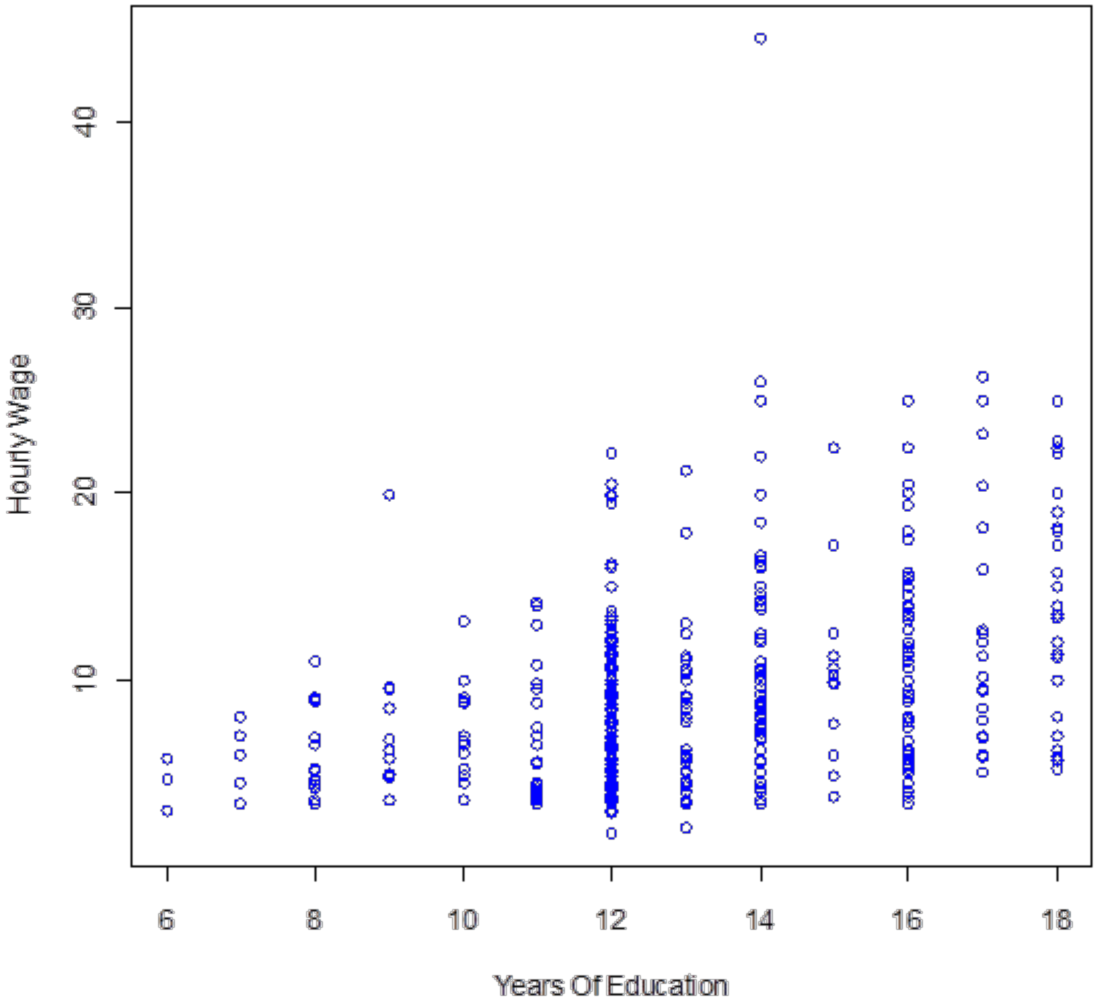
EPI 809
(Biostatistics II)

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?

