our testing procedure will only acknowledge this truth 18% of the time.

The problem is that there's too much variability in only 100 flips of a coin. 45% is not that far from 50% in terms of the variability of our test statistic.

If your rule chose x = 30, power = .00153 (but the good news is that your significance level was really low!).

If your rule chose x = 49, power = .8172 (but the bad news is that your significance level is really high!)

d) Choose another value of x so that the power is now higher than before. (x , remember, is the number which, if you do the experiment and get x heads or fewer, you'll decide that p does not equal .5.)

We could improve the power by being a little less finicky and skeptical and let go of the null hypothesis a little more easily. So instead of x = 40, suppose x = 42. Then
pbinom(42, 100,.45) = 0.3086 .

We've almost doubled the power!

e) Using the x you chose in (d), calculate the significance level.

The significance level of our "new" test is P(reject when p = 5) = P(x <= 42 when p = .5) = pbinom(42, 100,.5) = 0.067. So although we doubled our power by moving x from 40 to 42, we've increased our significance level.

In practice, a better method for increasing power -- one that does not negatively impact the significance level -- is to increase the sample size. I'll leave it to you to figure out what your power/significance level would have been if we flipped the coin 1000 times and rejected the null hypothesis if we got 40% or fewer heads.

I

B. Load the survey data set into R (use the command read.table("survey.txt", header=T, sep="\t")). This data consists of responses to a survey by a Stats 11 class.
a) Plot weight against height (weight should be on the y-axis.) There are two outliers. To find out where they are in the data set, do the following:
i) type identify(height, weight)
ii) place your cursor over one of the points and click. Do the same for the other point.
iii) Hit the escape key.
If all went well, two numbers should appear, and these numbers are the locations of these two points in the data set.

You probably discovered that there are outliers at points 58 and 59. If these people are reporting their heights correctly, they are over 25 feet tall. We can be pretty certain this is incorrect and delete these outliers with no qualms.

b) Create two new variables, height2 and weight2 that have these two outlier points removed. (For example, if you wanted to remove the 6th point, type height2 <- height[-6]

Some had trouble deleting the outliers. It is very important that you either delete in the right order, or delete them both at once. The best thing to do is

```
height2 <- height[-c(58,59)]
weight2 <- weight[-c(58,59)]
```

Some of you instead did

height2 <- height[-58]
height2 <- height[-59]

The problem with this is the second command writes over the first command, and you end up with height2 still having the 58th observation in it.

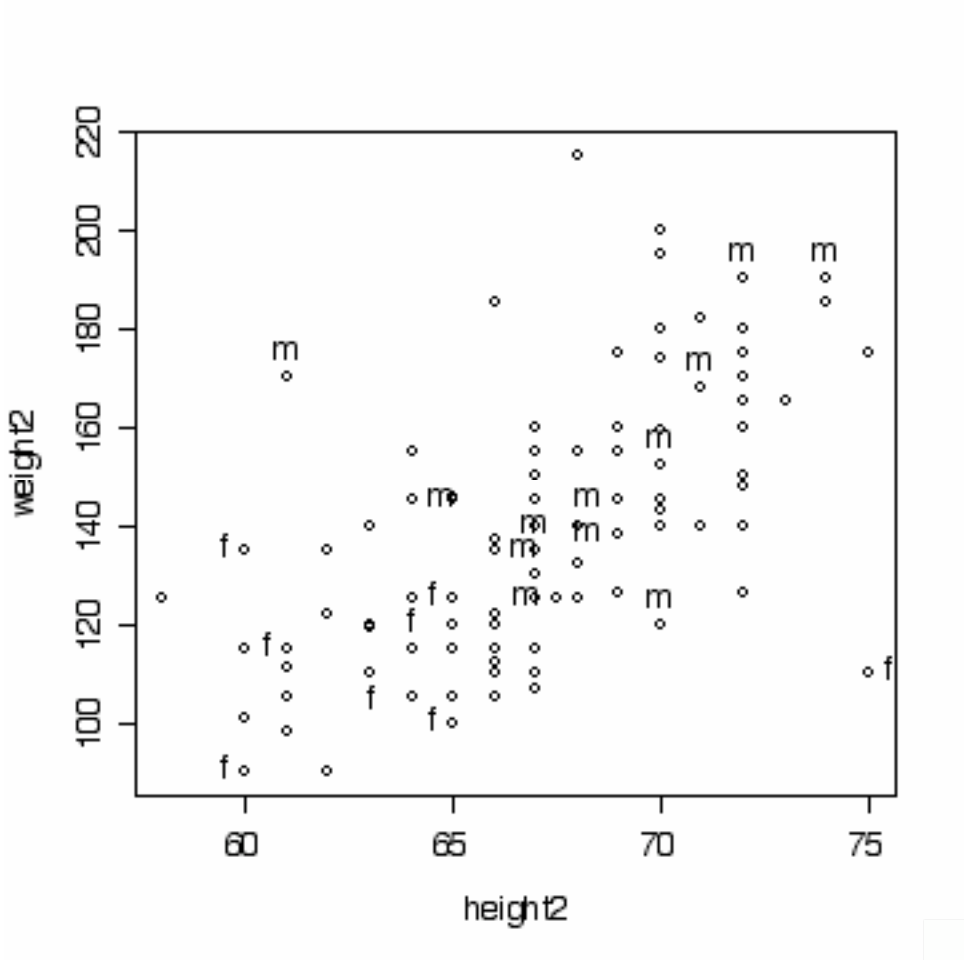
Another mistake was
height2 <- height[-58]
height2 <- height2[-59]

The first command drops the 58th observation, and so the 59th (an outlier) slides into its place and is now in the 58th slot of height2. But the second command drops the 59th observation o fheight2, even though height2 has only 58 observations. So it leaves the outlier in place.

These are common mistakes. What I want to warn you against, however, is compounding this mistake. When you make a plot of the data and notice the outliers are still there, you should do some investigating and figure out what went wrong.

