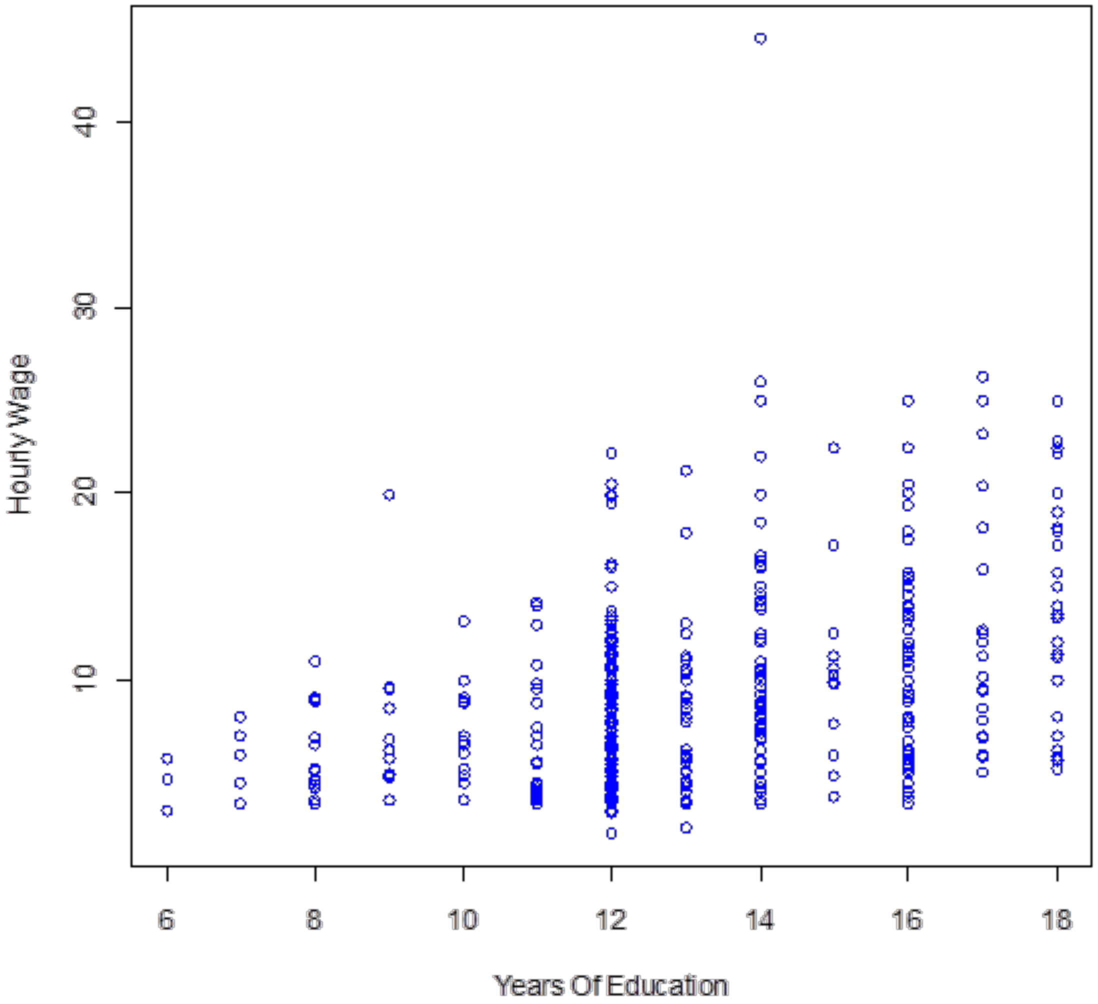
EPI 809
(Biostatistics II)

Simple Linear Regression



Key Concepts we will discuss

- Conditional Expectation (expected value of Y give X)
- Variance & Co-variance
- Correlation
- Linear regression (linear approx. to the conditional expectation)