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

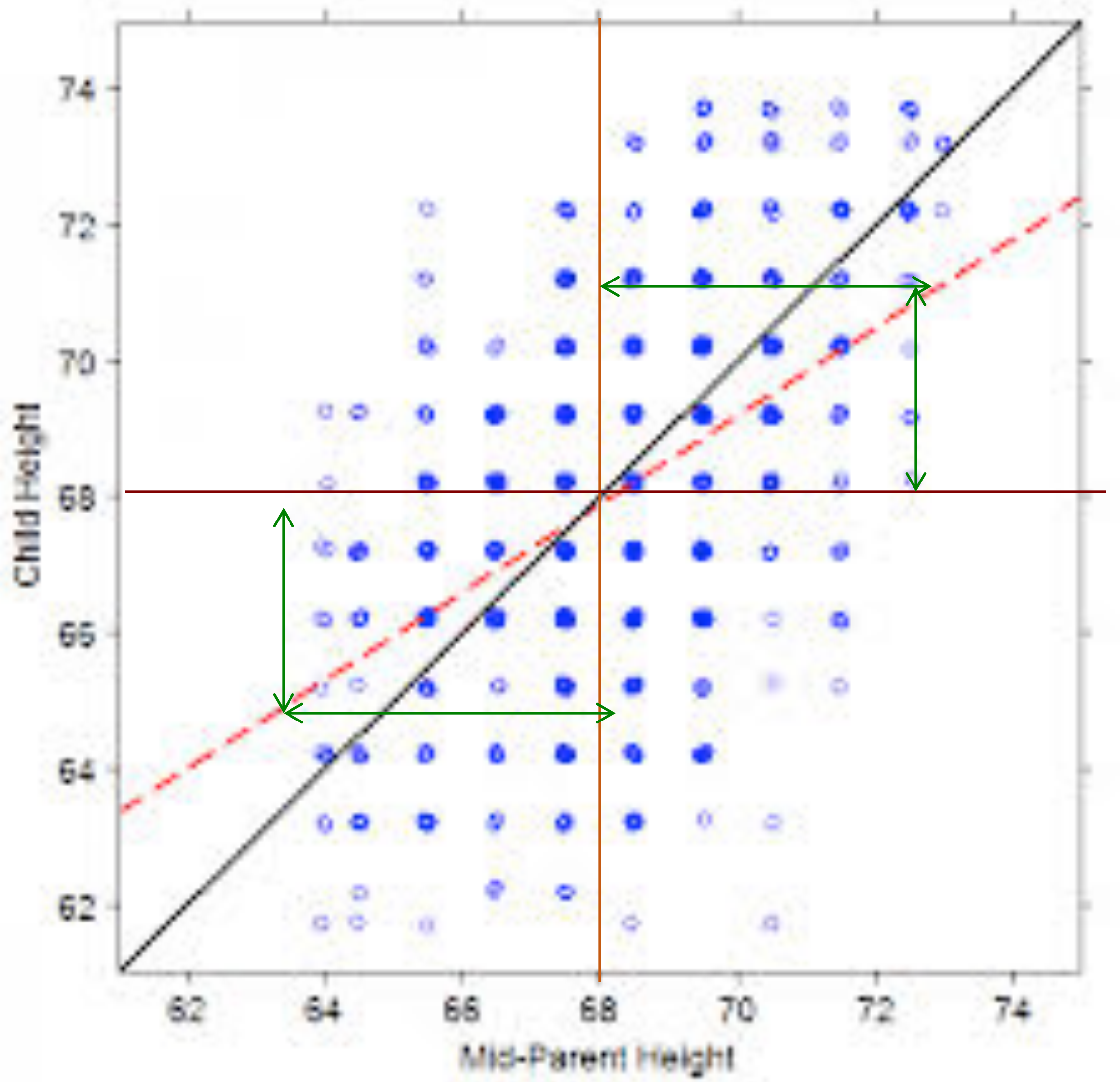
Response

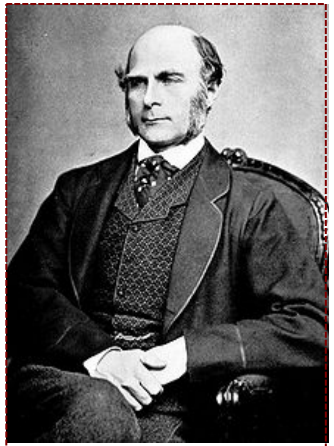
Conditional Expectation
(Expected Value of Y given X)

Error Term

A Bit of History…Regression toward the mean…

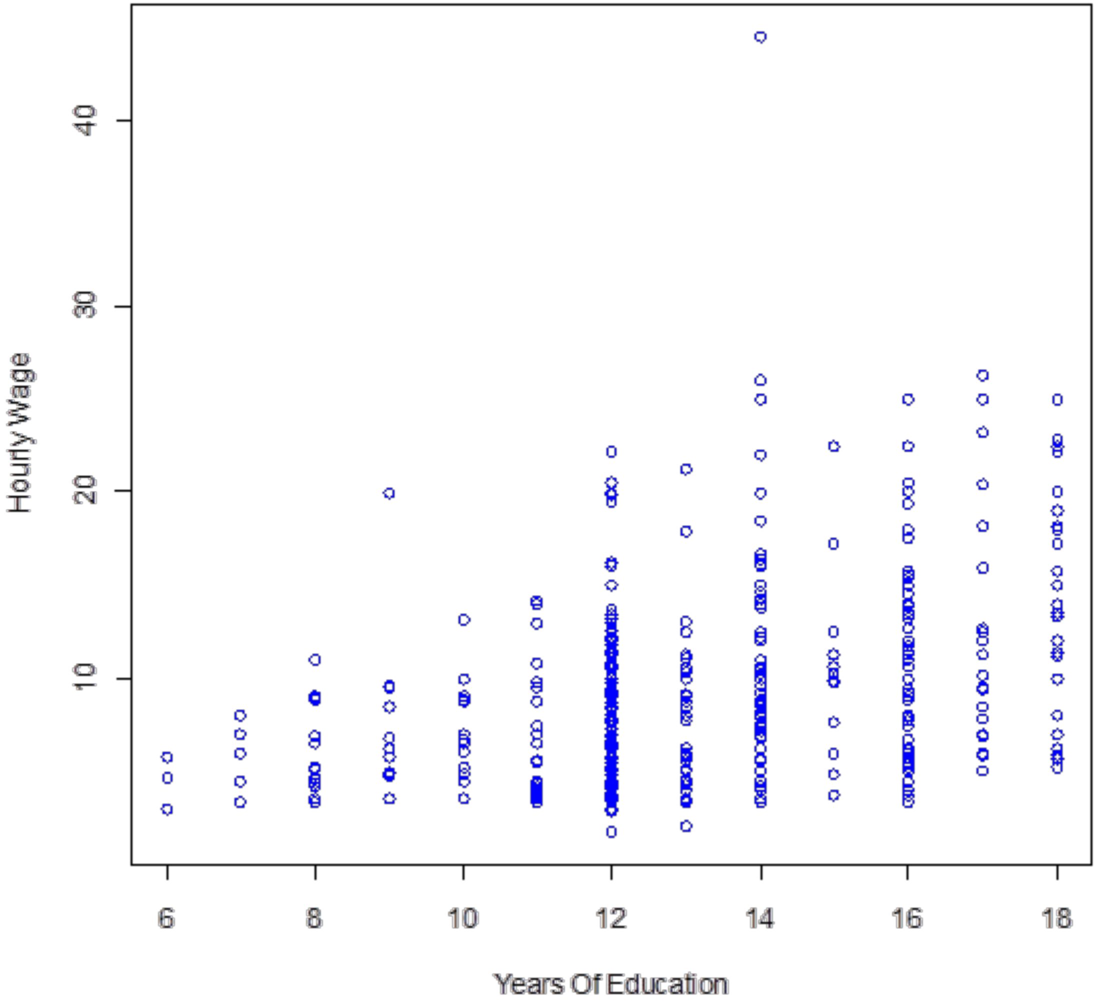
http://www.amstat.org/publications/jse/v9n3/stanton.html