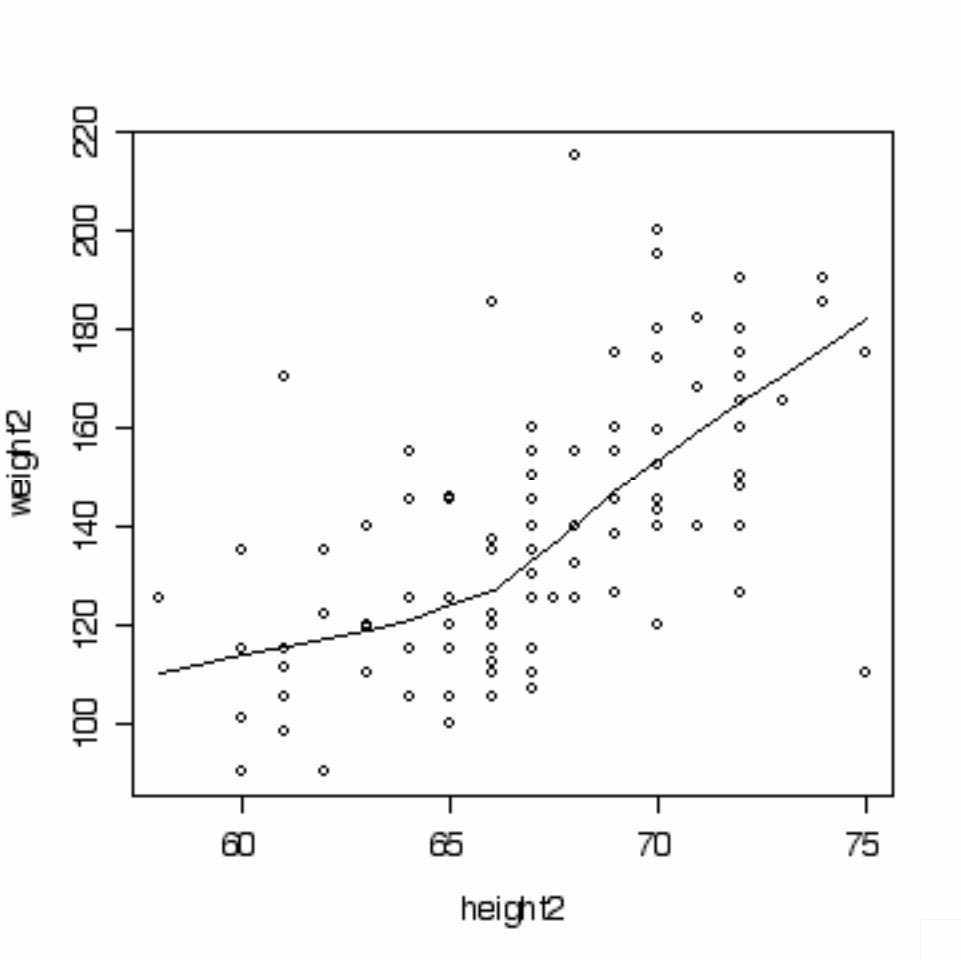
c) Plot weight2 against height2 and comment on the relationship between height and weight (which shoudl be obvious without the graph.) This plot combines men and women. Can you see any evidence of two distinct groups? (the identify command can be fun here. Type identify(height2,weight2, labels =gender) and click on a few points; then hit the escape key.

Without the 'identify' command, you'd be hard pressed to see a difference. (Some of you said you did, but didn't describe how or what). What you see is that the men tend to be in the upper-right of the plot, and the women in the lower left.



I apologize for forgetting to remind you of one fact. Because you're now working with "shortened" variables (height2 and weight2), you need to also shorten any other variables you bring into play. So you need to do
gender2 <- gender[-c(58,59)]
identify(height2, weight2, labels=gender2)

This plot appears to show a linear, positive trend between weight and height. Taller people tend to weigh more (not surprisingly). Some of you did a great thing and put a lowess curve on the plot.



This suggests a "kink" in the trend. Some of you said that while this suggested that the linear model might not hold, it could also just be an artifact of the natural variation in such a data set. True. But given that we know that the plot combines men and women, it could also be that there are actually two different trends: one trend describes the relationship between height and weight for women, the other for men.

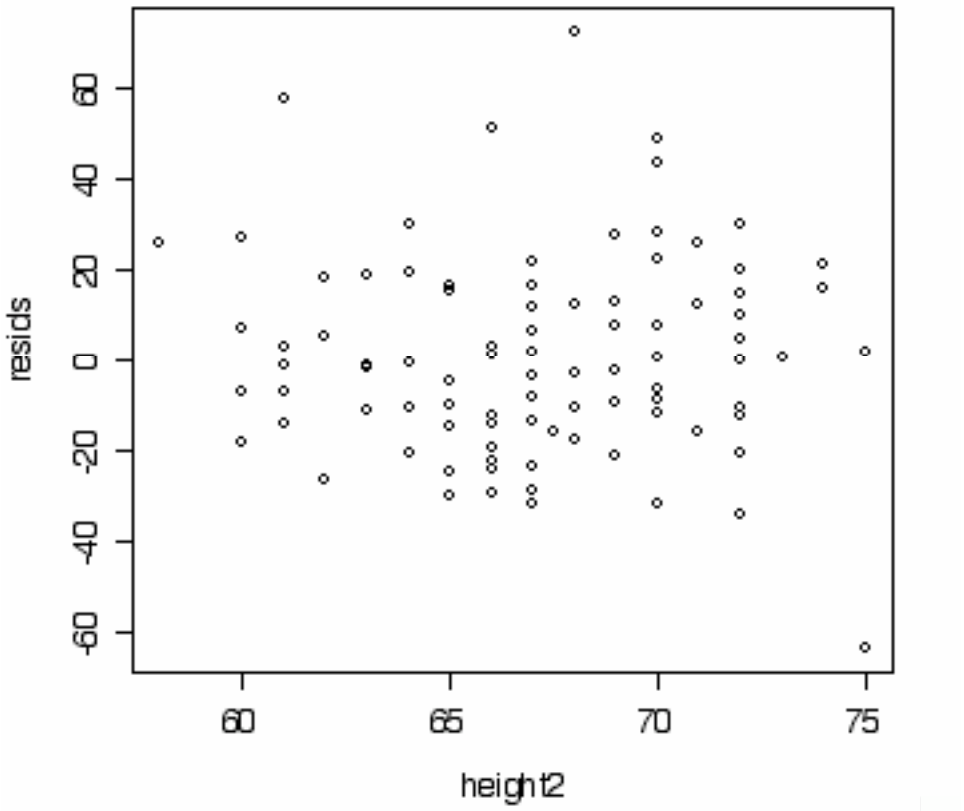
d) Find the least-squares regression line and interpret the results. (See the bottom of p. 108 and top of 109).