Regression

Main Question: How Does Y change as X does?

$$Y \quad = \quad f(X) \quad + \quad E$$

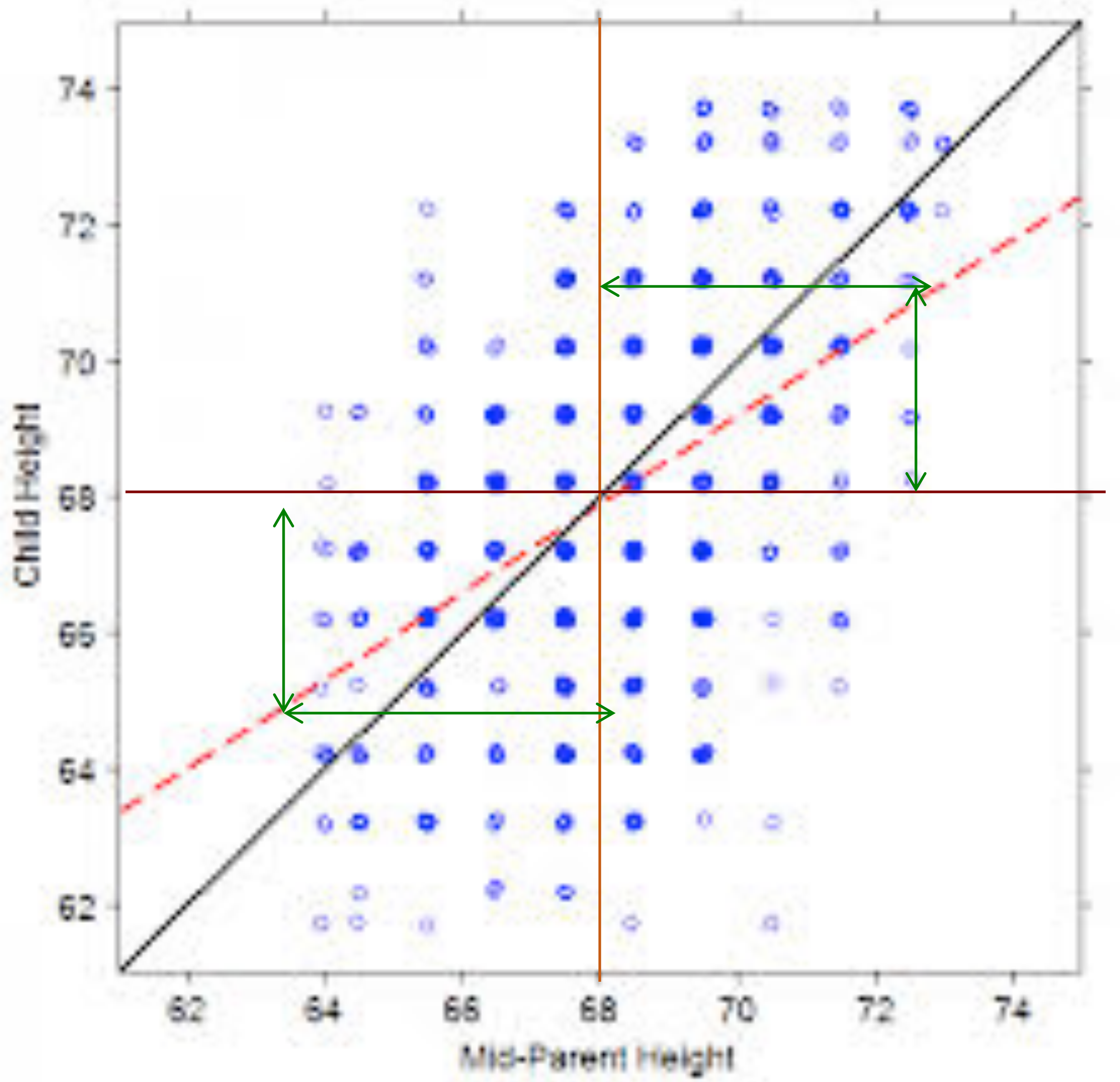
Response

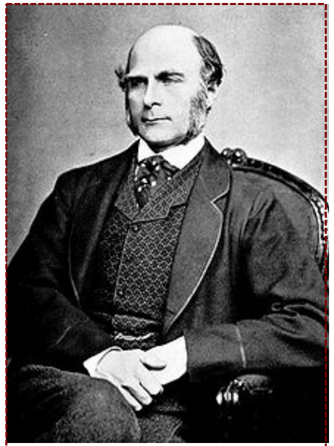
