# 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**

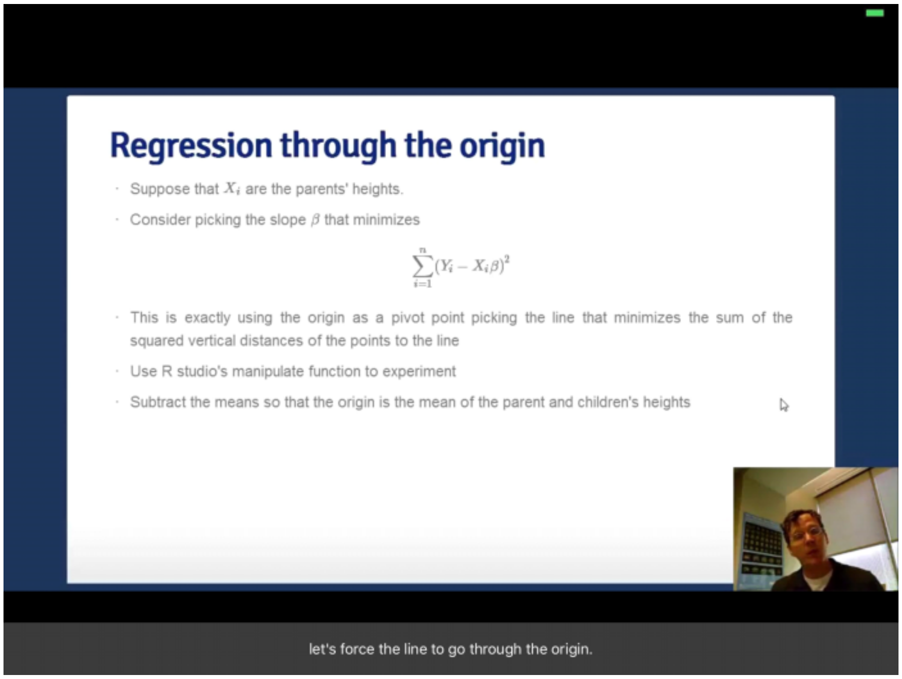
# Basic Least Squares

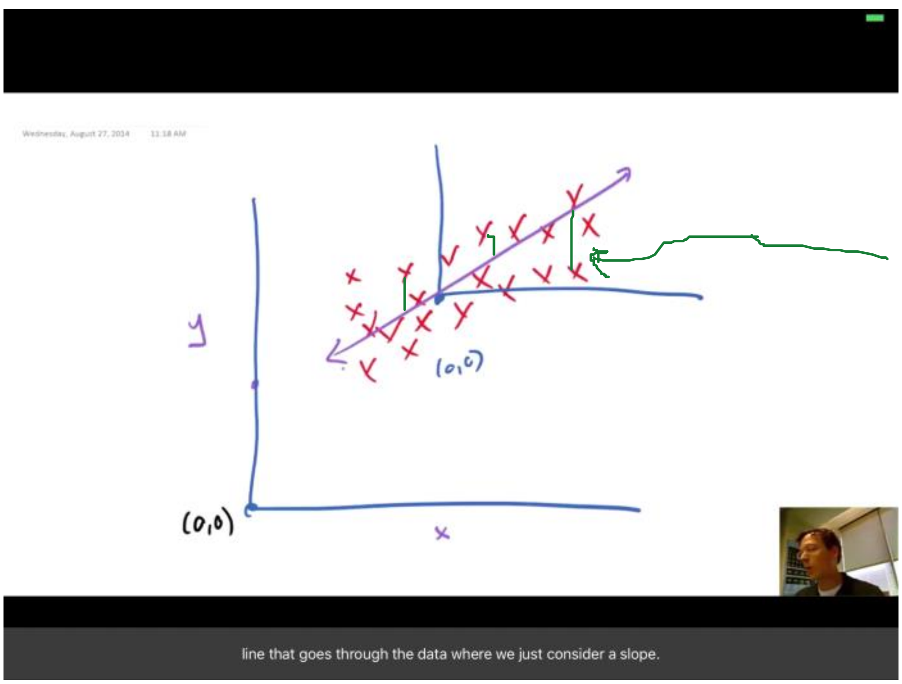
Tuesday, September 6, 2016 9:37 AM

- 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

- 

- 

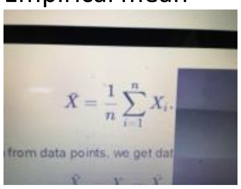
Vertical Distance between point and line
- Regression through the origin doesn't mean setting from the origin; rather, you place the axes in the middle of the data so makes the origin make sense