The regression line is Predicted Weight = -154.05 + 4.3648*Height

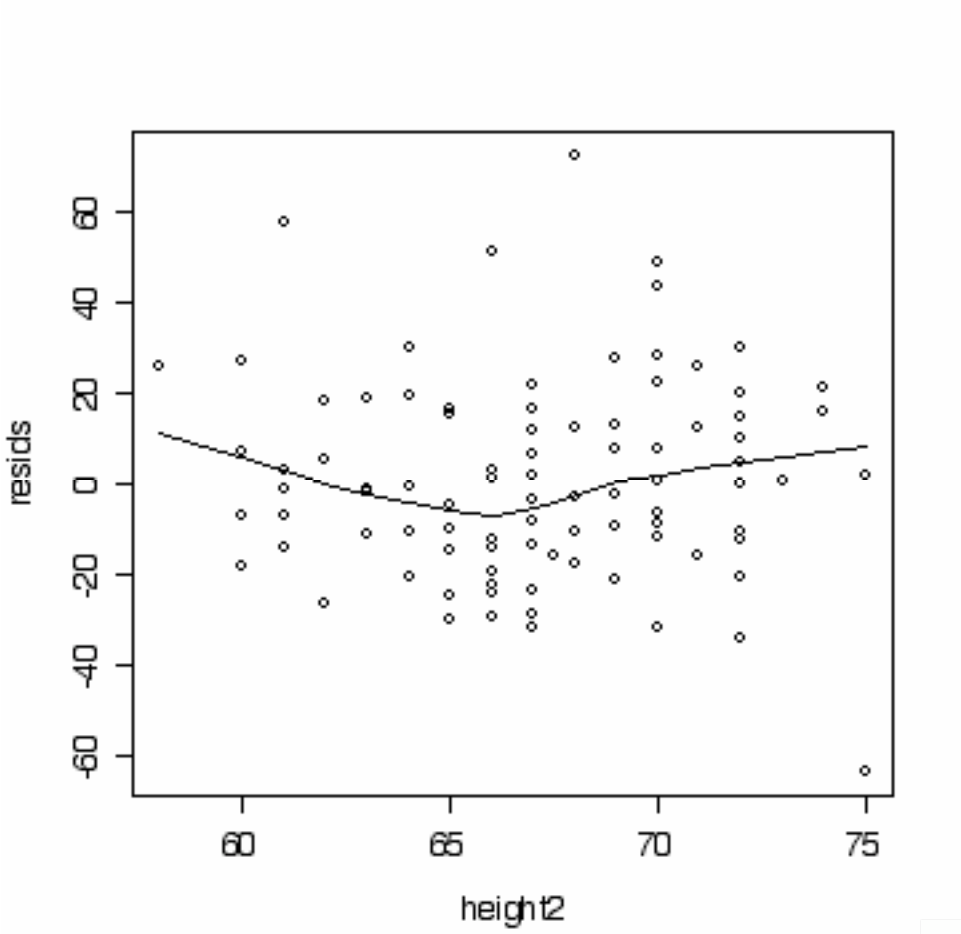
Very few of you interpreted this model. The key things to consider are the intercept and slope. The intercept, here, is not too useful since there are few people wandering this planet who are 0" tall.

The slope, though, tells us that people who differ in height by 1" differ in average weight by 4.3 pounds. Thus, each inch of height "weighs" about 4.3 pounds, on average. Note that it is very wrong to say "as height increases, weight increases" since we did not (and could not) observe height change.

e) Plot the residuals against height2. Do you see any deviations from the model? (if output.lm is the output of your least squares fit, then resid <- residuals(output) will give you the residuals. Then do plot(resid, height2).



At first glance, this looks pretty good. But its sometimes hard for the eye to pick out trends, so lets put a lowess curve through it:



Now we see what looks like a quadratic relationship -- which suggests that we left something out of our model or didnt' fit it right. I think, however, that the problem is combining men and women. We'll be able to explore this more when we've covered multiple regression.

f) What assumptions must you make if we are to be able to interpret the p-values provided in the output?

I apologize that we didn't really get to cover this in time for the homework. Most of you carried on valiantly and I appreciate that. The assumptions you need to make are that the data are a random sample, that heights are normally distributed, and that the variation about the regression line is the same for all heights.

g) Algebraically solve the least squares regression line to find the equation using weight to predict height.

weight = -154.05 +4.346 height
(weight + 154.05) = 4.346 height
height = 35.45 + 0.230 weight

If this is truly the regression line for using weight to predict height, it suggests that people who weight one pound more are .230 inches taller, on average.

h) Now fit a regression line using weight to predict height. How does it compare to your algebraic solution in (g)? Why do you think this is?

survey2.lm <- lm(height2~ weight2)
predicted height = 55.11 + 0.0865* weight

The two lines are very different. The reason for this is that the regression line minimizes the vertical distances between the points and the lines. This means that you can't simply "solve for" the other variable to turn the line around. So there are really two regression lines.

i) Now fit a least squares regression line using height to predict weight but using the original variables (that include the outliers). Do the outliers have a big affect on the fitted model?

