Simple Linear Regression

Ice Cream Sales vs. Temperature

- If you were asked to describe the pattern between ice cream sales and temperature, you might say “ice cream sales seem to increase as temperature increases.”
- Temperature is called the independent variable, or the regressor, and Sales, the dependent variable, or the regressand.
- The independent variable is also known as the explanatory, predictor, or input variable, and the dependent variable, the response, or output variable.

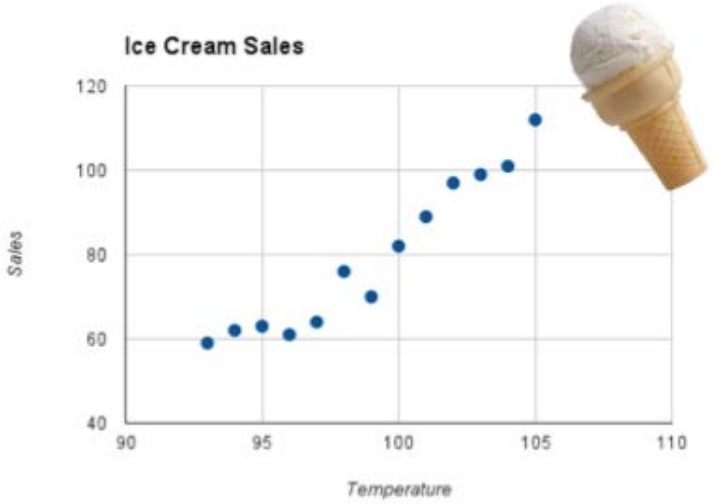


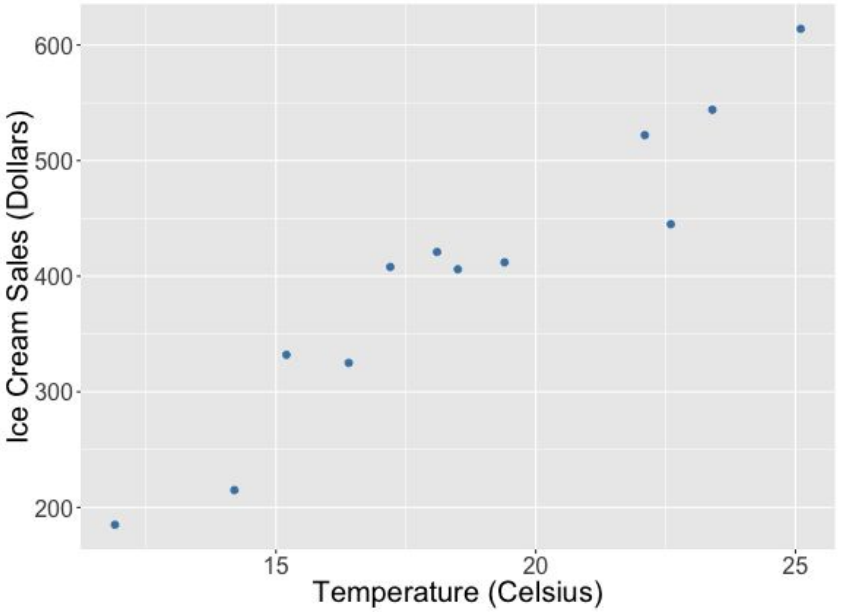
Image source

Parameters that Define Relationships

- Direction
  - Positive (direct)
  - Negative (indirect)
- Form
  - Linear
  - Non-linear
- Strength (weak, strong, moderate)
- Caution: Outliers

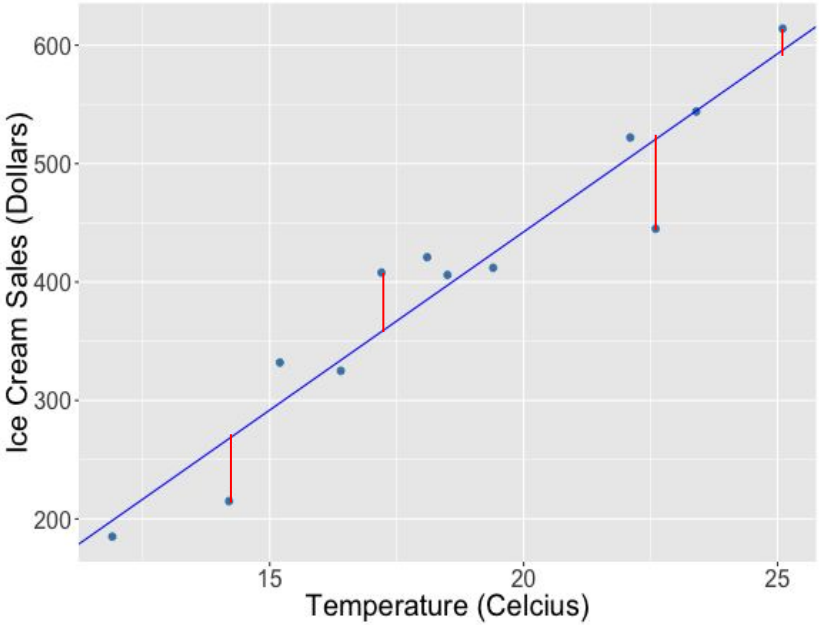
Simple Linear Regression

- Linear Regression is the study of linear, additive relationships between variables
- With simple linear regression, we fit a line to data, thereby describing the linear relationship between exactly two variables.



