A Brief History of Regression
Heights of Fathers and their Sons

- The scatter plot to the right depicts data collected by Pearson and his colleagues in the early 1900’s.
- It consists of 1078 pairs of heights of father and their sons.
- The plot is shaped like an American football, with a dense center and fewer points around the perimeter.
The histograms of the fathers’ and sons’ heights are both bell-shaped.
The histograms mostly overlap.
But sons are about an inch taller than their fathers, on average.

```r
> summary(heights)

          Father          Son
  Min.   :59.00   Min.   :58.50
  1st Qu.:65.80   1st Qu.:66.90
  Median :67.80   Median :68.60
  Mean   :67.69   Mean   :68.68
  3rd Qu.:69.60   3rd Qu.:70.50
  Max.   :75.40   Max.   :78.40
```
Correlation in their Heights

The correlation in their heights is exactly what leads to the American football (i.e., ellipsoidal) shape.

```r
> pearson <- read.csv("pearson.csv")
> cor(pearson$Son, pearson$Father)
[1] 0.5011627
```
Histogram of the Differences

The bulk (95%) of the data lie between -4.4 and 6.4 inches.

> summary(heights)

<table>
<thead>
<tr>
<th></th>
<th>Father</th>
<th></th>
<th>Son</th>
<th></th>
<th>Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.</td>
<td>59.00</td>
<td>Min. 58.50</td>
<td>Min.</td>
<td>-9.0000</td>
<td></td>
</tr>
<tr>
<td>1st Qu.</td>
<td>65.80</td>
<td>1st Qu.66.90</td>
<td>1st Qu.</td>
<td>-0.8000</td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>67.80</td>
<td>Median 68.60</td>
<td>Median</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>67.69</td>
<td>Mean 68.68</td>
<td>Mean</td>
<td>0.9974</td>
<td></td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>69.60</td>
<td>3rd Qu.70.50</td>
<td>3rd Qu.</td>
<td>2.7750</td>
<td></td>
</tr>
<tr>
<td>Max.</td>
<td>75.40</td>
<td>Max. 78.40</td>
<td>Max.</td>
<td>11.2000</td>
<td></td>
</tr>
</tbody>
</table>
The Regression Effect

- We might expect the sons of tall fathers to be tall as well.
- This histogram shows the heights of sons of 72 inch fathers.
- Most (68%) of these sons are less than 72 inches tall!
The Regression Effect (cont’d)

- This is surprising!
  - Sons are an inch taller than their fathers, on average.
  - But sons of tall fathers are an inch shorter than their fathers!

```r
> tall_fathers <- heights %>% filter(Father >= 72)
> mean_tall_fathers <- tall %>% summarize(father = mean(Father), son = mean(Son), diff = mean(Diff))
> mean_tall_fathers

father     son     diff
1 72.8178 71.4575 -1.36027
```
History of the Regression Effect

- The regression effect was first documented by the statistician Francis Galton, who had thought (hoped, even) that tall fathers would have tall sons.
- These data show that tall fathers’ sons were not quite as tall.
- Galton, who is sometimes called the father of eugenics, called this effect “regression to mediocrity”.
- Galton also noticed that short fathers had sons who were somewhat taller than their generation on average. Today, this is called the regression effect.
- Individuals who are below average after a first measurement tend to move towards the mean after a second, and vice versa. Why?
The Regression Effect, Explained

- Imagine pre-test and a post-test measurements for a set of individuals who receive a null treatment (i.e., a placebo).
- Some individuals will test below the mean, and others will test above.
- Assuming perfect measurements, those who test below (or above) in the pre-test will do so for one of two reasons. Either:
  - Their measurements are truly below (or above) the mean, or
  - Random fluctuations
- In the post-test, if they are truly below (or above) the mean, they will likely measure that way again. But if their measurement error was due to random fluctuations, they will move in the direction of the mean!
- So, conditioned on measuring below (or above) the mean in the pre-test, measurements will be closer to the mean in the post-test!
Fitting a Regression Line in R

The blue line follows the angle of the cloud of points, and is called the regression line.

> attach(pearson)
> plot(Father, Son, col = "red")
> fit <- lm(Son ~ Father)
> abline(fit, col = "blue")
> detach(pearson)

Coefficients:

|             | Estimate | Std. Error | t value | Pr(>|t|) |
|-------------|----------|------------|---------|----------|
| (Intercept) | 33.89280 | 1.83289    | 18.49   | <2e-16 *** |
| Father      | 0.51401  | 0.02706    | 19.00   | <2e-16 *** |
The Regression Line, in Standard Units

- This scatter plot depicts the data in standard units.
- The black line has a slope of 1:
  - A one unit increase in father’s height leads to corresponding one unit increase in son’s.
- The slope of the regression line is less than 1. In fact, it is $r \approx 0.5$:
  - A one unit increase in father’s height leads to corresponding one-half unit increase in son’s.