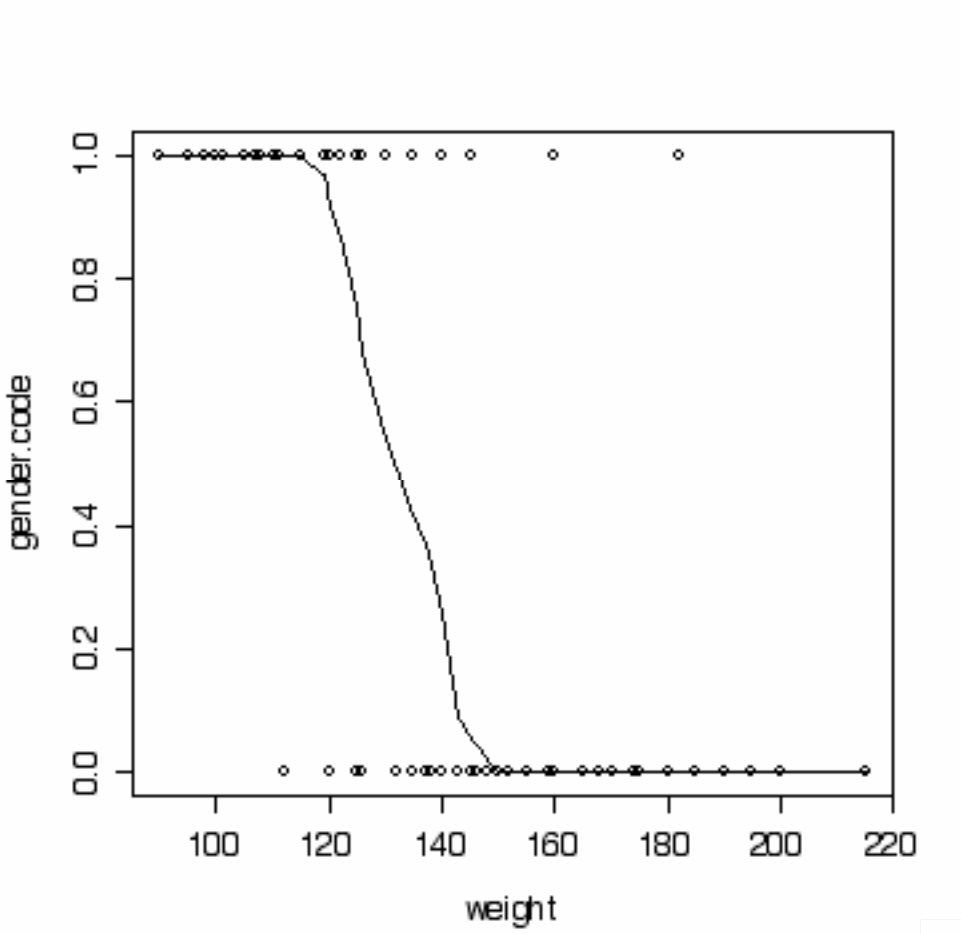
You should see a tremendous difference. The slope, for example, changes signs and suggests that the taller you are, the less you will weight, on average.

C1. In this problem we'll turn things on their side and use height to predict gender.This problem will use the same survey data. as in B. But first, we need to create a new variable. The gender variable has two values, "m" and "f". We want to recode this as 0 and 1. Do the following:
gender.code <- c(length=length(gender)) (creates a null vector with same length as gender.
gender.code[gender=="m"] <- 0
gender.code[gender=="f"] <- 1

Now the men are coded with 0's and the women with 1's.
a) Make side-by-sdie boxplots of weight for men and women. Compare.

The men are, typically, heavier than the women. Not surprising. But there's a fair amount of overlap. If there were no overlap it would be easy to predict gender from weight. But the overlap means for intermediate weights this may not be so easy.

b) Although there is considerable overlap in weight, still the means are different. Suppose I randomly select a person from this class and tell you their weight. Can you predict which gender they are most likely to be? This is a bit unusual compared to other problems we've seen: the response variable is binary and has values 0/1 while the predictor is continuous. One way to think about this is that you want to know the probability that a person is female given that you know their weight. So for each weight, we can estimate it by finding what percentage of people at that weight are women. Plot gender.code against weight (weight on x axis) and fit a lowess line. What does the line tell you about the probability that a randomly selected person is a woman if you know their weight?



Remember that the lowess curve is calculating average y-values in each small neighborhood of an x value. So if x is, say 100, it's averaging the y-values of people who weight around 100 pounds. Now the y-values are all 0's and 1's, so an average of 0s and 1's is a percent. This means that the lowess curve is estimating the percentage of women at each weight value.

One interpretation of this, as many of you said, is that for weights of about 150 and higher, we should predict the person is male (since nearly 100% of them are) and for weights lower than 120 or so, predict female.

Some people swapped the axes, which makes this question harder to answer.

c) Find the mean weight of men and of women.

```
> mean(weight[gender=="f"])
```

```
[1] 118.9615
> mean(weight[gender=="m"])
[1] 155.7759
```

Note that these values are very close to where the lowess curve flattens out.

d) Fit a regression line using weight to predict gender. Interpret the coefficients and compare to your answer in (c).