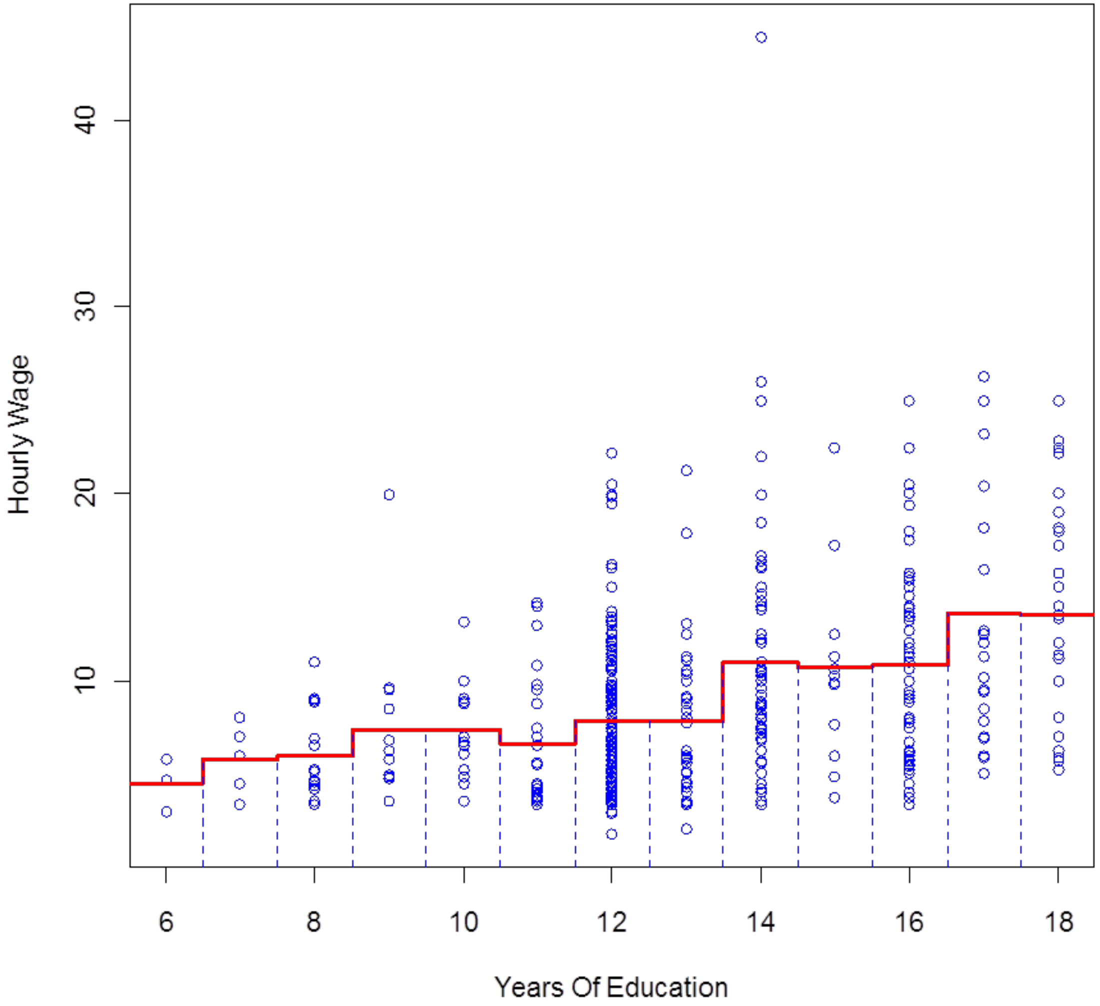


Sir Francis Galton

Estimating A Conditional Expectation Function



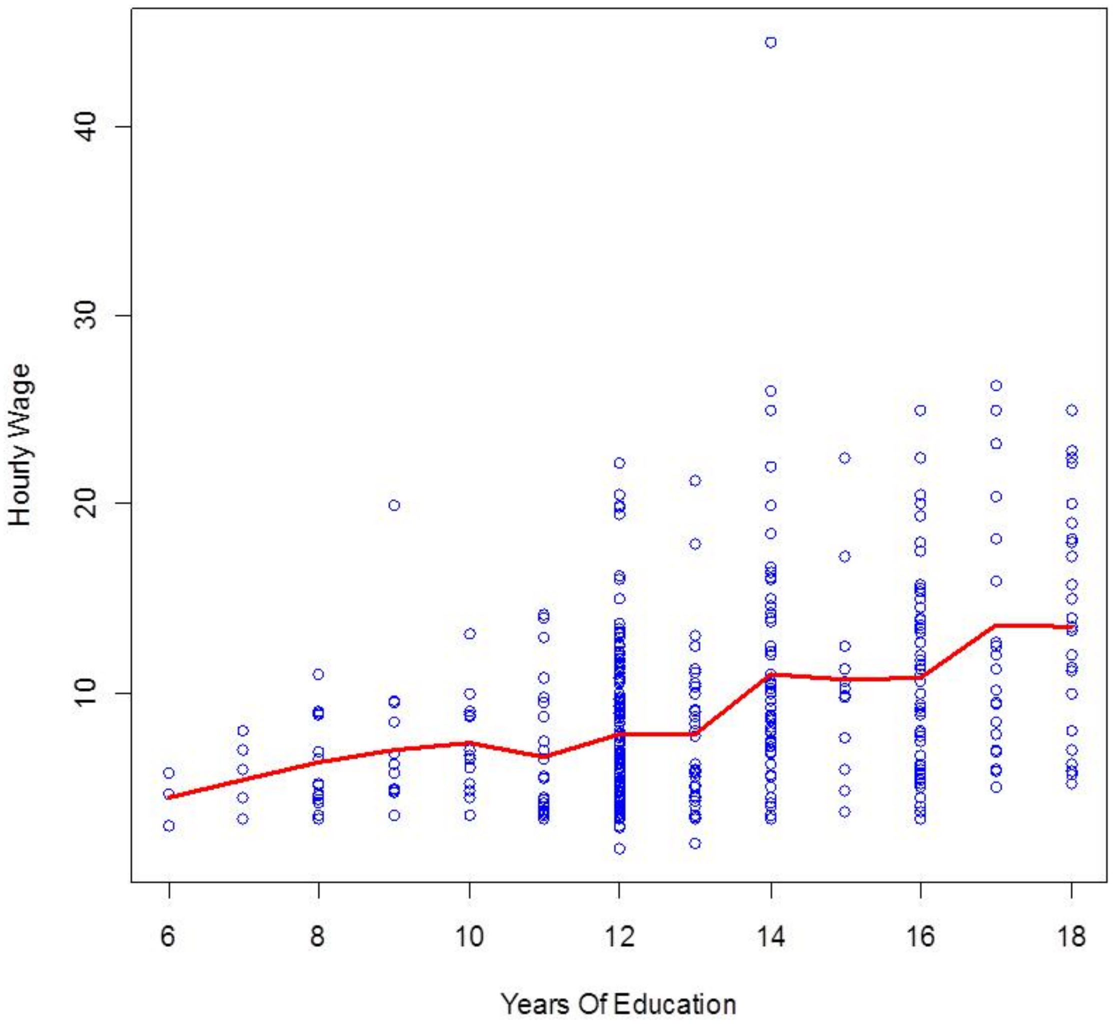
Estimating A Conditional Expectation Function