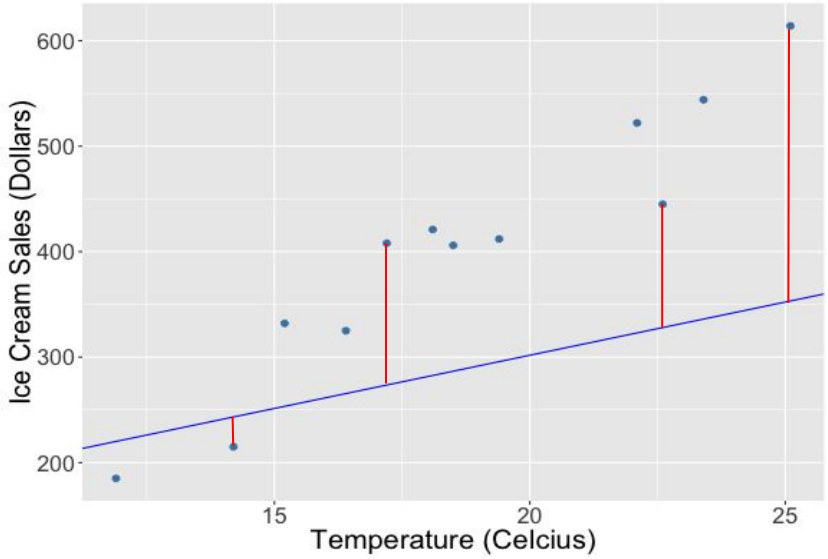
The Formal Problem Statement

- Find the line that “best” fits the data
- More precisely: given a set of (*x*, *y*) pairs, find a line such that the squared distance between each of the points and the line is minimized.
- This distance is called the residual. So the formally, the regression problem is to minimize the sum of the squared residuals.

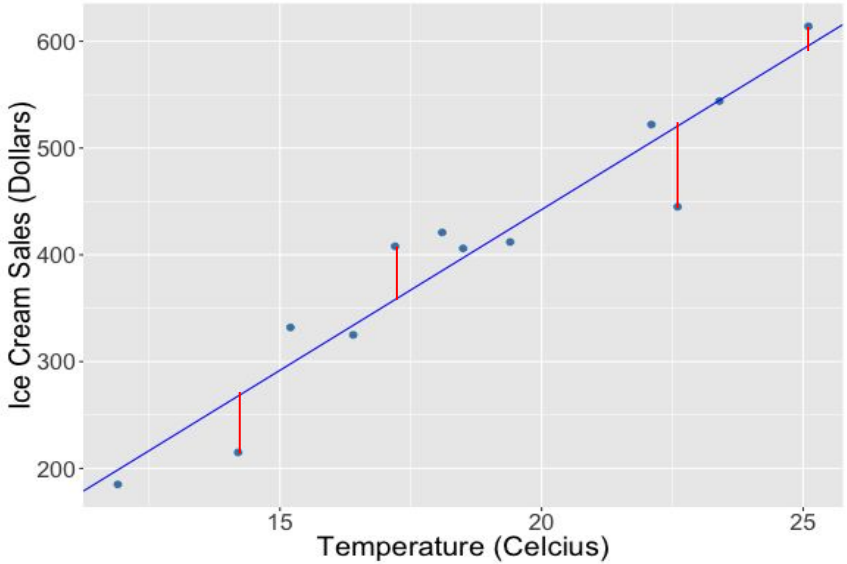


Fitting the “best” line

- The errors would be much larger if we fit this line to our data



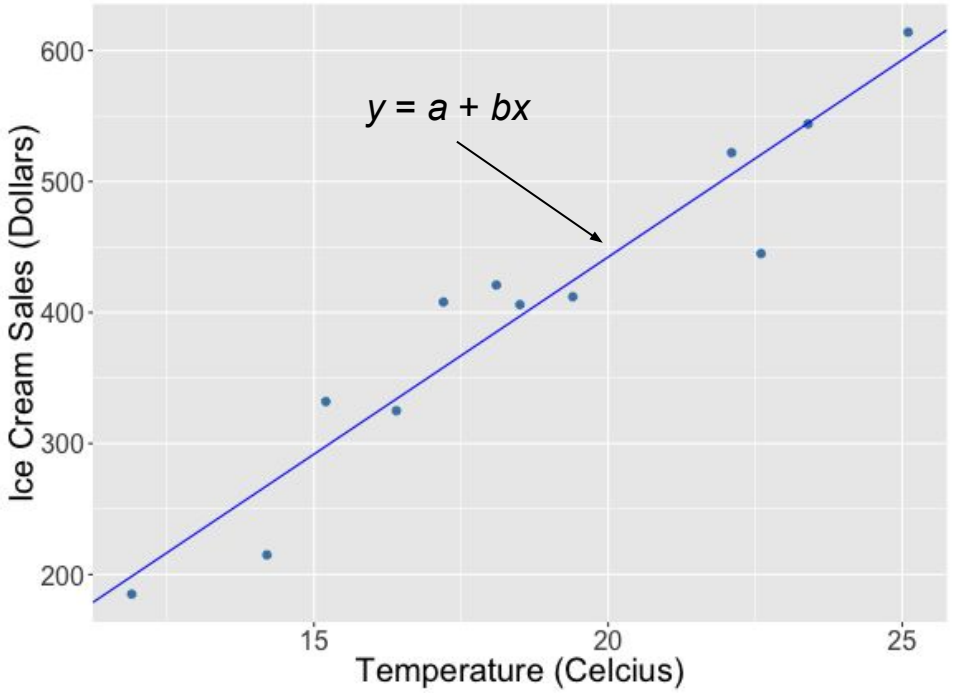
- This line minimizes the sum of the squared residuals



The Regression Equation

The regression equation takes the following form: $y = a + bx$

- y is ice cream sales in dollars
- a is the y intercept of the line (the values of y when x is zero)
- b is the slope of the line
- x is temperature in celsius



Linear Regression in R

- The regression equation for Ice Cream Sales versus Temperature is:

  Sales = -122.99 + 28.43 (Temperature)

- *b* = 28.43 is the slope. For a one degree increase in temperature, sales are predicted to increase by 28.43 dollars.
- *a* = -122.99 is the y intercept. This value has no particular meaning; it definitely does not mean that when temperature is zero, sales are predicted to be -122.99 dollars!