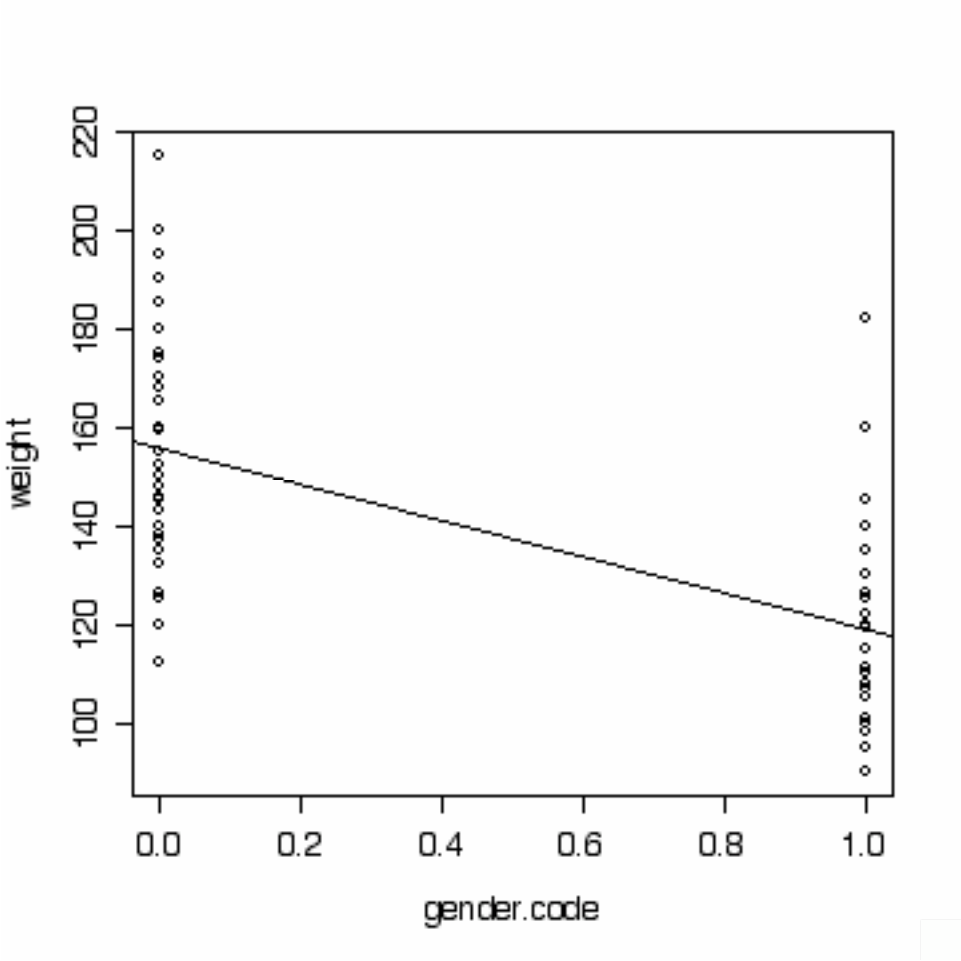
predicted gender = 2.23 - 0.012(weight)

OK, thanks to cutting and pasting errors, I screwed this up. As many of you pointed out, there really is not relationship between the coefficients and your answer to (c).

The only relation is this: if you put average x in, you get average y out. So if we put in average weight (138.37), we should get average y (which here would be the percent of women in the class) out: 2.21-0.012*138.37) = .549.

The interpretation of these coefficients is tricky. First, many of you pointed out that the relation is far from linear and the linear model fits really horribly. True. We'll return to this with a nice solution later when we study logistic regression. But the main thing to understand is that the slope is telling estimating the percentage of women at each weight, and hence for every additional pound, there is, on average, 1.2% fewer women at that weight. This means the probability that a person is female decreases by about 1.2% per pound. Assuming the model fits -- which it doesn't.

C2. Now we reverse things.
a) Fit a regression using gender to predict weight. Interpret the coefficients. (It might help to look at the plot of gender (y axis) vs. height.)



predicted weight = 155.776 -36.81432*gender

Look at that! The intercept is meaningful now. It is the predicted weight when gender.code=0 , which means that it's the average weight of the men! (This is what question (d) was meant to refer to. Now refer to your answer in (c).)

The slope represents the difference in the average weights of men and women. Many of you said that it represents a "one unit change in gender", but explore this further. Gender only has two values (in this dataset). So a one unit change means going from 0 to 1. And So this means that the mean weight of females (gender=1) is 36.8 pounds less than the mean weight of males (gender=0).

b) Do a t-test to test whether the mean weight of men and the mean weight of women are equal. Assume that the standard

deviations ARE EQUAL in these two populations. What assumptions must you make to carry out this test?

I think only one person did the test assuming that the standard deviations are equal. Here's what happens if you do:

```
> weight.m <- weight[gender=="m"]
> weight.f <- weight[gender=="f"]
> t.test(weight.f, weight.m, var.equal=T)

     Two Sample t-test

data:  weight.f and weight.m
t = -9.7754, df = 108, p-value = < 2.2e-16
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -44.27924 -29.34941
sample estimates:
mean of x mean of y
 118.9615  155.7759
```

We reject the null hypothesis (not surprisingly) and conclude that the difference in mean weight cannot be due to chance alone. Note that the observed value of the t-statistic is -9.7754.

c) Now compare the value of the t-statistic in part (b) with the t-statistic that corresponds to the slope you got in part (a). What does this tell you about the meaning of the t-test associated with this slope? What assumptions must you make in order for this t-test to be valid?

```
> gender.lm <- lm(weight~gender.code)
> summary(gender.lm)

Call:
lm(formula = weight ~ gender.code)

Residuals:
    Min      1Q  Median      3Q     Max
-43.776 -13.962  -3.869  10.585  63.038
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  155.776      2.589  60.161   <2e-16 ***
gender.code  -36.814      3.766  -9.775   <2e-16 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 `
' 1

Residual standard error: 19.72 on 108 degrees of freedom
Multiple R-Squared: 0.4694,   Adjusted R-squared: 0.4645
F-statistic: 95.56 on 1 and 108 DF,  p-value: < 2.2e-16
```

Note that the observed value of the t-statistic is 9.775.

The moral of the story:
A two-sample t-test (assuming variances are the same in both groups) is the same as doing a regression in which the predictor variable is a "dummy variable" in which a 1 indicates membership to one population and a 0 membership to the other.

The assumptions of a two-sample t-test are:

- both populations are normally distributed
- samples are selected independently from each other, and each group is independent of the other
- variances are the same in both populations

Because the tests are equivalent, we need the same set of assumptions to make the interpretation of the t-statistic in the regression "work".

D.
a) Fit a regression line to Galton's data of father/son heights. (Once you've downloaded the file to the right directory, simply type source("filename")

Predicted sons' height = 33.8866 + .51409 * father's height

b) Use the regression line to find out how many standard deviation above average are typical sons whose fathers are 1 standard deviation above average?

```
> mean(fheight)
[1] 67.6871
> sd(fheight)
[1] 2.744868
```

A father who is one standard deviation above average is mean(fheight) + sd(fheight) = 70.4" tall.

The regression says that the mean height of the sons of these fathers is
33.88866 + .51409*70.4 = 70.0806
And this height is
(70.0806-mean(sheight))/sd(sheight) = .4961557
So these sons are only about half a SD above average.

c) How many standard deviations below average are typical sons whose fathers are 1 standard deviation below average?