Conditional Expectation
(Expected Value of Y given X)

Error Term

A Bit of History…Regression toward the mean…

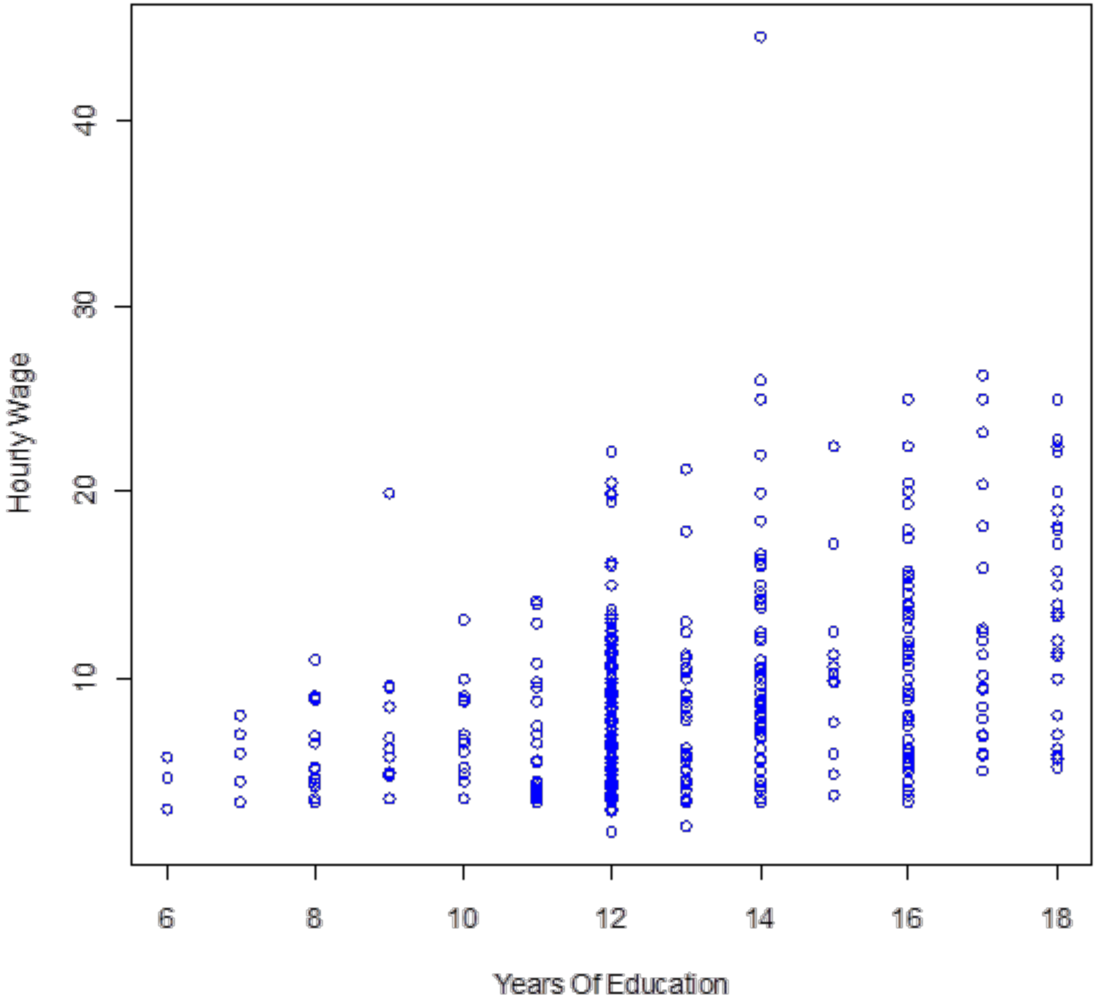
http://www.amstat.org/publications/jse/v9n3/stanton.html