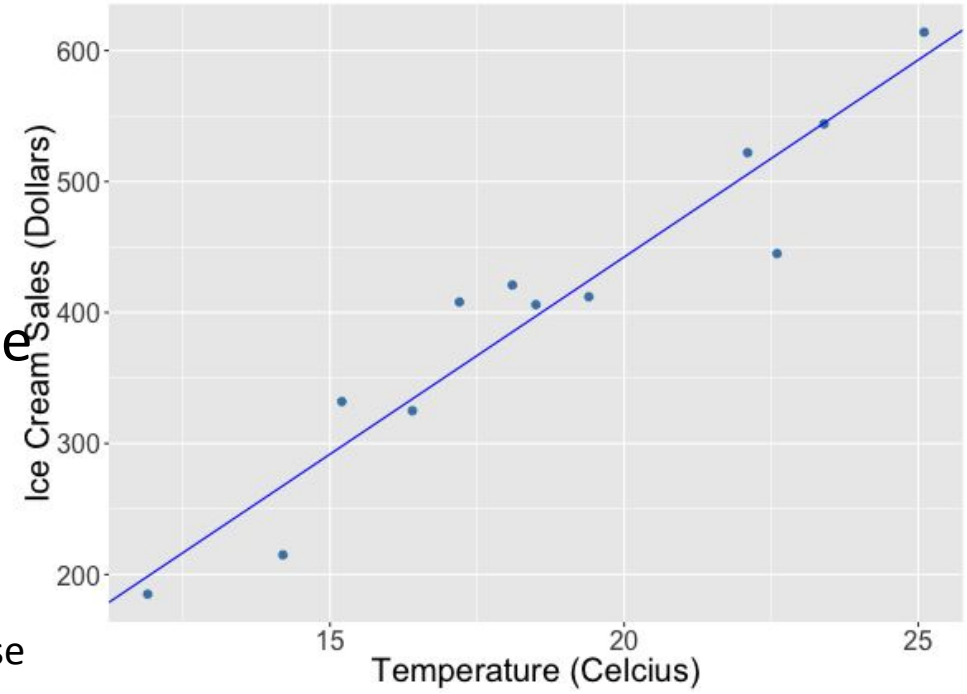
```
> attach(ice_cream)
> lm(sales ~ temp)

Call:
lm(formula = sales ~ temp)

Coefficients:
   (Intercept)            temp
       -122.99           28.43
```

Interpreting the Regression Line

In dollars: $y = a + bx$

- The intercept a = -122.99
- The slope is b = 28.43

Each point on the regression line is the result of multiplying temperature by b and adding a



Caution: It is dangerous to make predictions outside the range of measured x values, because we don't know anything about the relationship between x and y outside this region.

The Problem

- Given $D = \{ (x_i, y_i) \mid i = 1, \ldots, n \}$
- Let y_i represent the ith actual value.
- Let y_i^p represent the ith predicted value.
  - $y_i^p = b_0 + b_1 x_i$
- Then, the regression problem can be formalized as:
  - Find b_0 and b_1 that minimize $\sum_i (y_i - y_i^p)^2$

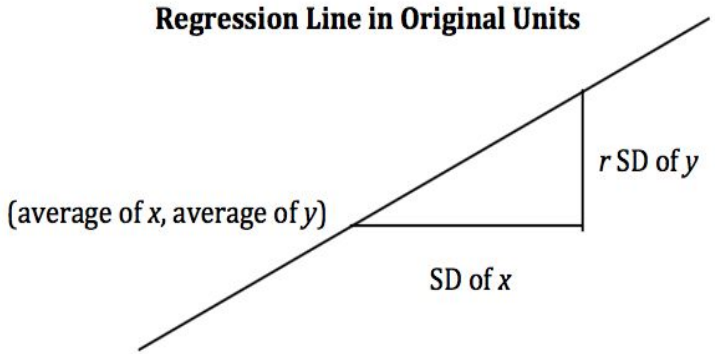
The Solution

- We want to minimize a function, so we use calculus to solve this problem.
  - Set the partial derivatives of this function equal to zero, and solve.
- After doing so, the solution to the problem is:
  - $b_1 = \sum_i (x_i - \overline{x})(y_i - \overline{y}) / \sum_i (x_i - \overline{x})^2$
  - $b_0 = \overline{y} - b_1 \overline{x}$

The Solution

- We want to minimize a function, so we use calculus to solve this problem.
  - Set the partial derivatives of this function equal to zero, and solve.
- After doing so, the solution to the problem is:
  - $b_1 = \sum_i (x_i - \overline{x})(y_i - \overline{y}) / \sum_i (x_i - \overline{x})^2$
  - $b_0 = \overline{y} - b_1 \overline{x}$
- Very interesting: $b_1 = nc_{XY} / ns_{XX}^2 = c_{XY} / s_{XX}^2$
  - c_{XY} is sample covariance
  - s_{XX} is sample variance
- But remember $r_{XY} = c_{XY} / s_{XX} s_{YY}$, so $b_1 = c_{XY} / s_{XX}^2 = r_{XY} (s_{YY} / s_{XX})$
  - So: if we regress on the *z*-scores of the data (instead of the data values themselves), so that $s_{XX} = s_{YY} = 1$, the slope of the regression line equals the correlation of *X* and *Y*!