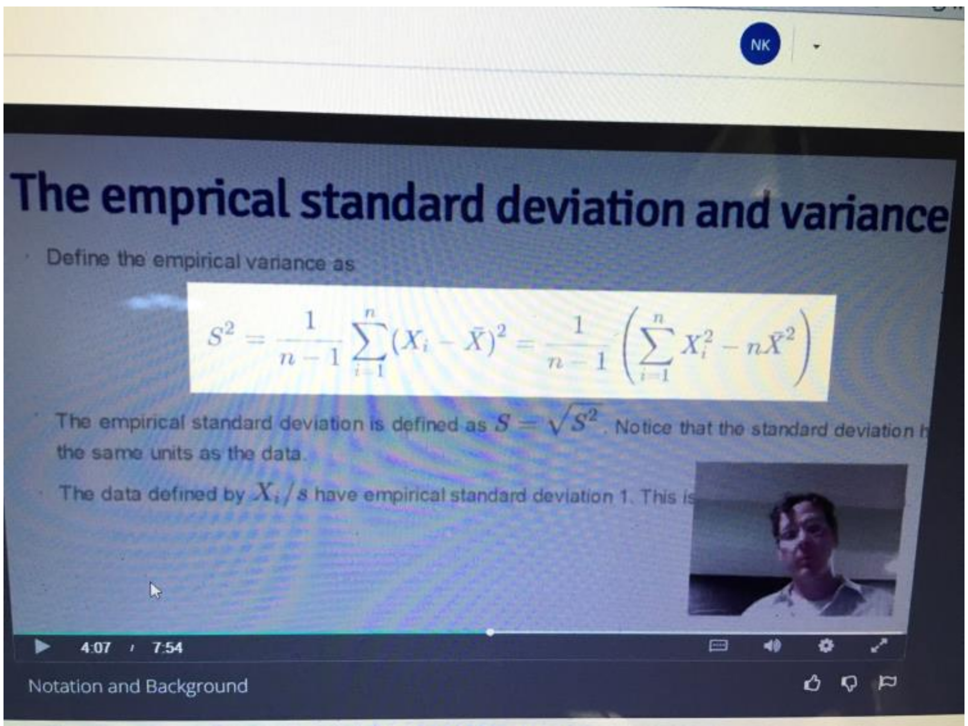
# Ordinary Least Squares

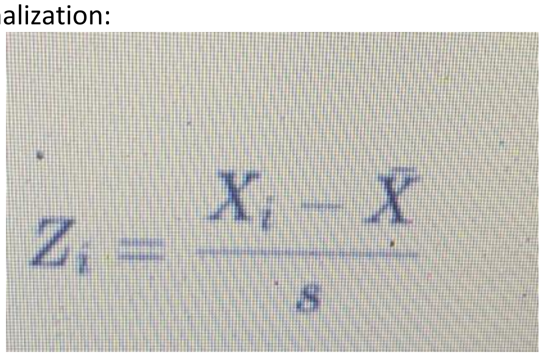
Tuesday, September 6, 2016 10:04 AM

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

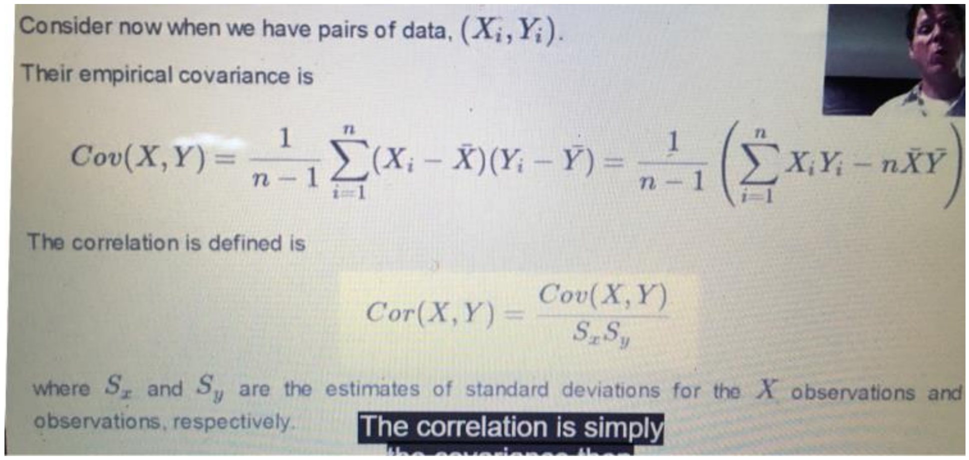
**Background and Notation**

- Empirical mean =
- 

- 

- Normalization:
    - 

- Empirical Covariance
    - Most central value

- 


- 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

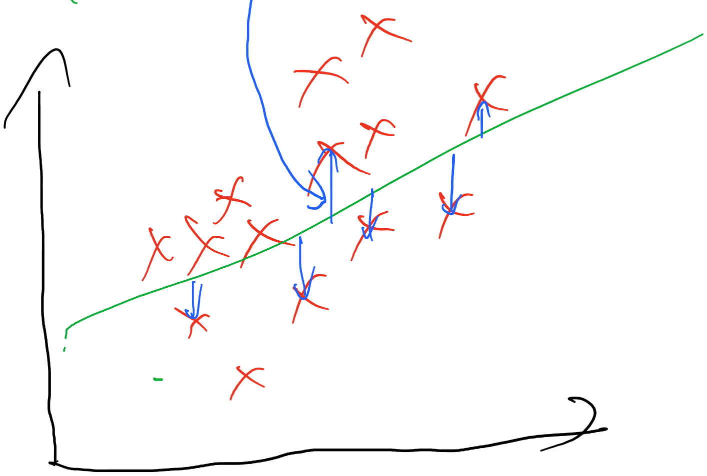
Tuesday, September 6, 2016 10:57 AM

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

Best Fit Line

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

$$\sum_{i=1}^{n} \{Y_i - (\beta_0 + \beta_1 X_i)\}^2$$



units $\frac{Y}{X}$ $\leftarrow \hat{\beta}_1 = Cor(Y, X)\frac{Sd(Y)}{Sd(X)}$

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

Units
$Y \leftarrow \beta_0$   $\leftarrow \beta_1$

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
  - ↳ $P(Y < x \mid X = x) \uparrow$ as $x \uparrow$
    $P(Y > x \mid X = x) \downarrow$ as $x \downarrow$

# Statistical Linear Regression Models

Tuesday, September 6, 2016 11:23 AM

- Making statistical models for a population

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

intercept Slope Gaussian Error

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

$$Var(Y_i \mid X_i = x_i) = \sigma^2$$

---

Interpreting Coefficients

$\beta_0$ = expected $Y_i$ when $X_i = 0$

$\beta_1$ = expected change $Y_i$ 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