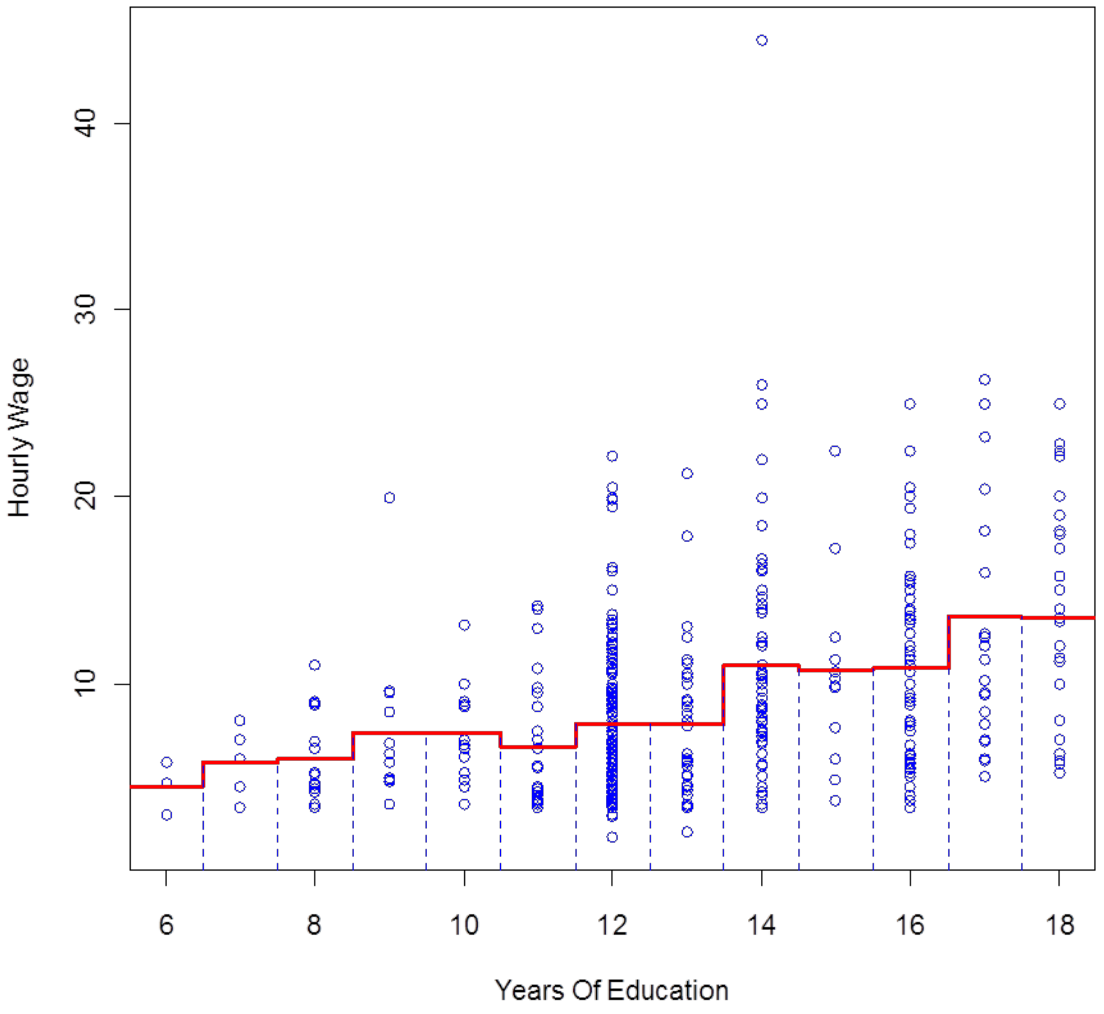


Sir Francis Galton

Estimating A Conditional Expectation Function



