

# Intro to Regression

Monday, September 5, 2016 7:03 PM

- Questions (posed in an example situation):
  - Use parents' heights to predict children's heights
  - Mean relationship between parent and children's heights
  - Variation in children's heights that appears unrelated to parent's height
  - Quantify genotypic effect
  - **General Q: What conclusions can we draw from one small dataset about other related data**

# Basic Least Squares

Tuesday, September 6, 2016 9:37 AM

- Find the middle of histogram representation
- What is the middle?

- $\sum_{i=1}^n (Y_i - \mu)^2$

- "Physical" center of mass of histogram
- Can experiment with R studio's manipulate library to make a slider to shift mu

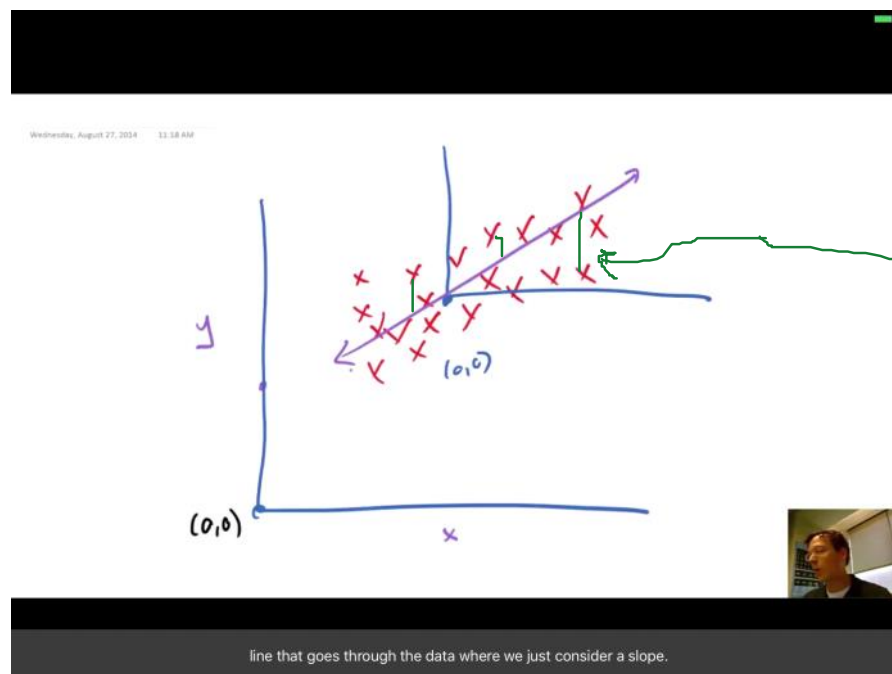
# Introductory Data

Tuesday, September 6, 2016 9:53 AM

## Regression through the origin

- Suppose that  $X_i$  are the parents' heights.
- Consider picking the slope  $\beta$  that minimizes
$$\sum_{i=1}^n (Y_i - X_i \beta)^2$$
- This is exactly using the origin as a pivot point picking the line that minimizes the sum of the squared vertical distances of the points to the line
- Use R studio's manipulate function to experiment
- Subtract the means so that the origin is the mean of the parent and children's heights

let's force the line to go through the origin.



- Regression through the origin doesn't mean setting from the origin; rather, you place the axes in the middle of the data so makes the origin make sense

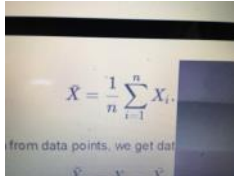
# Ordinary Least Squares

Tuesday, September 6, 2016 10:04 AM

- Workhorse of statistics
- Complicated outcomes and explain behavior (such as trends) via linearity
- Simplest application of OLS is fitting a line through some data