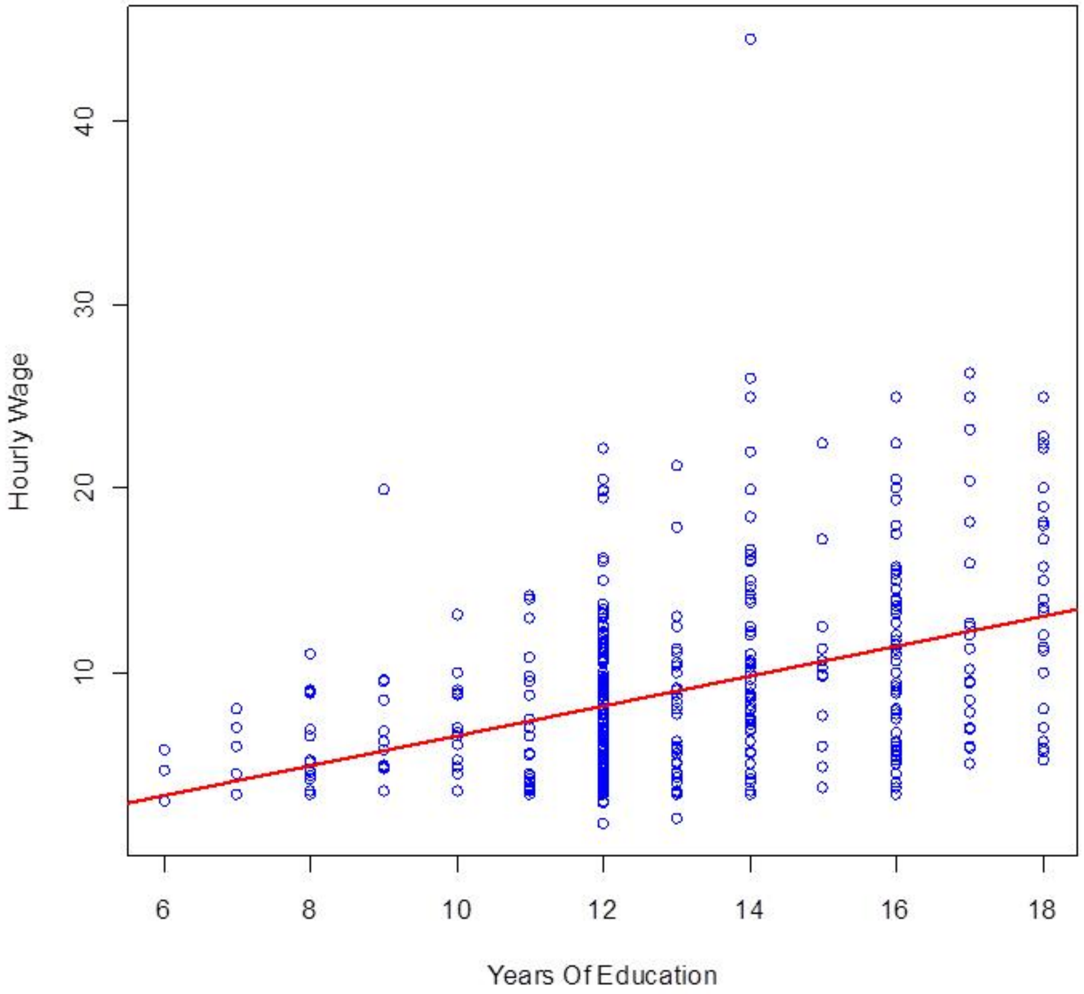
Estimating the Cond. Expectation Using By Windows