These sons are about a half SD below average.

Note that both groups of sons are close to the average than their father's were.

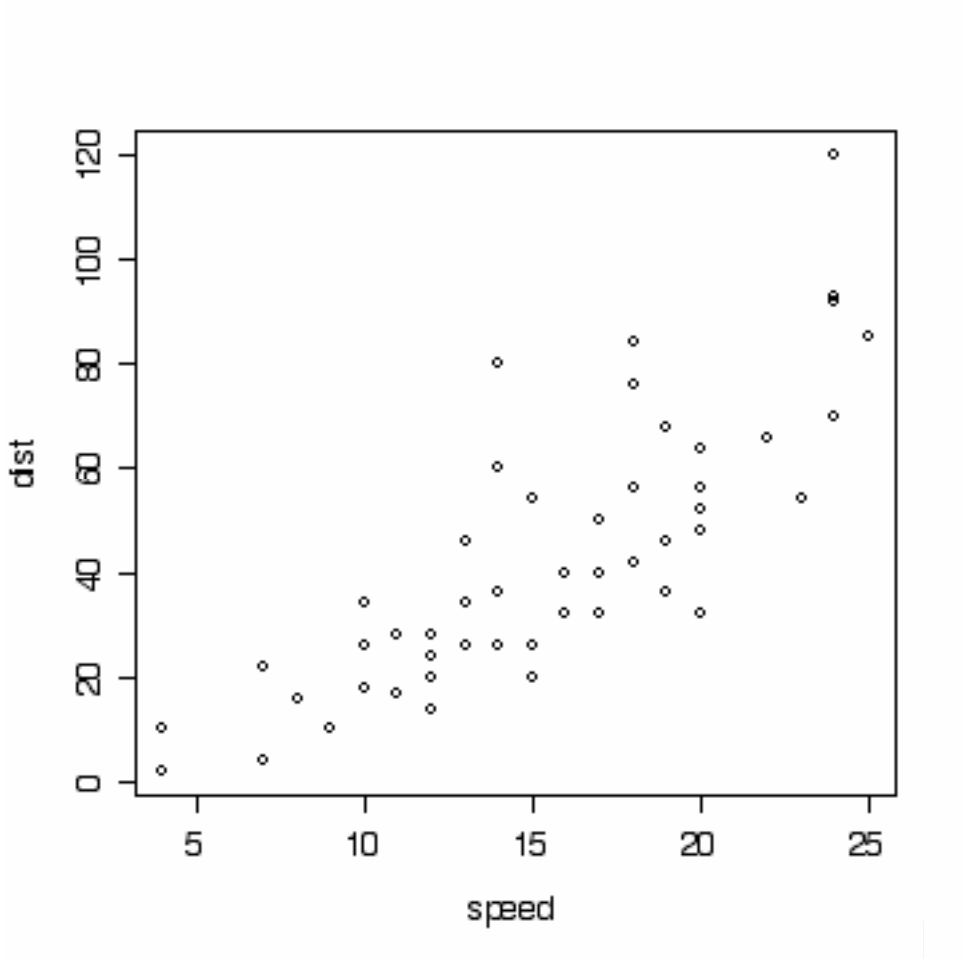
Galton called this phenomenon (that taller than average and shorter than average fathers produce more "mediocre" sons) "regression to the mean."

d) Suppose a certain large class takes two midterms. What does regression to the mean say about the people who get A's on the first midterm? (For the sake of conversation, suppose that to get an A you need to be 2 standard deviations above average). Assume that r = .4 and the standard deviations of both exams are the same.

The regression to the mean phenomenon means that people who tend to be far above average on the first midterm will be closer to average on the second. This means that some people who got A's on the first MT will not on the second midterm.

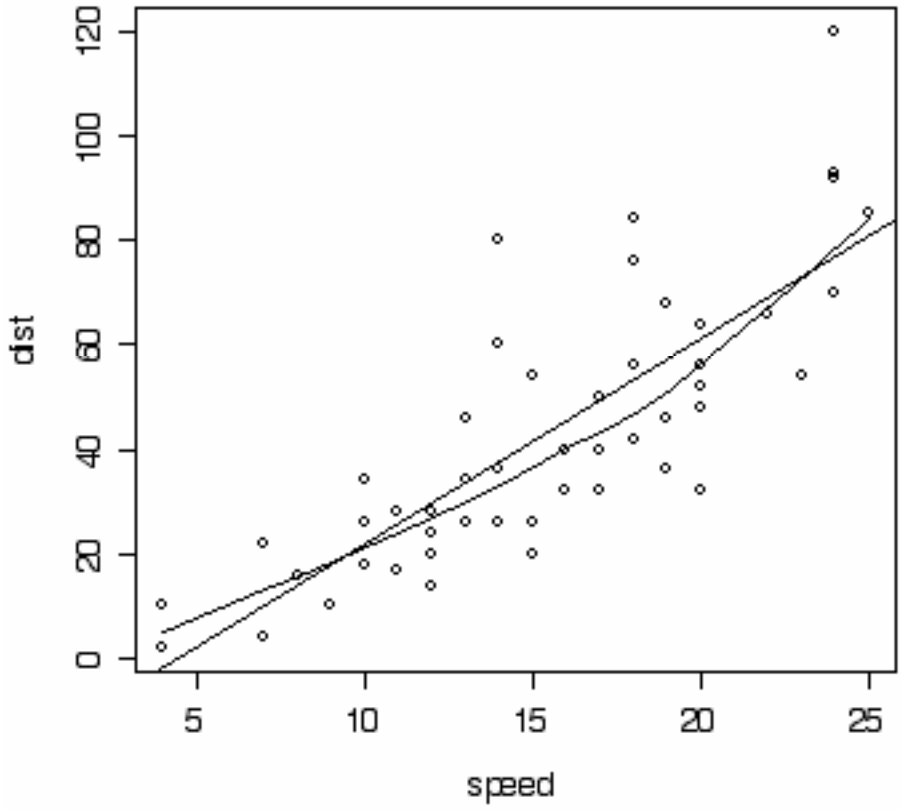
We can be more precise. The slope is r(SDy)/SDx = .4 here. So of all of the people who scored 3 SDs above average on the first midterm, the would on average score only .4*3 = 1.2 SDs above average on the second.