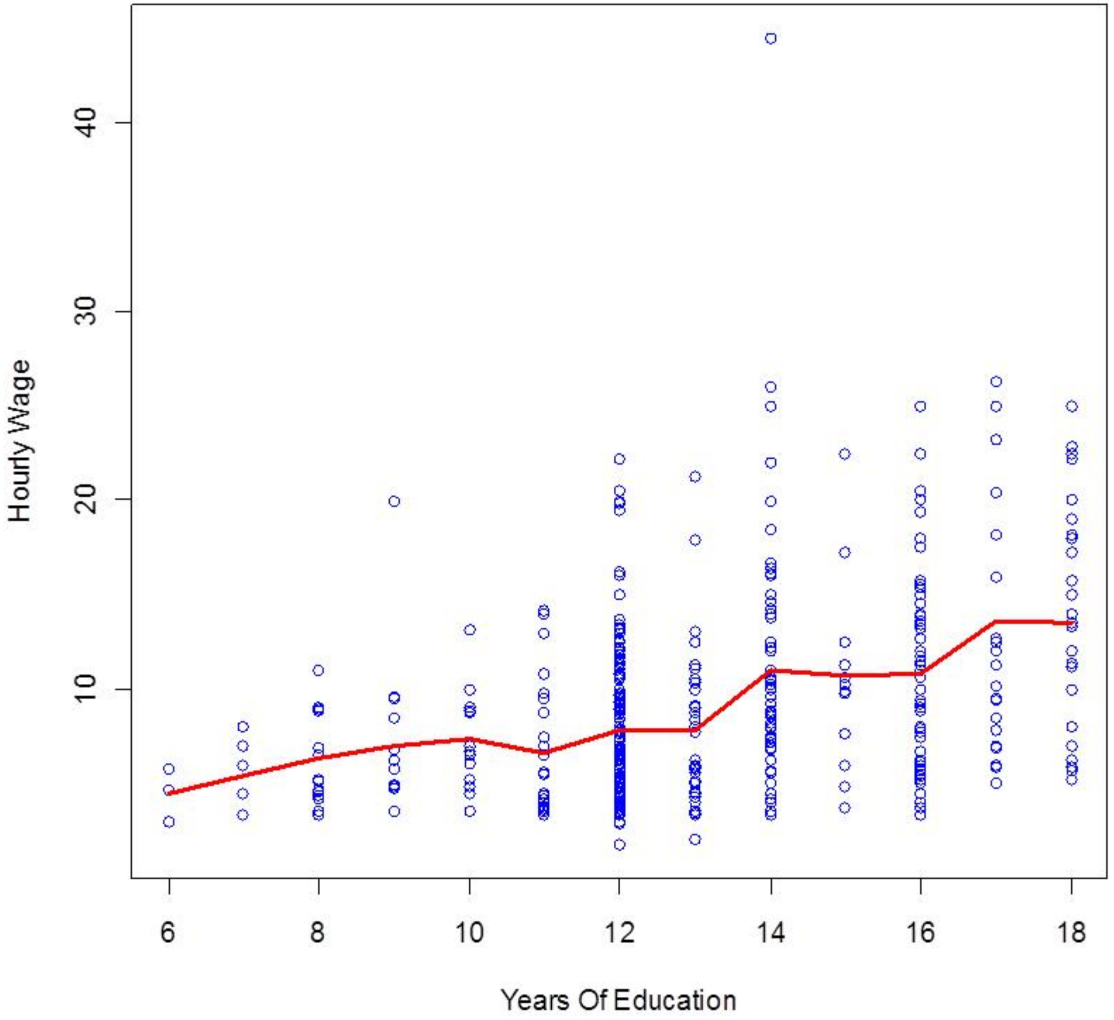
Estimating A Conditional Expectation Function