- 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



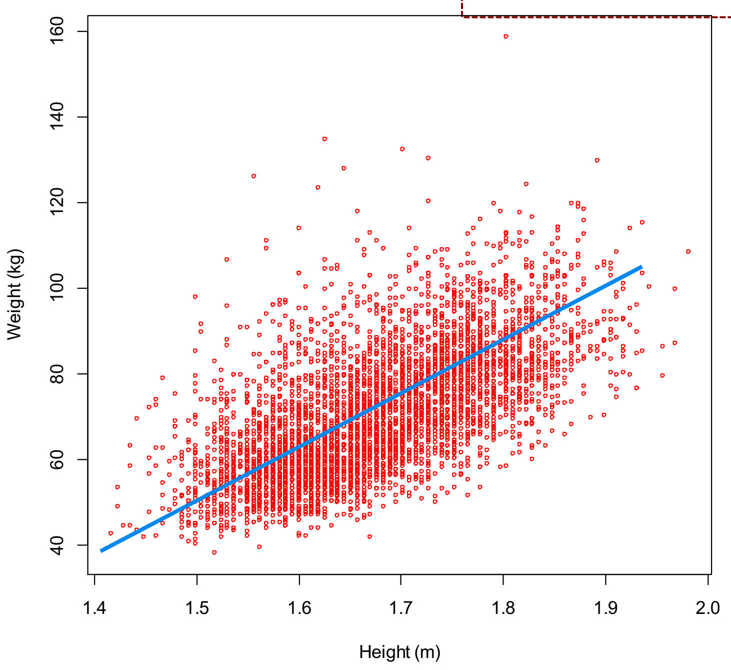
Estimating A Conditional Expectation Function



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.

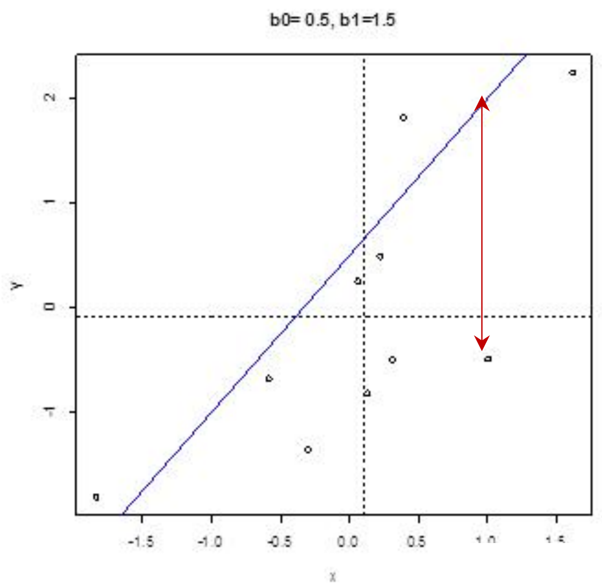
Residual

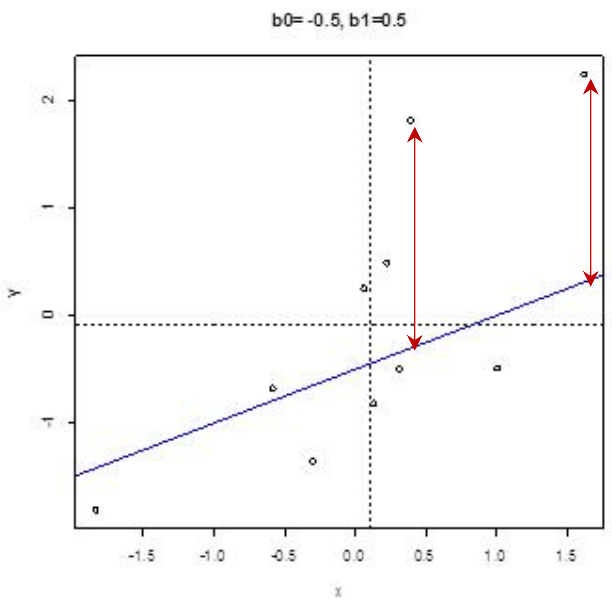
Interpretation:

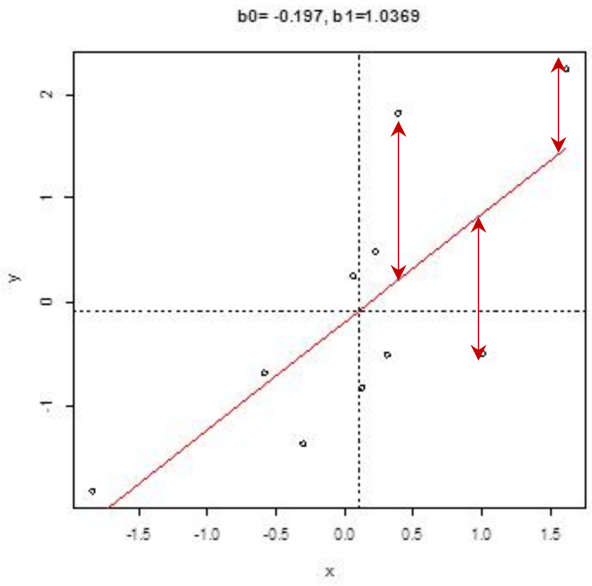
=> β_1: Linear rate of change of y with respect to X

=> $\beta_0 + x_i\beta_1$: 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

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

Covariance(X,Y)

Variance(X)

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

Excel-file (Galton.xls) + INCLASS

Inference in the Linear Regression Model

Sampling Distribution of Estimates

- Data usually constitutes a random sample from a conceptual population.

- Example: a conceptual population may be all individuals in the US between 18-65 years of age.

- In the population there is a regression of Weight on Height

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

- The actual values of the parameters are unknown to the researcher.

- So, we collect a sample and estimate the parameters using some estimation method (e.g., OLS).

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\varepsilon}_i$$

- However, our estimates are not the population parameters, these are simply our best guess about the parameter values given the data.

- There is uncertainty about our estimates given by the fact that we have used a finite sample and not data from the whole population.