The Slope of the Regression Line



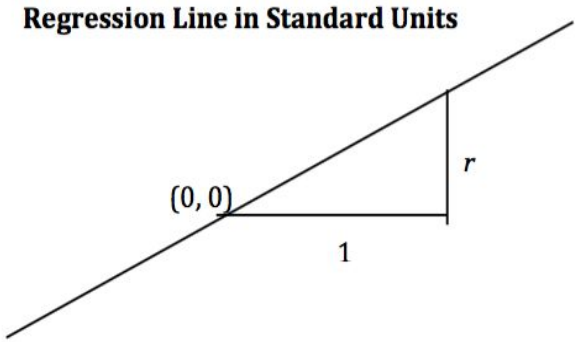


Image source

Interpreting the Regression Line

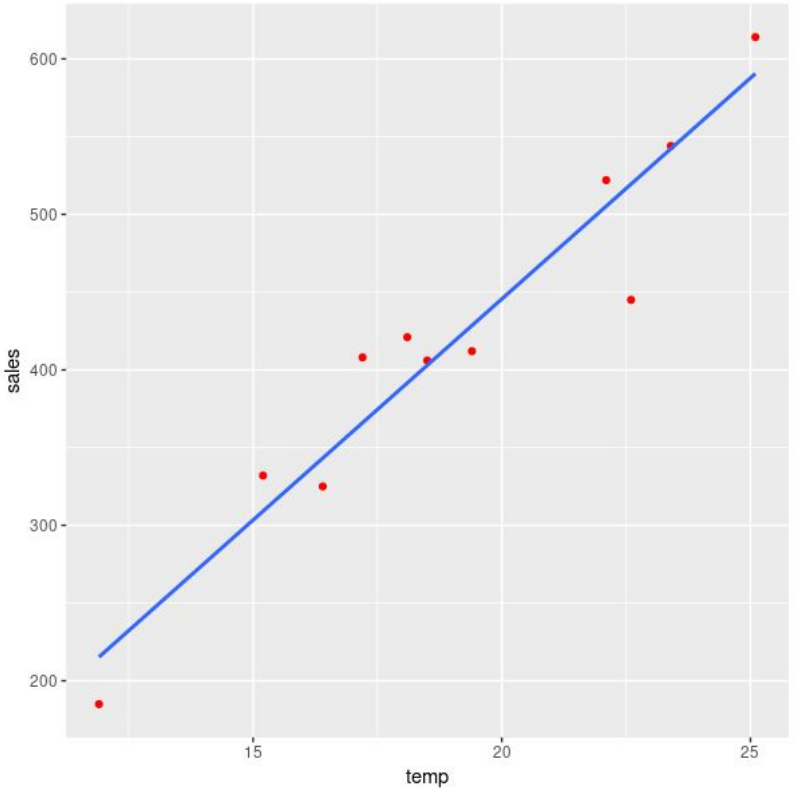
- $b_1 = \sum_i (x_i - x)(y_i - y) / \sum_i (x_i - x)^2 = c_{XY} / s_{XX}^2$
  - The slope differs from zero the more *Y* covaries with *X*.
  - The slope tends towards zero the more *X* alone varies.
- $b_0 = \bar{y} - b_1 \bar{x} = \bar{y} - (c_{XY} / s_{XX}^2)\bar{x}$
  - The intercept defines a line with slope b_1 that passes thru the point $(\bar{x}, \bar{y})$.
  - When *x* equals zero, the intercept is the mean of the dependent variable, $\bar{y}$.
  - If *x* never equals zero, then the intercept has no intrinsic meaning.
  - Caution: It is dangerous to make predictions outside the range of measured *x* values, because we don't know anything about the relationship between *x* and *y* outside this region.

Interpreting the Regression Line

In dollars: $y = a + bx$

- The intercept a = -159.47
- The slope is b = 30.9

Each point on the regression line is the result of multiplying temperature by b and adding a

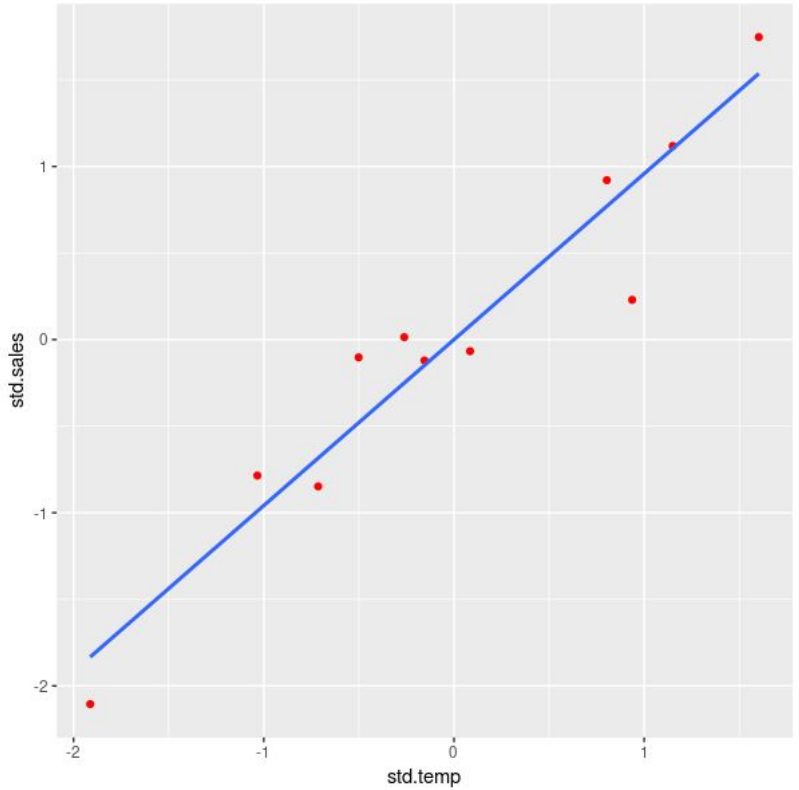


Interpreting the Regression Line (cont'd)

In standard units: $y = rx$

- The intercept is 0
- The slope is r

Each point on the regression line is the result of multiplying temperature in standard units by r (and adding 0)