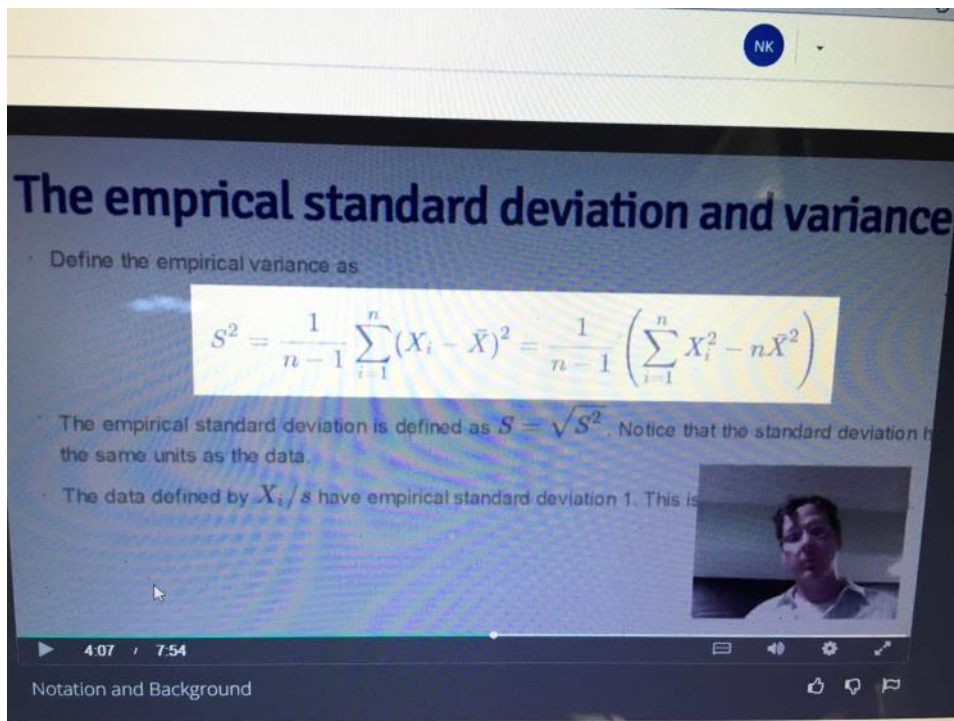
## Background and Notation

- Empirical mean =



$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

from data points, we get data



**The empirical standard deviation and variance**

Define the empirical variance as:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1} \left( \sum_{i=1}^n X_i^2 - n\bar{X}^2 \right)$$

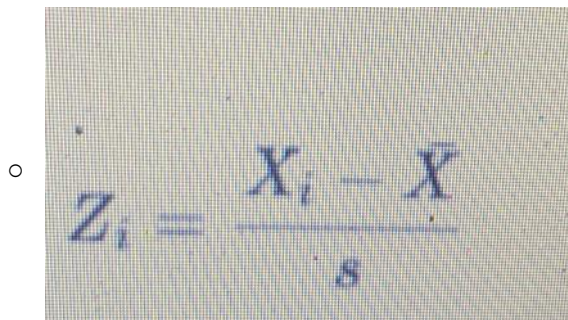
The empirical standard deviation is defined as  $S = \sqrt{S^2}$ . Notice that the standard deviation has the same units as the data.

The data defined by  $X_i/s$  have empirical standard deviation 1. This is

4:07 / 7:54

Notation and Background

- Normalization:



$$Z_i = \frac{X_i - \bar{X}}{s}$$

- Empirical Covariance
  - Most central value

Consider now when we have pairs of data,  $(X_i, Y_i)$ .

Their empirical covariance is

$$Cov(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{n-1} \left( \sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y} \right)$$

○ The correlation is defined is

$$Cor(X, Y) = \frac{Cov(X, Y)}{S_x S_y}$$

where  $S_x$  and  $S_y$  are the estimates of standard deviations for the  $X$  observations and observations, respectively.