Sampling Variability of Estimates

Population	Sample	Estimates
Population	Sample 1 (n)	$\hat{\beta}_0 = 40.3\ ;\ \hat{\beta}_1 = 0.92$
	Sample 2 (n)	$\hat{\beta}_0 = 41.1\ ;\ \hat{\beta}_1 = 0.95$
	Sample 3 (n)	$\hat{\beta}_0 = 39.7\ ;\ \hat{\beta}_1 = 0.99$

Standard Errors & t-statistic

$\Rightarrow$ In practice we have only one sample.

$\Rightarrow$ We estimate the variance of the estimates using formulas that reside on a few assumptions.

$\Rightarrow$ The standard errors (SE) printed in the parameter estimate table are estimates of the square-root of the variance of the estimated parameter (variance over conceptual repeated sampling).

Standard Error of Estimates $$SE(\hat{\beta}) = \sqrt{Var(\hat{\beta})}$$

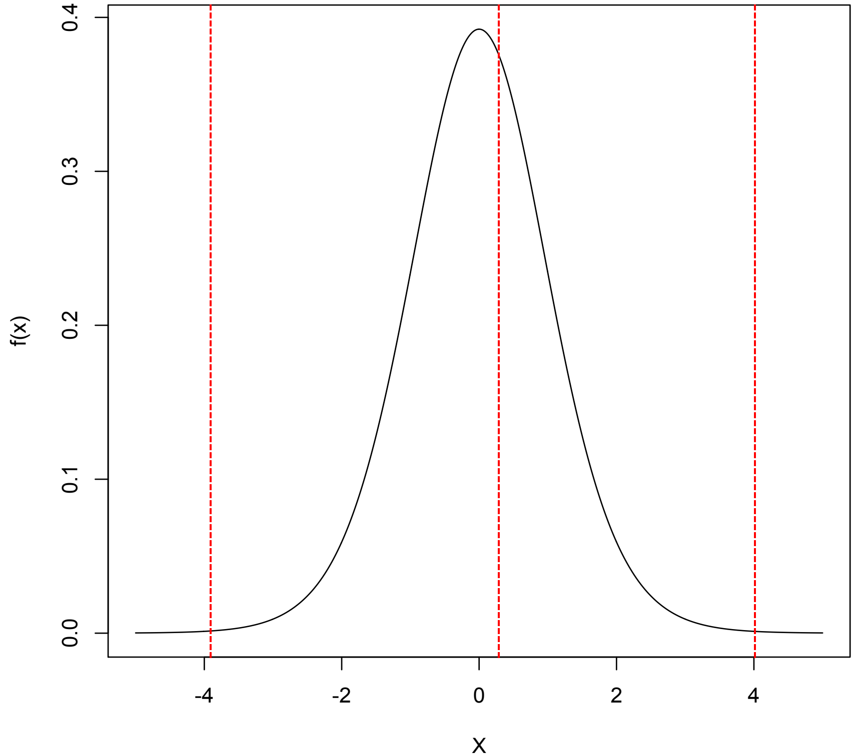
Standardized Estimates (t-statistic) $$t = \frac{\hat{\beta}}{SE(\hat{\beta})}$$

Hypothesis Testing & pValue

Hypothesis $\begin{cases} H_0 : \beta_1 = 0 & y_i = \beta_0 + \varepsilon_i \\ H_A : \beta_1 \neq 0 & y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \end{cases}$

Distribution of the *t*-statistic under H_0

$$\left.\frac{\hat{\beta}}{SE(\hat{\beta})}\right| H_0 \sim t(DF = N-2) \overset{\bullet}{\sim} N(0,1)$$