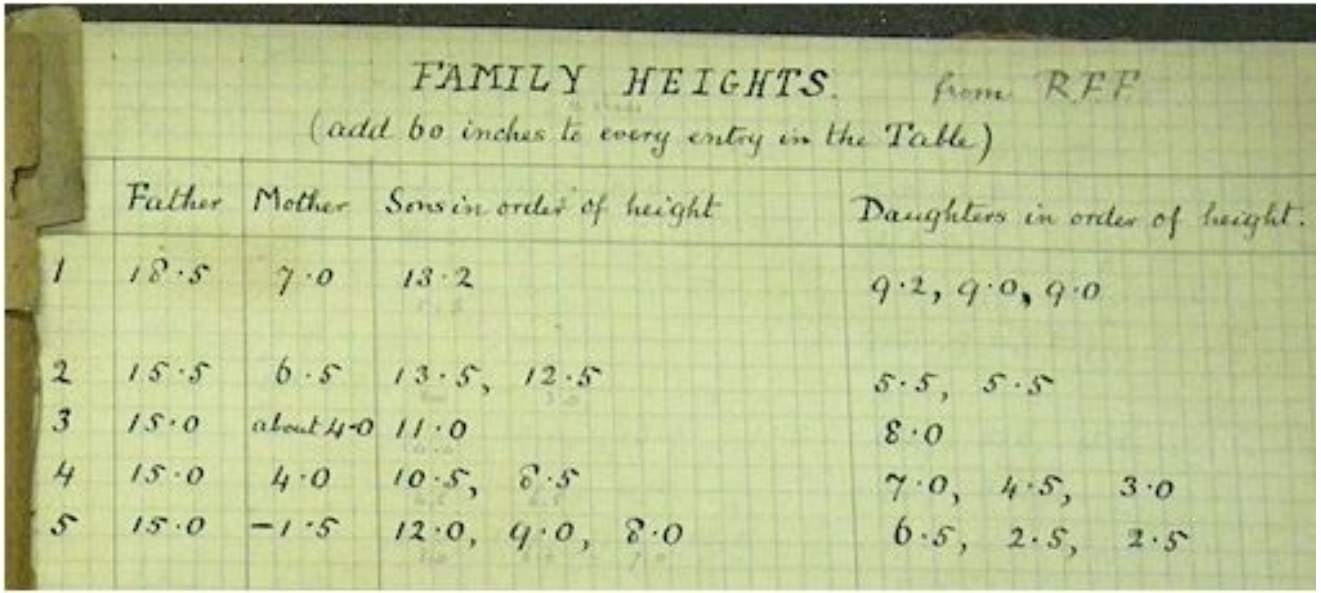
BTW, the sum of the residuals is zero ...

$$
\begin{aligned}
\sum_{i=1}^{n}(y_i - \hat{y}_i) &= \sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i)) \\
&= \sum_{i=1}^{n}(y_i - ((\bar{y} + b_1\bar{x}) - b_1 x_i)) \\
&= \sum_{i=1}^{n}(y_i - \bar{y}) + b_1 \sum_{i=1}^{n}(x_i - \bar{x}) \\
&= \sum_{i=1}^{n} 0 + b_1 \sum_{i=1}^{n} 0 \\
&= 0
\end{aligned}
$$

But the sum of the residuals of any line through ($\bar{x}$, $\bar{y}$) is zero!

A Brief History of Regression

Francis Galton



FAMILY HEIGHTS. from R.F.F.

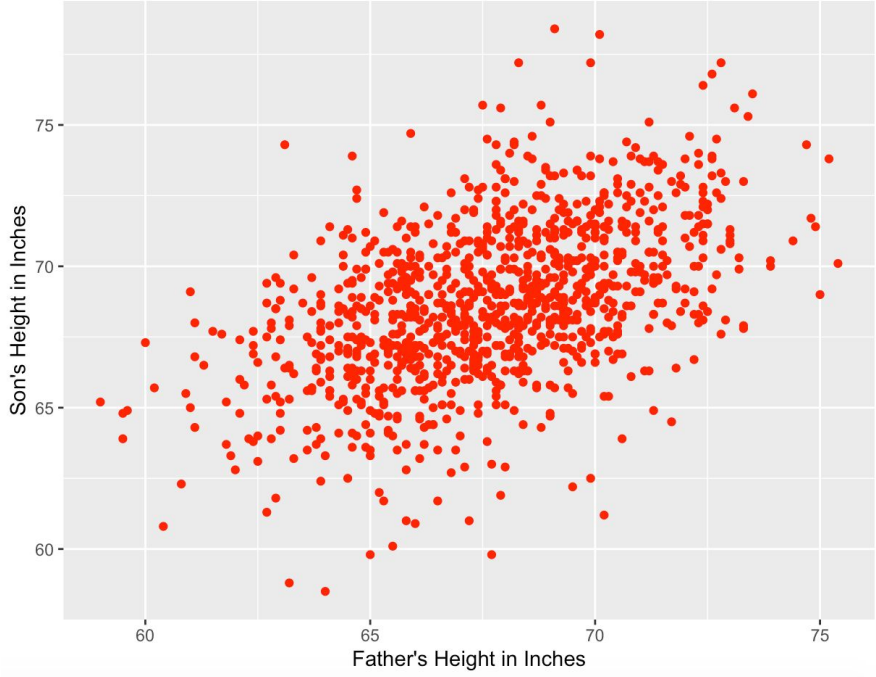
(add 60 inches to every entry in the Table)

	Father	Mother	Sons in order of height	Daughters in order of height.
1	18·5	7·0	13·2	9·2, 9·0, 9·0
2	15·5	6·5	13·5, 12·5	5·5, 5·5
3	15·0	about 4·0	11·0	8·0
4	15·0	4·0	10·5, 8·5	7·0, 4·5, 3·0
5	15·0	−1·5	12·0, 9·0, 8·0	6·5, 2·5, 2·5

