Lecture 5

We begin by examining some interesting (?) two-variable data sets.

the first is historic, and was one of the primary movers and shakers that led Galton to "invent" regression. Galton was interested in exploring the role that genetics plays in determining physical characteristics, and this data set was intended to help him understand the extent to which height was genetically determined.

It was casually understood that tall fathers tend to produce tall sons. And also, in fact, that subsequent generations were growing taller. in fact, in this sample, the sons are an average of about 1 inch taller than the fathers. But Galton tried to quantify this. If you're 1" taller than average, how much taller than average is your son likely to be? If you're 1" shorter than average?

```
> mean(fheight)
[1] 67.6871
> mean(sheight)
[1] 68.68407
```

As you can see, for any given fathers' heights, there's substantial variation.

How would you describe this variation? How would you describe the shape?

Three things you should look for:

- Shape --- is it linear or not? Exceptional points?
- Trend -- positive, negative, hard to say?
- Strength--- how much variability about the trend?

This is a good example of a linear relation. In fact, one of the clearest examples you'll ever see. Rarely do you see such a nice, elliptical cloud.

One outcome of linear regression is, for the special case of linear relations, to extract the trend from this apparent cloud of points.

Strength is measured, at least for a linear relation, in terms of the correlation. The correlation for a sample is a number, r, such that -1 < = r <= 1. If the points fall precisely on a line, then r = +/- 1. If they are "patternless", then r = 0 (or close to it.)

But this interpretation works only for *linear* relations. Your homework shows how various non-linear relations can "mask" for almost any corelation. In addition, 0 correlations do not mean NO relationships. They mean no *linear* relation.

One implication of this is that while two indpt (unrelated) variables will always have a 0 correlatoin, a 0 correlation doesn't necessarily mean they're indpt (unrelated).

One common goal of a regression is to make a prediction. So for example, given the height of a father, I want to predict how tall his son will be.

This plot makes it very clear that such a prediction must be vague. None-the-less, we can get some information out of it.

For example, given a father's height, what's the average height of the sons?

Let's take 66" -- jus ta bit below average. Turns out there no son's whose fathers were exaclty 68 inches. So we need to be a little more vague. Let's look at those betwen (65.5, 66.5)

```
> son.66 <- sheight[fheight>65.5 & fheight < 66.5]
> length(son.66)
[1] 138
> mean(son.66)
[1] 67.66685
```

So one answer to the question: "How tall or the songs of men who are 66" tall?" Is: on average, they tend to be about an inch taller: taller: 67.67"

Let's take an average father: (about 67.7")

```
> son.favg <- sheight[fheight>67.2 & fheight < 68.2]
> length(son.favg)
[1] 152
> mean(son.favg)
[1] 68.84971
```

So if father is average, the son is pretty close to average -- although again about an inch taller than the father.

How about the 69" fathers?