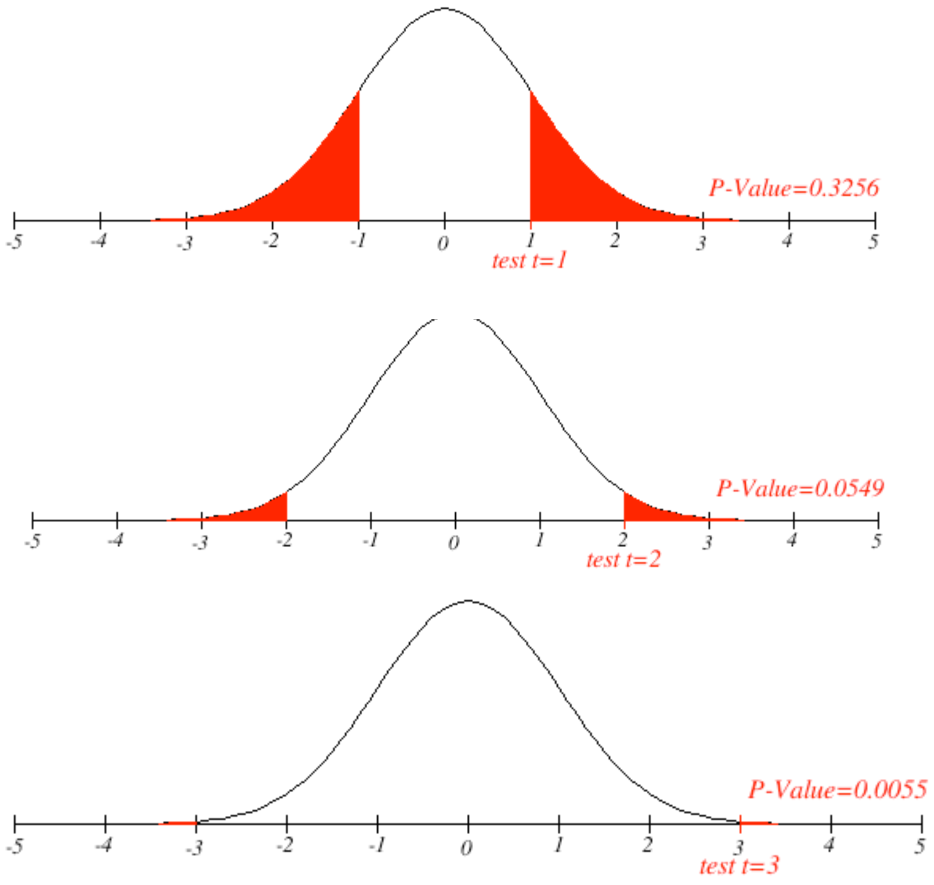


Hypothesis Testing & pValue

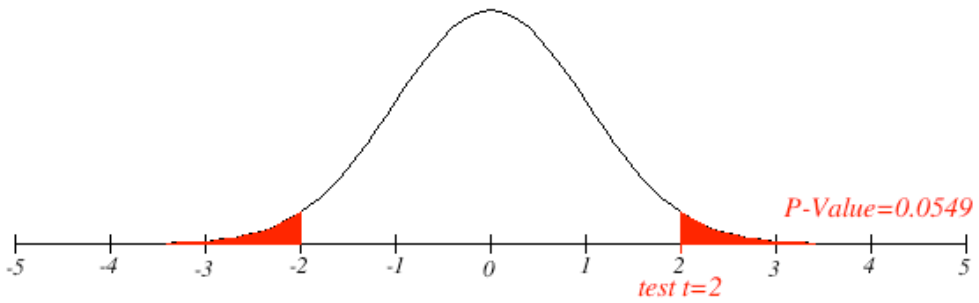
$$\left.\frac{\hat{\beta}}{SE\left(\hat{\beta}\right)}\right| H_0 \sim t\left(DF = N-2\right) \quad 9.4.9$$

Decision Rule: If |t|>k => Reject the Null.   How do we choose k?

Statistical Decision	True State of the Null Hypothesis	
	H_0 True	H_0 False
Reject H_0	Type I error	Correct
Do not Reject H_0	Correct	Type II error



Hypothesis Testing & pValue



pValue: Probability of observing a t-statistic at least as extreme as the one observed if the null hypothesis holds.

Significance: P(Type I Error|H_0) , minimum pValue below which we reject (typically 0.05 or 0.01)

ANOVA

ANOVA

Ha: $y_i = \mu + x_i\beta + \varepsilon_i$

H0: $y_i = \mu + \varepsilon_i$

Main question (s):

How much of the variance un-explained by H0 can be explained by Ha?

Is the additional variance explained by Ha large enough, considering the difference in the number of parameters (df)?

Degrees of freedom

- We begin with n data-points
- The null hypothesis y=mu+e has involves 1 parameter (mu), thus, it has n-1 residual degree of freedom and
- Our Ha involves 2 parameters, thus it has n-2 residual degree of freedom.

ANOVA

Ha: $y_i = \mu + x_i\beta + \varepsilon_i$

H0: $y_i = \mu + \varepsilon_i$

Main question (s):

How much of the variance un-explained by H0 can be explained by Ha?

Is the additional variance explained by Ha large enough, considering the difference in the number of parameters (df)?

Degrees of freedom

- We begin with n data-points
- The null hypothesis $y_i = \mu + \varepsilon_i$ involves 1 parameter (μ), thus, it has n-1 residual degree of freedom and 1 'model df'
- Our Ha involves 2 parameters, thus it has n-2 residual degree of freedom.
- In ANOVA, when we compare H0 and Ha, we have:
    - Model degree of freedom: Difference in the # of parameters in Ha relative to H0
    - Residual DF= n-number of parameters in Ha.

ANOVA: Variance Partition

Model: $y_i = \mu + x_i\beta + \varepsilon_i$

Child's height

Mid-parental height

Data: https://github.com/gdlc/EPI809/blob/master/GALTON.txt

Total SS: $SSy = \sum_{i=1}^{n}(y_i - \bar{y})^2$

Model Residuals: $\hat{\varepsilon}_i = y_i - \hat{\mu} - x_i\hat{\beta}$

Residual Sum of Squares: $RSS = \sum_{i=1}^{n}\hat{\varepsilon}_i^{\,2} = \sum_{i=1}^{n}\left(y_i - \hat{\mu} - x_i\hat{\beta}\right)^2$ (Variability not explained by the model)

Model Sum of Squares: MSS=SSy-RSS= $\sum_{i=1}^{n}\left(\hat{\mu} + x_i\hat{\beta} - \bar{y}\right)^2$

Graphical Representation of Variance Partition

$$\left(y_i - \bar{y}\right) = \left(y_i - \hat{y}_i\right) + \left(\hat{y}_i - \bar{y}\right)$$

$$\left(y_i - \bar{y}\right) = \left(y_i - \hat{y}_i\right) + \left(\hat{y}_i - \bar{y}\right)$$

$$\left(y_i - \bar{y}\right) = \hat{\varepsilon}_i + \left(\hat{y}_i - \bar{y}\right)$$