E. Read #3 on p. 132 . You can get the data very simply; just type data(cars). Type names(cars) to see the variables. Type attach(cars) to use the variables by their first name. The data consist of stopping distances for a car at various speeds.
a) Make a plot to help us determine how much distances is needed to stop for a given speed.



Make sure you put the right variable on the x-axis. As you saw in Problem B (parts g and h), it matters which variable is the predictor and which the response. This shows us the not-too-surprising result that the faster you're moving, the more distance it takes to stop.

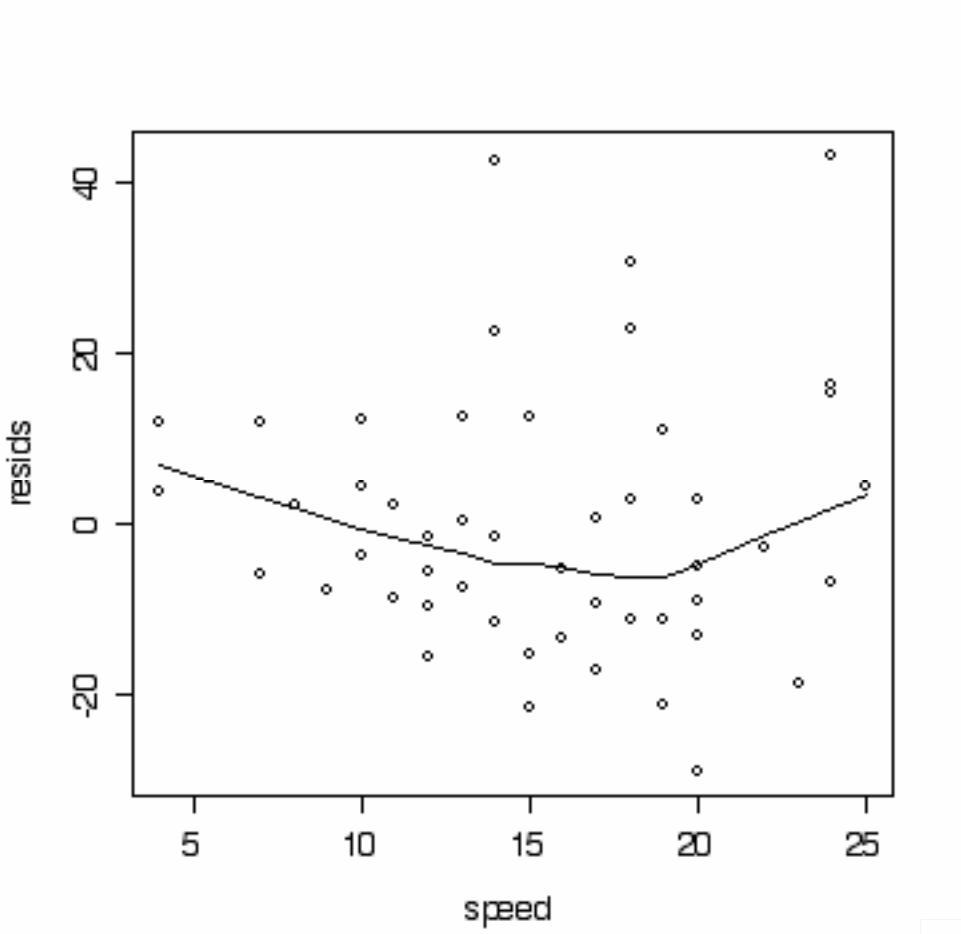
b) plot the least squares regression line and comment on whether the model fits. Superimpose the lowess line and compare.



The lowess curve shows a curve. It suggests that at mid-level speeds, the regression line will be over-predicting, and at the highest and lowest speeds the regression line will under-predict.

c) Make a plot of the residuals to confirm your answer to (b).

```
> resids <- residuals(speed.lm)
> plot(speed, resids)
> lines(lowess(speed,resids))
```



The curve is pretty apparent in the residuals (but you might need the lowess curve to see it).

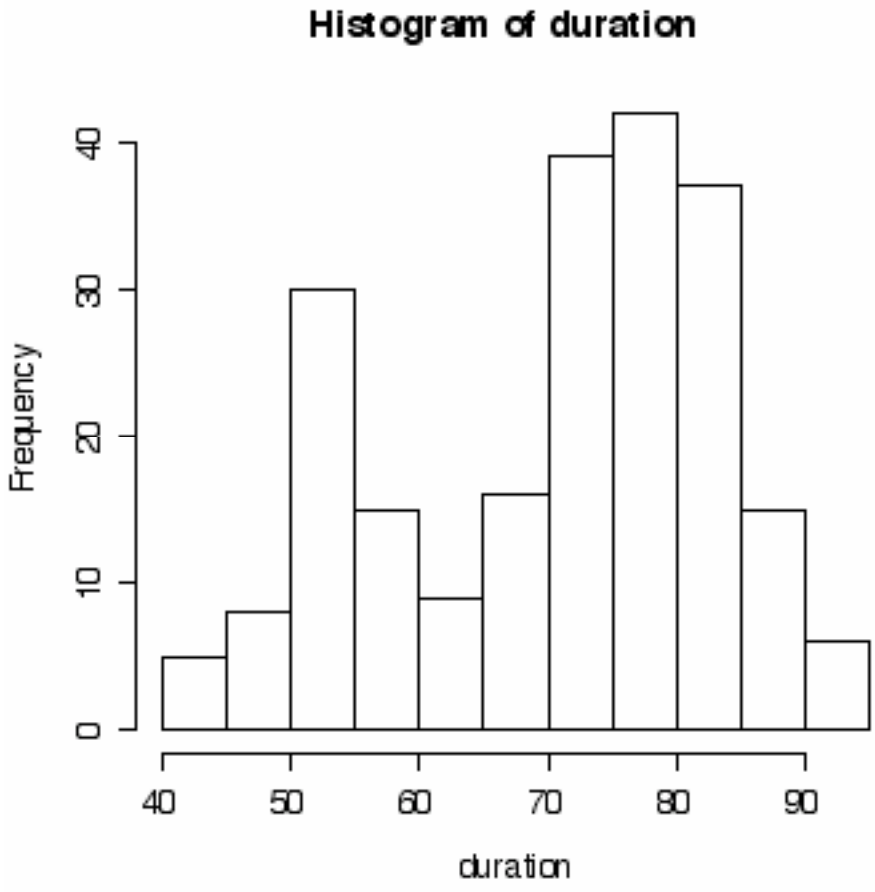
d) Now we'll try transforming the y-variable. Fit a curve of distance^2 to speed. Repeat for sqrt(distance) and log(distance). Which fits best and why?