Image source

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

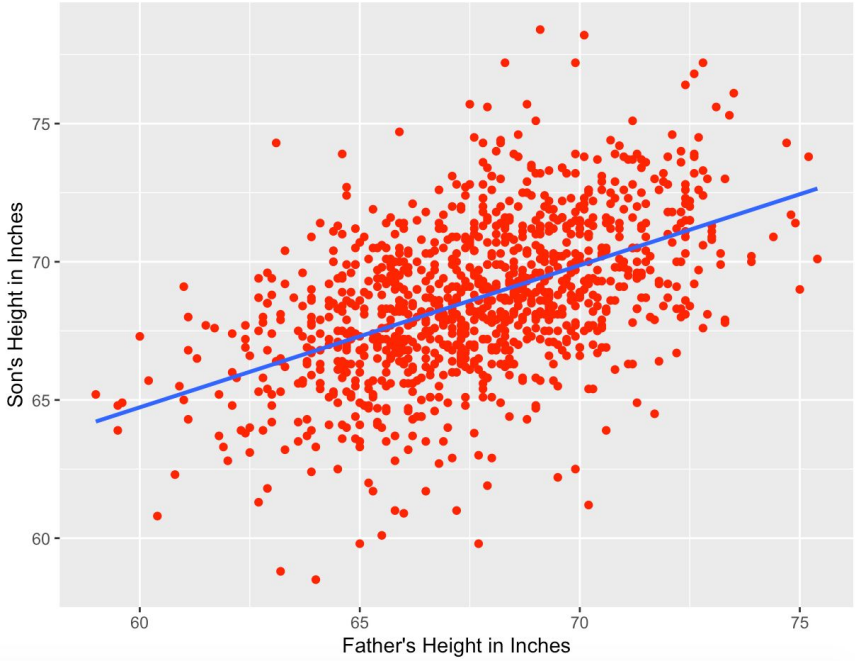


Fitting a Regression Line in R

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

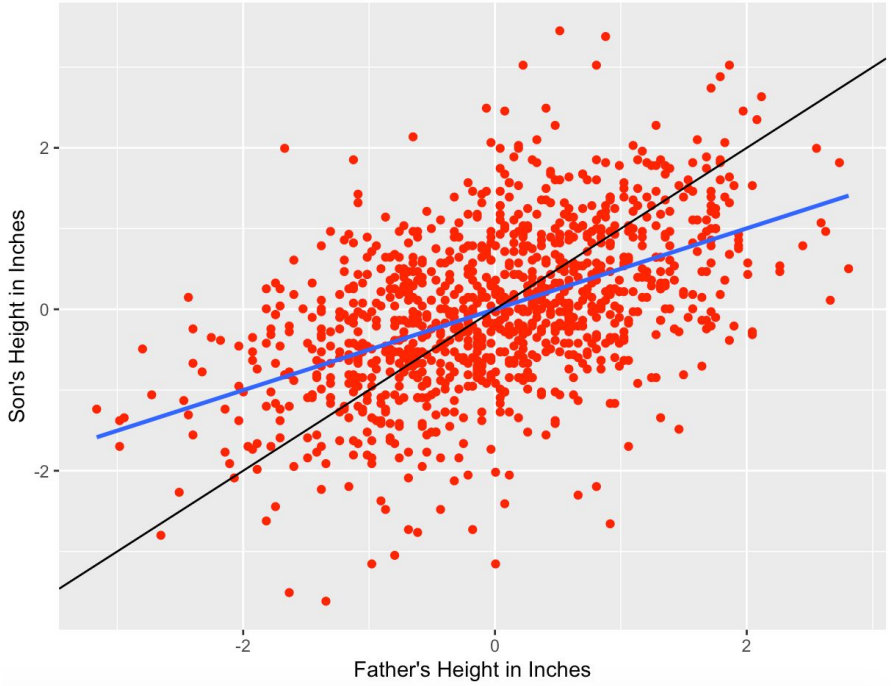
```
plot(Father, Son, col = "red")
fit <- lm(Son ~ Father)
abline(fit, col = "blue")
```

```
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.

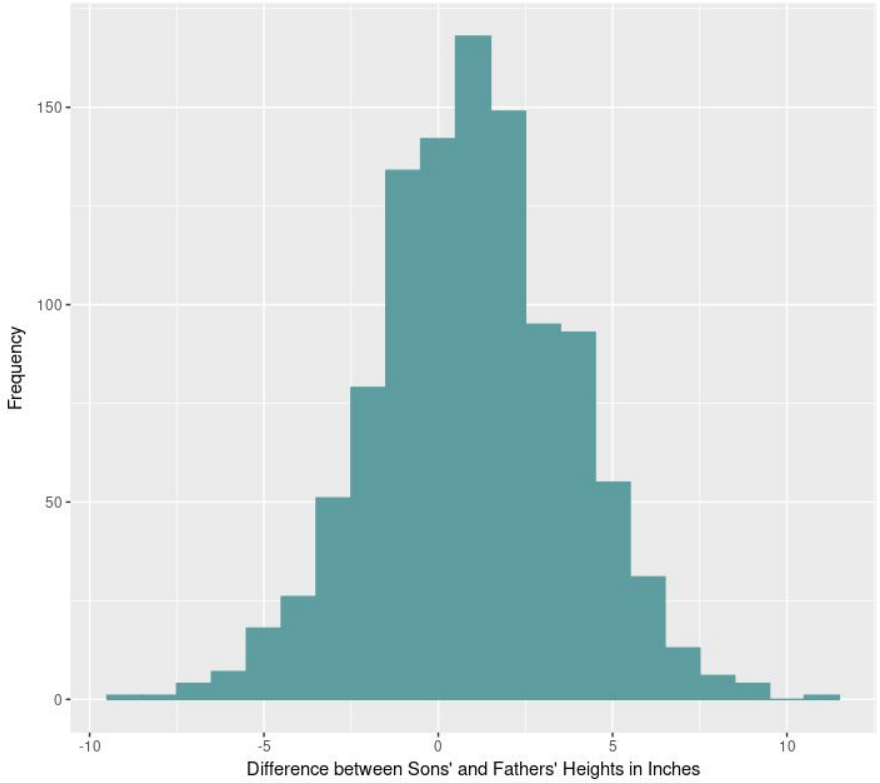


Histogram of the Differences

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