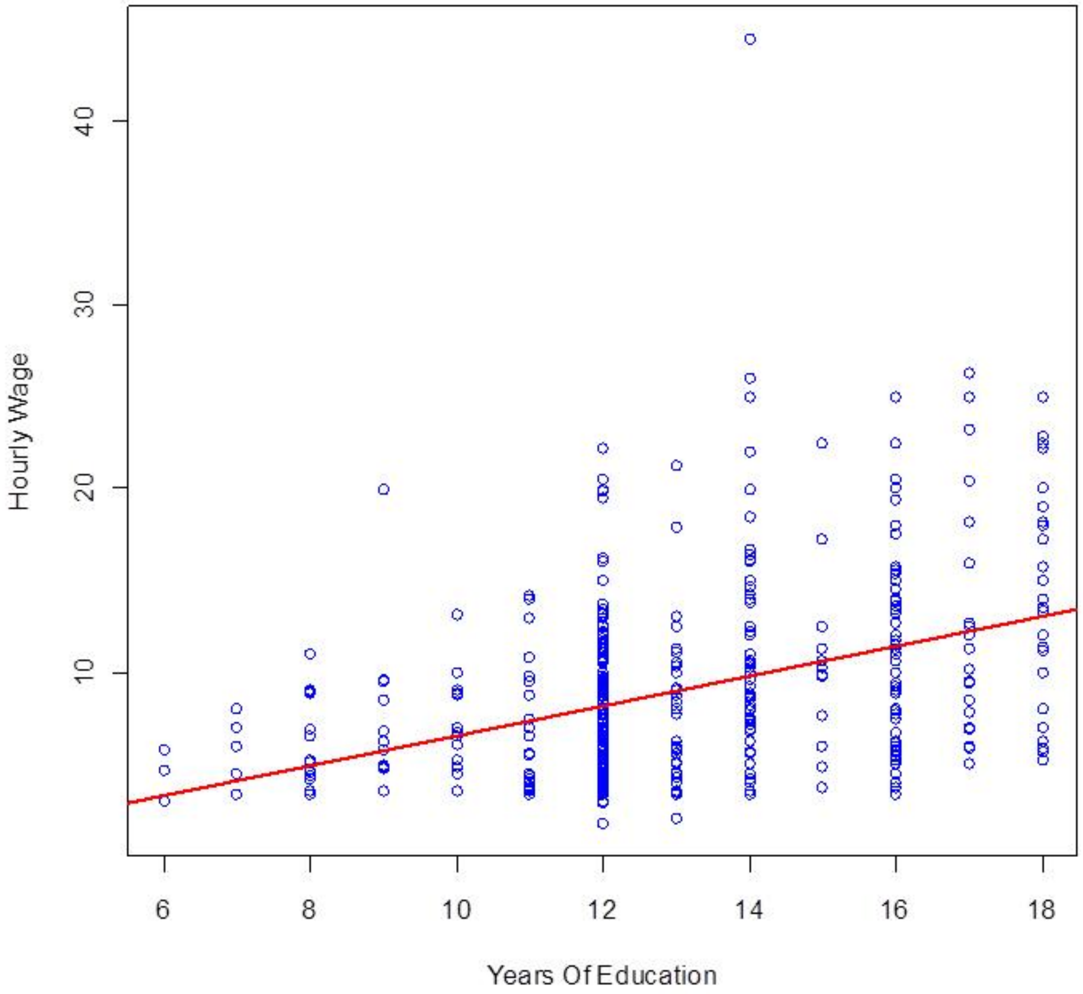
Estimating the Cond. Expectation Using By Windows

- Define ‘windows’ for the X-variable.
- For each window estimate the mean value of Y for individuals with value of X falling within the window.