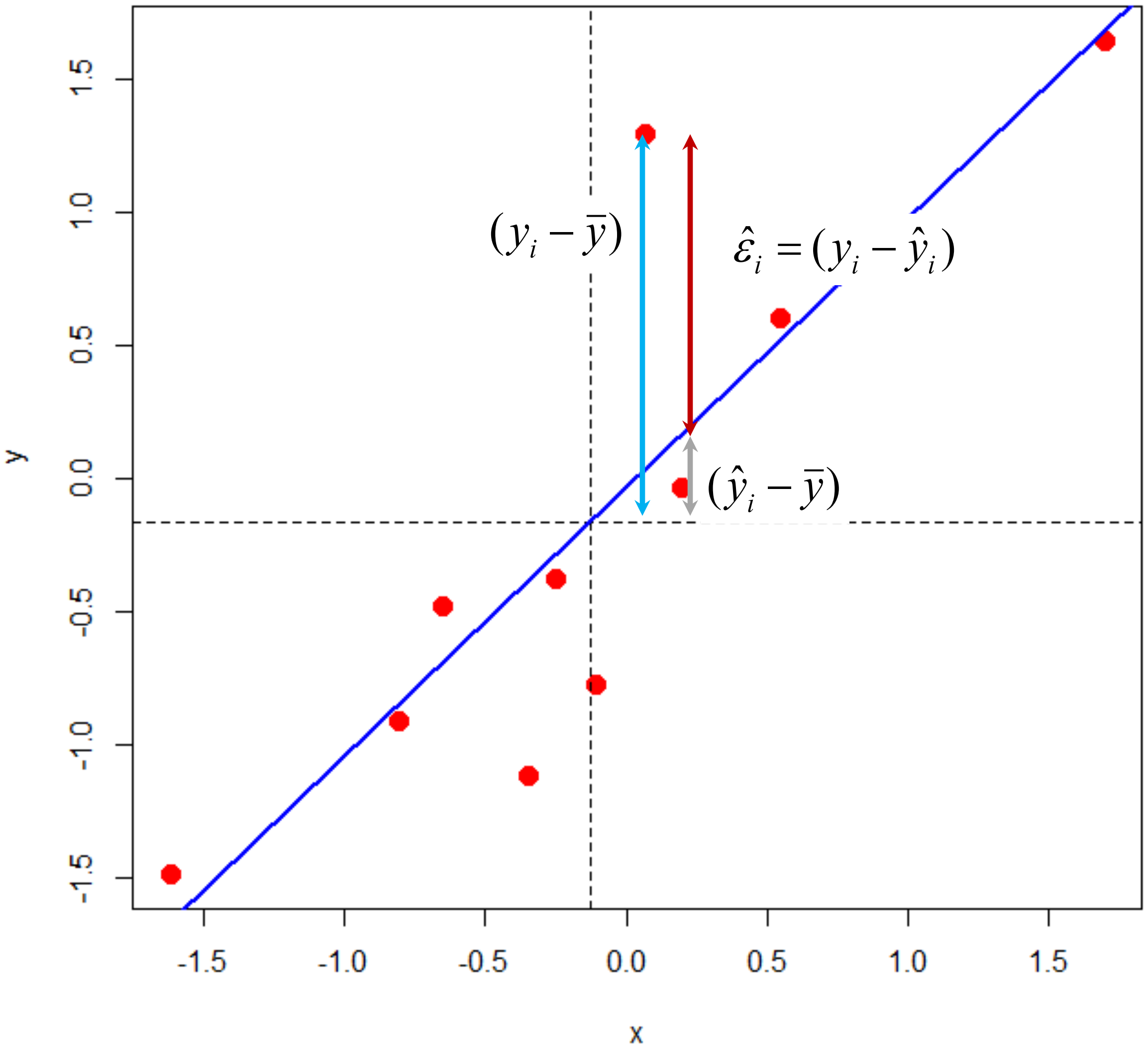
Total= Un-Explained + Explained

Total: $\quad SST = \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2$

$Un-Explained: \quad \mathrm{SSE} = \sum_{i=1}^{n} \hat{\varepsilon}_i^2$

$Explained: \quad \mathrm{SSR} = SST - SSE$

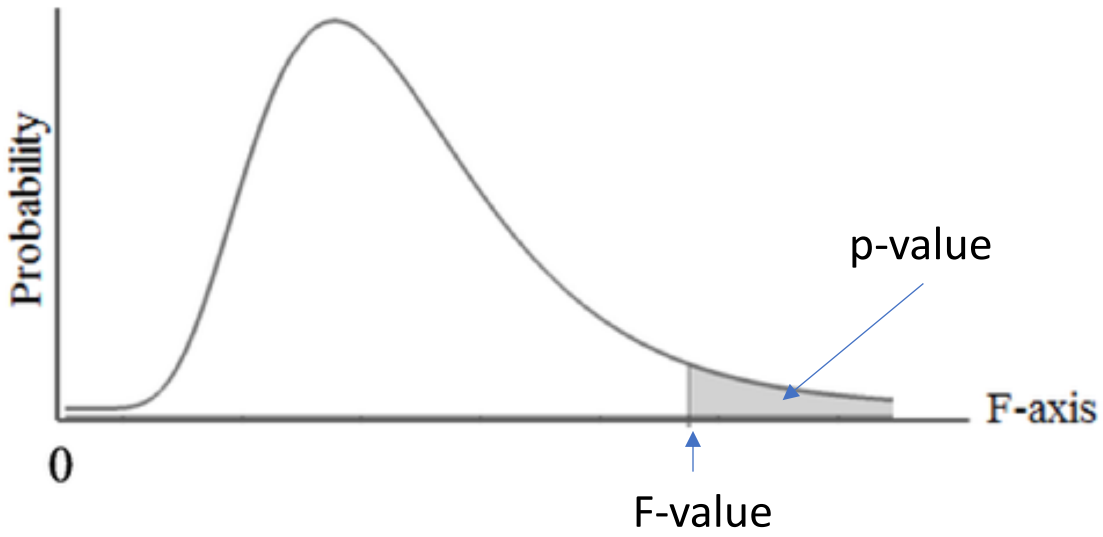
$$R^2 = \frac{SSR}{SSR + SSE}$$



ANOVA

Source	df	SS	MS	F	P-value
Model	p-1	MSS	MSS/(p-1)	$\frac{[MSS/(p-1)]}{[RSS/(n-p)]}$	(see below)
Residual	n-p	RSS	RSS/(n-p)		
Total	n				

Under the null, the F-statistic, F=[MSS/(p-1)]/[RSS/(n-p)], the F-statistic follows an F distribution wit p-1. and n-p DF.




```
pf(F-stat,DF1,DF2,lower.tail=F)
```

ANOVA & st

Work on this example:

https://github.com/gdlc/EPI809/blob/master/GALTON.txt