sqrt(dist) fits best. This should remind you of freshman physics -- speed and stopping distance are related through a quadratic relation.

F. Download the oldfaithful data into R. This consists of information about eruptions of the Old Faithful Geyser in October 1980. The length variable is the number of minutes that an eruption lasts and duration is the time in minutes to the next eruption. Old Faithful has many tourists and this data (well, data

like these) are used to derive an equation that helps park rangers tell tourists when to expect the next eruption.

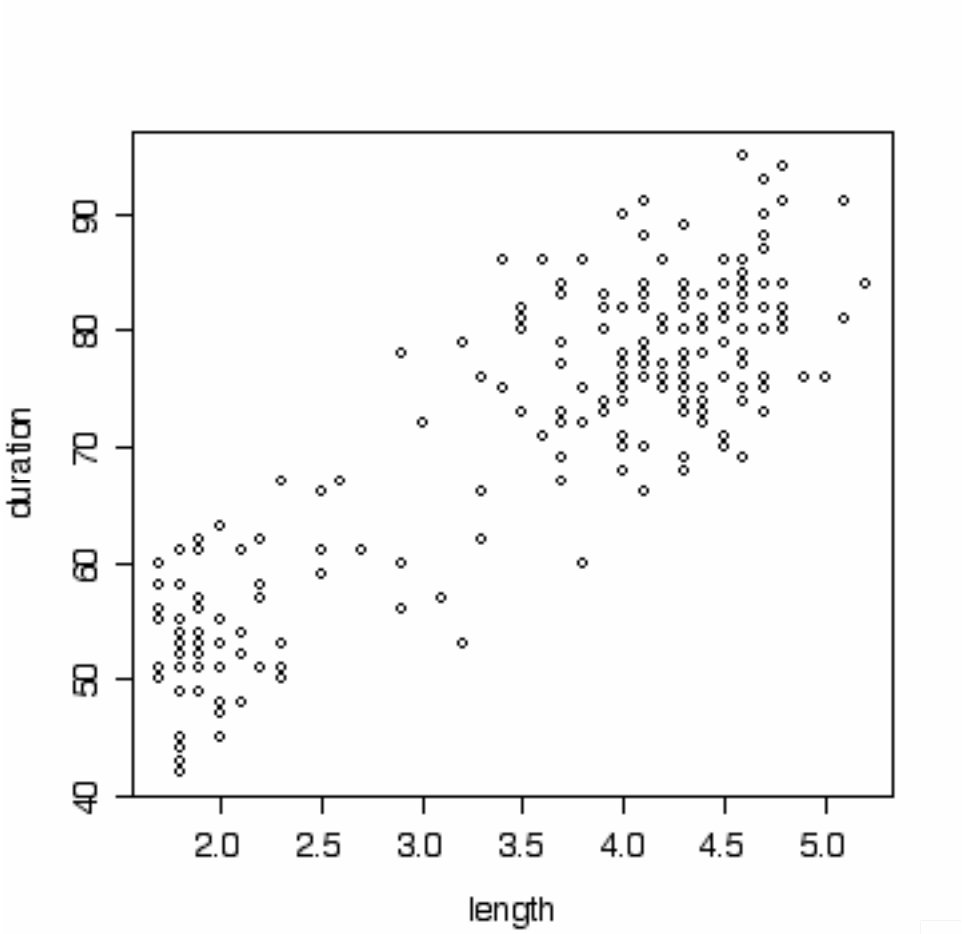
a) Make a histogram of duration. Describe the distribution. What would you say is the typical waiting time for the next eruption?



The most important feature of this distribution is that it's bimodal. This suggests that a single-number summary is perhaps misleading. A better summary would be to say that waiting times are typically 50 minutes or 75 minutes (roughly), and it seems to depend on another variable to determine which of the two populations we're dealing with.

b) Make a graph to show the relation between the length of the eruption and the time until the next eruption. Describe this relationship in words. Does it make sense?

c) Find an equation that predicts the time to the next eruption based on the length of the current eruption.



This plot tells us that how long you need to wait for the next eruption can be better predicted if you know how long the previous eruption lasted. Longer eruptions typically result in a longer wait until the next eruption. This is not surprising --- it takes more time for the gasses to "refill" after a long eruption.

The linear model is
predicted duration = 33.97 + 10.36*length.

So if an eruption lasted 3 minutes, we'd predict that the next will occur in 3*10.36+33.97 = 65 minutes.

Also, on average, every additional minute that the eruption lasts means, on average, an additional wait of 10.4 minutes for the next one. Further, the intercept shows us that you'll have to wait at list 34 minutes between eruptions.

G. (optional). The EPA publishes fuel-efficiency data for a variety of types of cars. http://www.epa.gov/fueleconomy/data.htm. Download the 2005 data here. It is a text file in which values are separated by commas, and so you load it into R by typing read.table("fuel.csv", header=T, sep=",").
You can view the variable names with the names() function. You can see what the names mean by reading this file, although you'll have to scroll down a bit.

In general, cars with bigger engines are less efficient. One measure of the size of an engine is the displacement which measures, in liters, the amount of fuel moved through the engine. Come up with the best equation you can to predict the highway mileage (hwy) based on displacement (displ). Evaluate your model. Use the identify function to identify any unusual cars. (For example, which car has the worst fuel efficiency? the best?)