Estimating A Conditional Expectation Function



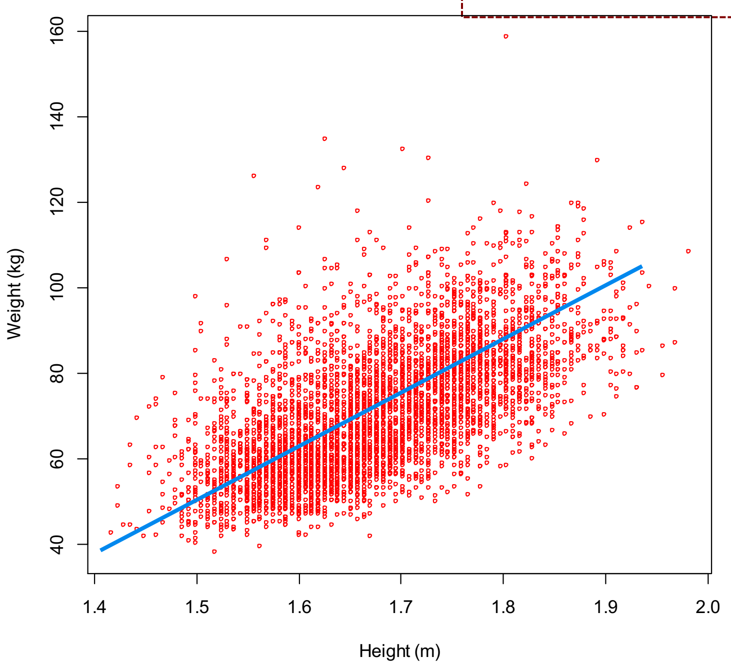
Estimating A Conditional Expectation Function



How do we estimate the line (a+bX) that fits the data best?

Simple Linear Regression

Simple Linear Regression



Model: $$y_i = \beta_0 + x_i\beta_1 + \varepsilon_i$$

Response

Intercept

Predictor

Regression Coef.