```
> summary(heights)
    Father           Son             Diff        
Min.   :59.00   Min.   :58.50   Min.   :-9.0000  
1st Qu.:65.80   1st Qu.:66.90   1st Qu.:-0.8000  
Median :67.80   Median :68.60   Median : 1.0000  
Mean   :67.69   Mean   :68.68   Mean   : 0.9974  
3rd Qu.:69.60   3rd Qu.:70.50   3rd Qu.: 2.7750  
Max.   :75.40   Max.   :78.40   Max.   :11.2000  
```

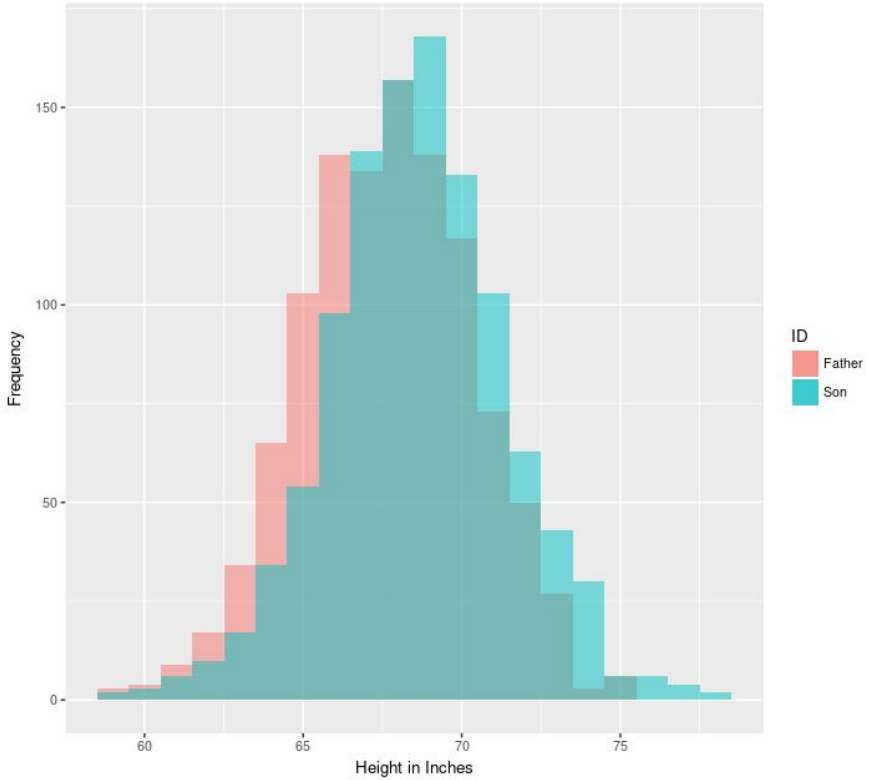
Sons are about an inch taller than their fathers, on average.



Histograms of their Heights

- The histograms of the fathers' and sons' heights are both bell-shaped.
- The histograms mostly overlap.
- Again, sons are about an inch taller than their fathers, on average.

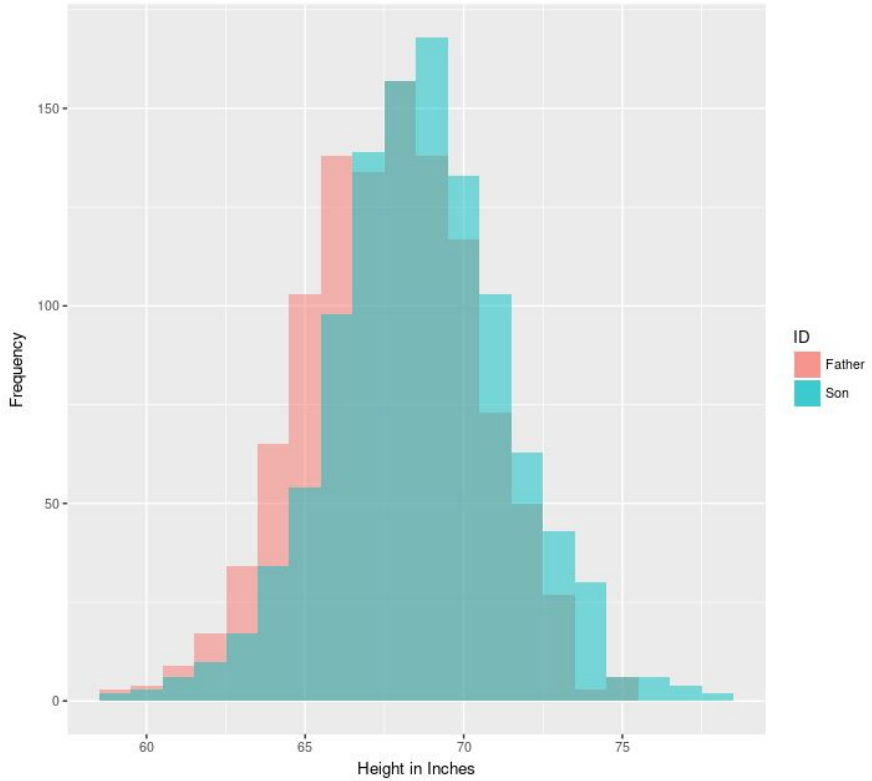
```
> summary(heights)
     Father           Son
 Min.   :59.00   Min.   :58.50
 1st Qu.:65.80   1st Qu.:66.90
 Median :67.80   Median :68.60
 Mean   :67.69   Mean   :68.68
 3rd Qu.:69.60   3rd Qu.:70.50
 Max.   :75.40   Max.   :78.40
```



Correlation in their Heights

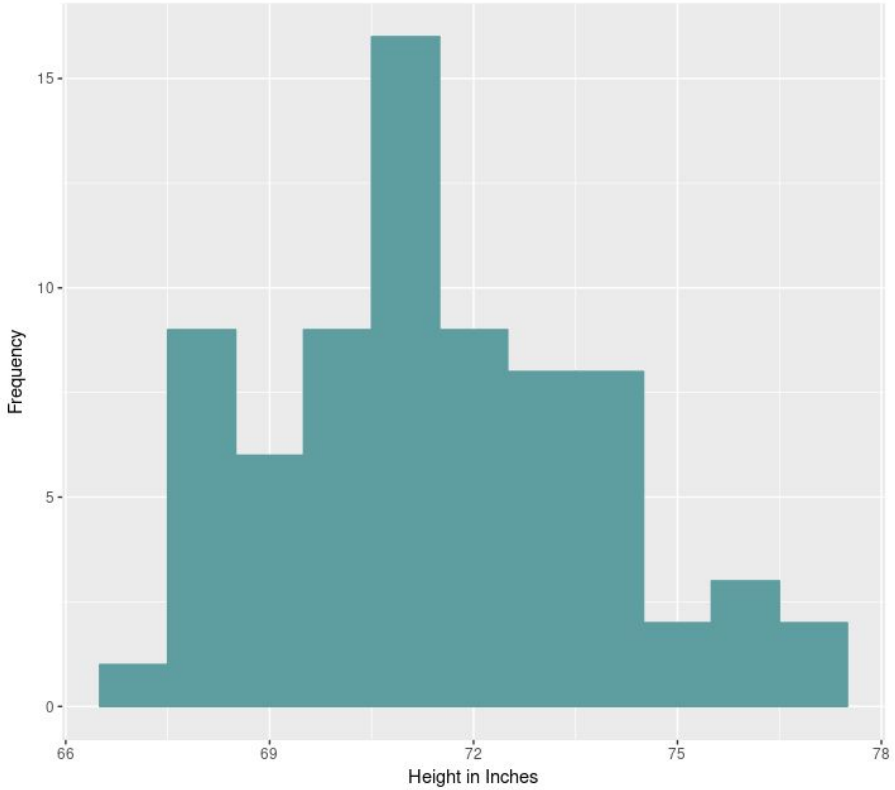
The correlation in their heights is exactly what leads to the American football (i.e., ellipsoidal) shape

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



The Regression Effect

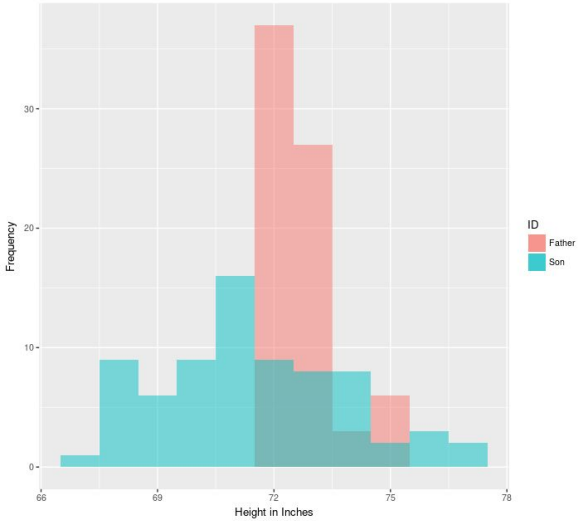
- We might expect the sons of tall fathers to be tall as well.
- This histogram shows the heights of sons of 72 inch fathers.
- Most (68%) of these sons are less than 72 inches tall!



The Regression Effect (cont'd)

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