**The correlation is simply**

- Facts about correlations
  - Between -1 and 1
  - = 1 or -1 only if pairs of data are on exact either positively or negatively sloped line
  - Stronger correlation = greater abs(cor)

# Linear Least Squares

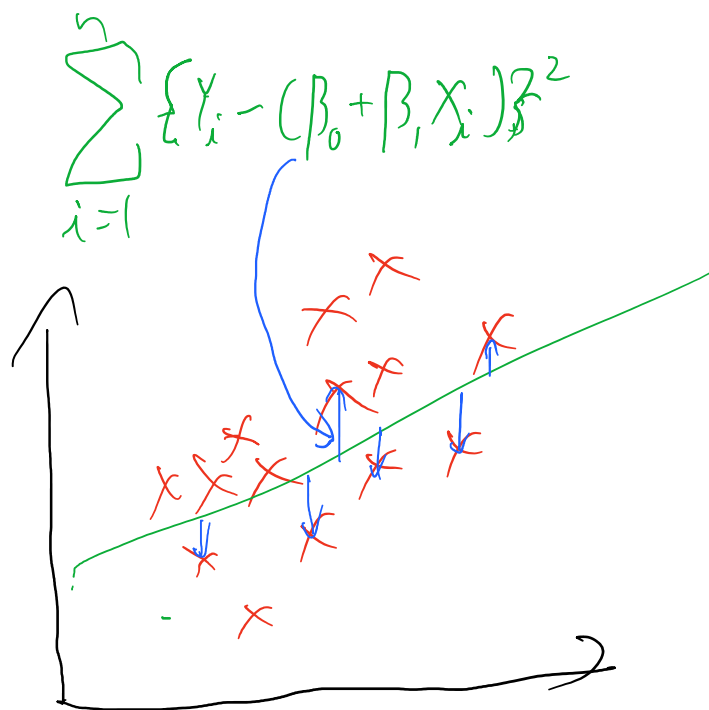
Tuesday, September 6, 2016

10:57 AM

- $Y_i = i^{\text{th}}$  child's height,  $X_i = i^{\text{th}}$  average of pair of parents' heights
- Child's Height  $= \beta_0 + (\text{Parent's Height} * \beta_1)$

Best Fit Line

Least Squares: minimize squared vertical distance between points and line.



$$\hat{\beta}_1 = \text{Cor}(Y, X) \frac{SD(Y)}{SD(X)}$$

units  $\frac{Y}{X}$  ←

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

units  $Y$  ←

$$\text{Units } Y \leftarrow \beta_0 + 1 - \beta_1$$

Line passes through point  $(\bar{X}, \bar{Y})$

Slope is same if you centered data and did regression through origin.

# Regression to the Mean

Tuesday, September 6, 2016 11:13 AM

• Things  $\uparrow$  one time becomes  $\downarrow$  next period

$\hookrightarrow P(Y < x | X = x) \uparrow$  as  $x \uparrow$

$P(Y > x | X = x) \downarrow$  as  $x \downarrow$



# Statistical Linear Regression Models

Tuesday, September 6, 2016

11:23 AM

- Making statistical models for a population

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \rightarrow \sim N(0, \sigma^2)$$

$\uparrow \quad \uparrow \quad \uparrow$   
intercept slope Gaussian Error

$$E[Y_i | X_i = x_i] = \mu_i = \beta_0 + \beta_1 x_i$$

$$\text{Var}(Y_i | X_i = x_i) = \sigma^2$$

---

Interpreting Coefficients

$\beta_0 = \text{expected } Y_i \text{ when } X_i = 0$

$\beta_1 = \text{expected change } Y_i \text{ for 1 unit of } X_i$

# Residuals

Tuesday, September 6, 2016 4:02 PM

- Residual variation after models have been designed
- Expected residual value is 0
- If you sum residuals multiplied by data, sum is 0
- Residuals can be thought of as outcome Y with linear prediction removed
- resid() function in R
- **Residual Variation**
  - Variation not explained by standard variation from model

# Inference in Regression

Tuesday, September 6, 2016 4:22 PM

- Process of drawing conclusions about a population using a sample
- Prediction
  - Predicting  $Y$  at  $X$  is