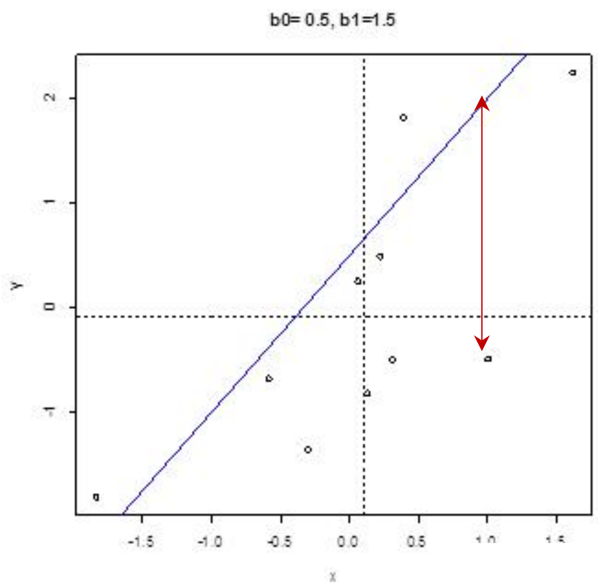
Residual

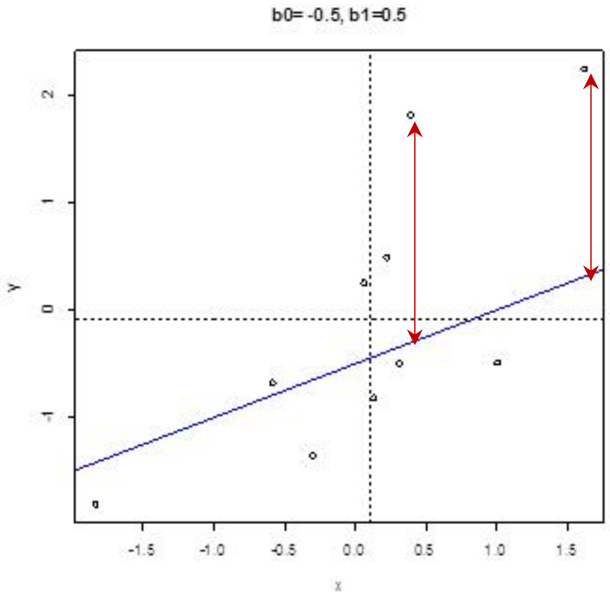
Interpretation:

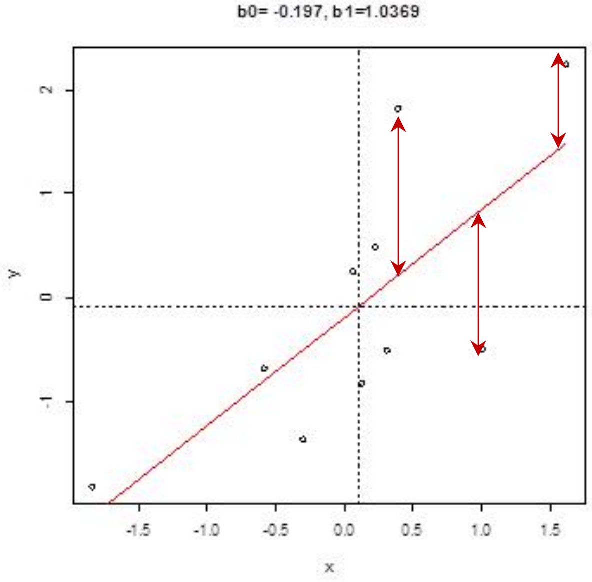
=> β_1: Linear rate of change of y with respect to X

=> $\beta_0 + x_i\beta_1$: Prediction equation (linear approx. to the conditional expectation)

Estimation Via Ordinary Least Squares







Estimation Via Ordinary Least Squares

Problem: Find the values of and of that minimize the residual sum of squares (OLS=Ordinary Least Squares).

$$\varepsilon_i = \left(y_i - \beta_0 - x_i\beta_1\right)$$

$$RSS\left(\beta_0, \beta_1, X, Y\right) = \sum_{i=1}^{n}\left(y_i - \beta_0 - x_i\beta_1\right)^2$$

$$\left(\hat{\beta}_0, \hat{\beta}_1\right) \underset{argmin}{=} \left\{ \sum_{i=1}^{n}\left(y_i - \beta_0 - x_i\beta_1\right)^2 \right\}$$

1st Order Conditions

$$\frac{\partial RSS}{\partial \beta_0} = \frac{\partial \sum_{i=1}^{n}\left(y_i - \beta_0 - x_i\beta_1\right)^2}{\partial \beta_0} = 2\sum^{n}\left(y_i - \beta_0 - x_i\beta_1\right)$$

Residuals add-up to zero

$$\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - x_i\hat{\beta}_1\right) = 0$$

$$\frac{\partial RSS}{\partial \beta_1} = \frac{\partial \sum_{i=1}^{n}\left(y_i - \beta_0 - x_i\beta_1\right)^2}{\partial \beta_1}$$

Cov. Residuals and X=0

$$\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - x_i\hat{\beta}_1\right)x_i = 0$$

Estimation Via Ordinary Least Squares

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{S_{XY}}{S_{XX}}$$

Covariance(X,Y)

Variance(X)

$$\hat{\beta}_0 = \overline{y} - \overline{x}\hat{\beta}_1$$

Excel-file (Galton.xls) + INCLASS