```
> 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". Today, this is called the regression effect.
- Galton also noticed that short fathers had sons who were somewhat taller than their generation on average.
- Individuals who are below (or above) 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 (no measurement error), 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 randomness.
- In the post-test, if they are truly below (or above) the mean, they will likely measure that way again.
- But if their pre-test measurements were 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!

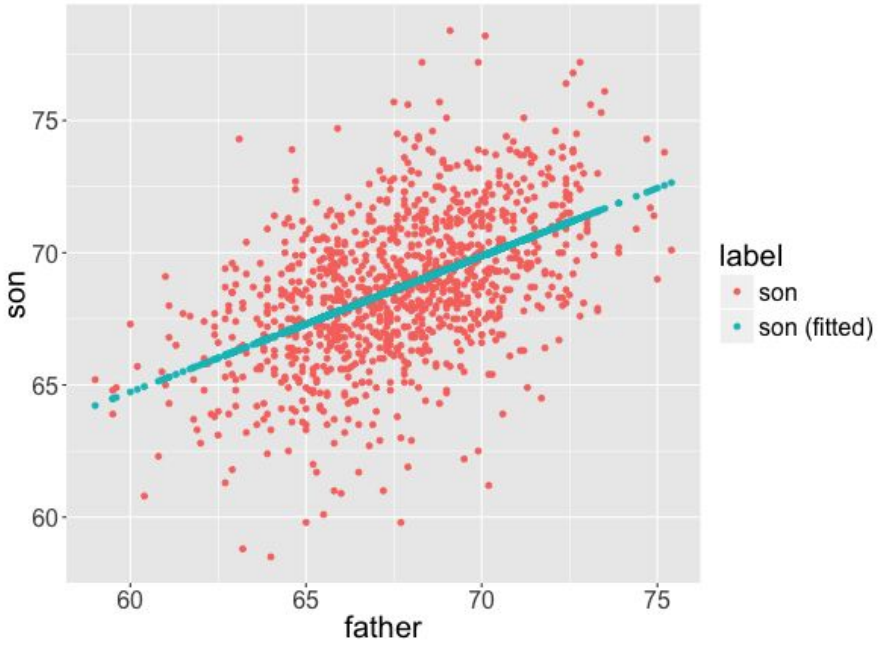
Extras

Interpreting the Regression Line

In inches: $y = a + bx$

- The intercept $a = 33.89$
- The slope is $b = 0.514$

Each point on the regression line is the result of multiplying a father's height in inches by b and adding a



Interpreting the Regression Line (cont'd)

In standard units: $y = rx$

- The intercept is 0
- The slope is r

Each point on the regression line is the result of multiplying a father's height in standard units by r