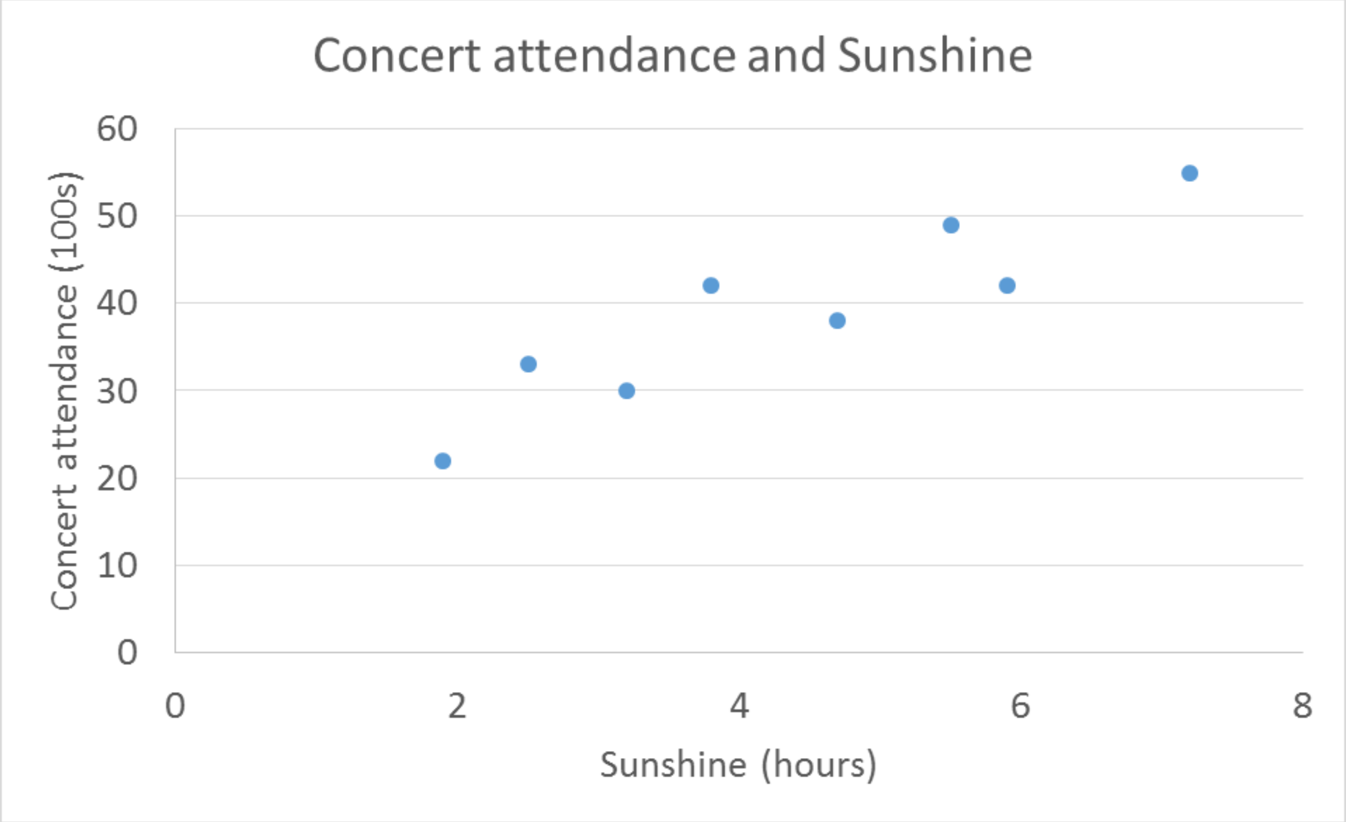
Linear Regression

Simple Linear Regression

Example

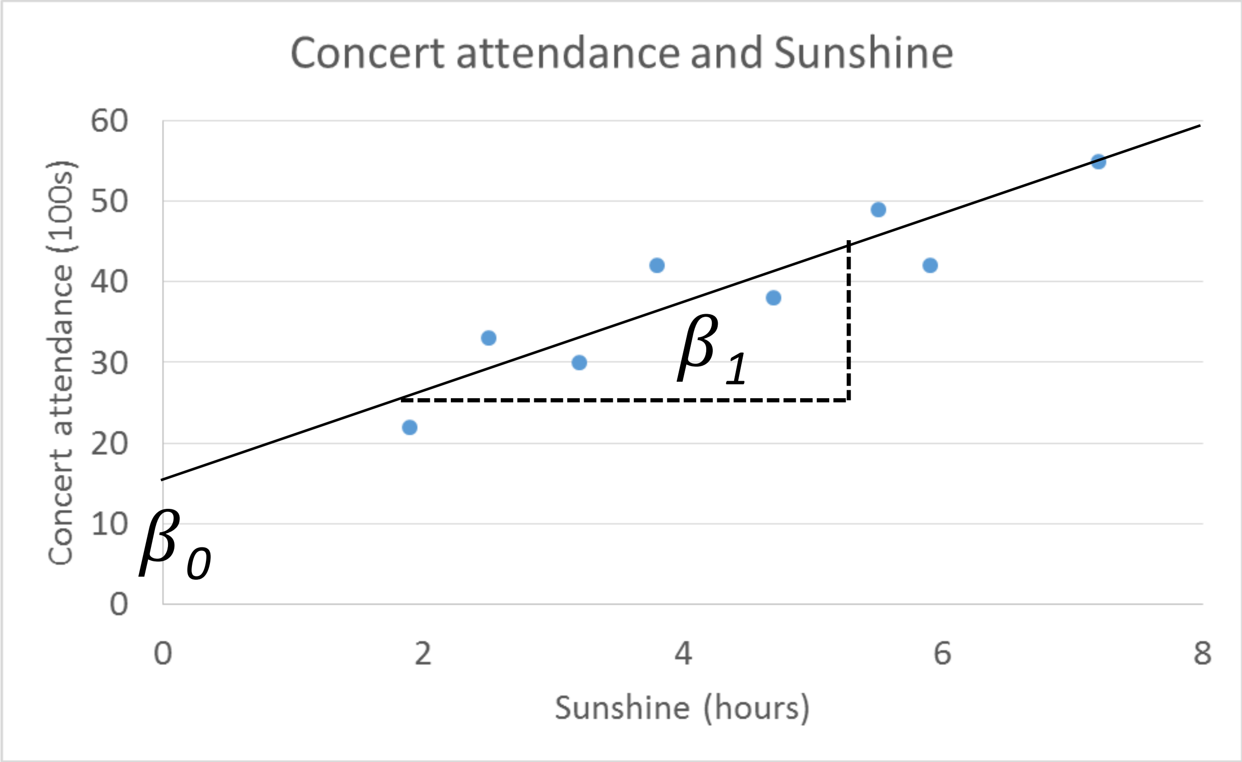
- Impact of weather on event attendance
- Correlated? Predictable?

Sunshine (hours)	1.9	2.5	3.2	3.8	4.7	5.5	5.9	7.2
Concert attendance (100s)	22	33	30	42	38	49	42	55

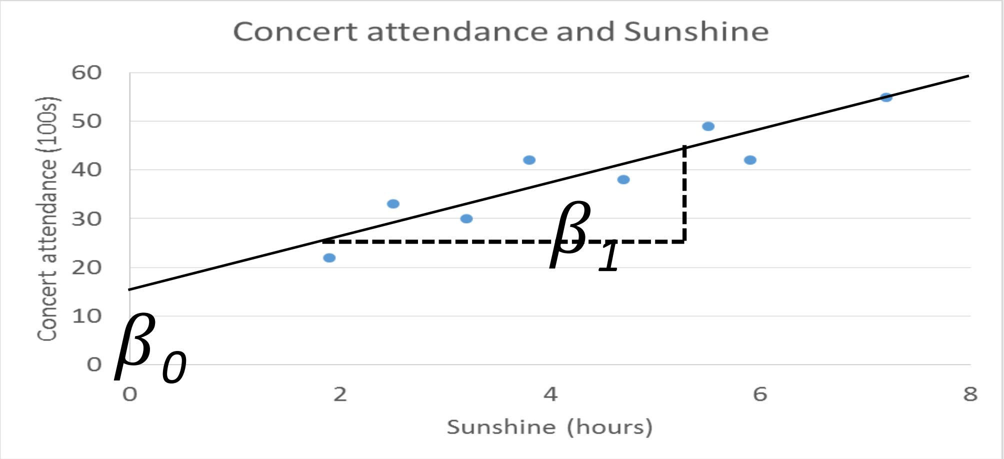


Simple Linear Regression

- Regression
  - Dependent variable is numeric
- Linear
  - Fit a line
  - Line : Coefficients
- Optimization
  - Many possible lines
  - Criteria : Minimize error
- Error
  - Sum of squared residuals



Linear Regression : Math



- Linear Regression
  - Dependent variable is numeric
  - Fit a line

- Line Fitting
  - More than 2 data points ➔ Over-specified problem
  - Criteria : Minimize error (sum of squared residuals)

$$y \approx \beta_0 + \beta_1 x$$

$$y = \beta_0 + \beta_1 x + \epsilon$$
$$\epsilon \sim N(0, \sigma^2)$$

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

- Optimization problem
  - Solve (using calculus)
  - Find coefficients (line) which minimizes the Residual Sum of Squares

$$\min_{\beta} RSS$$

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

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

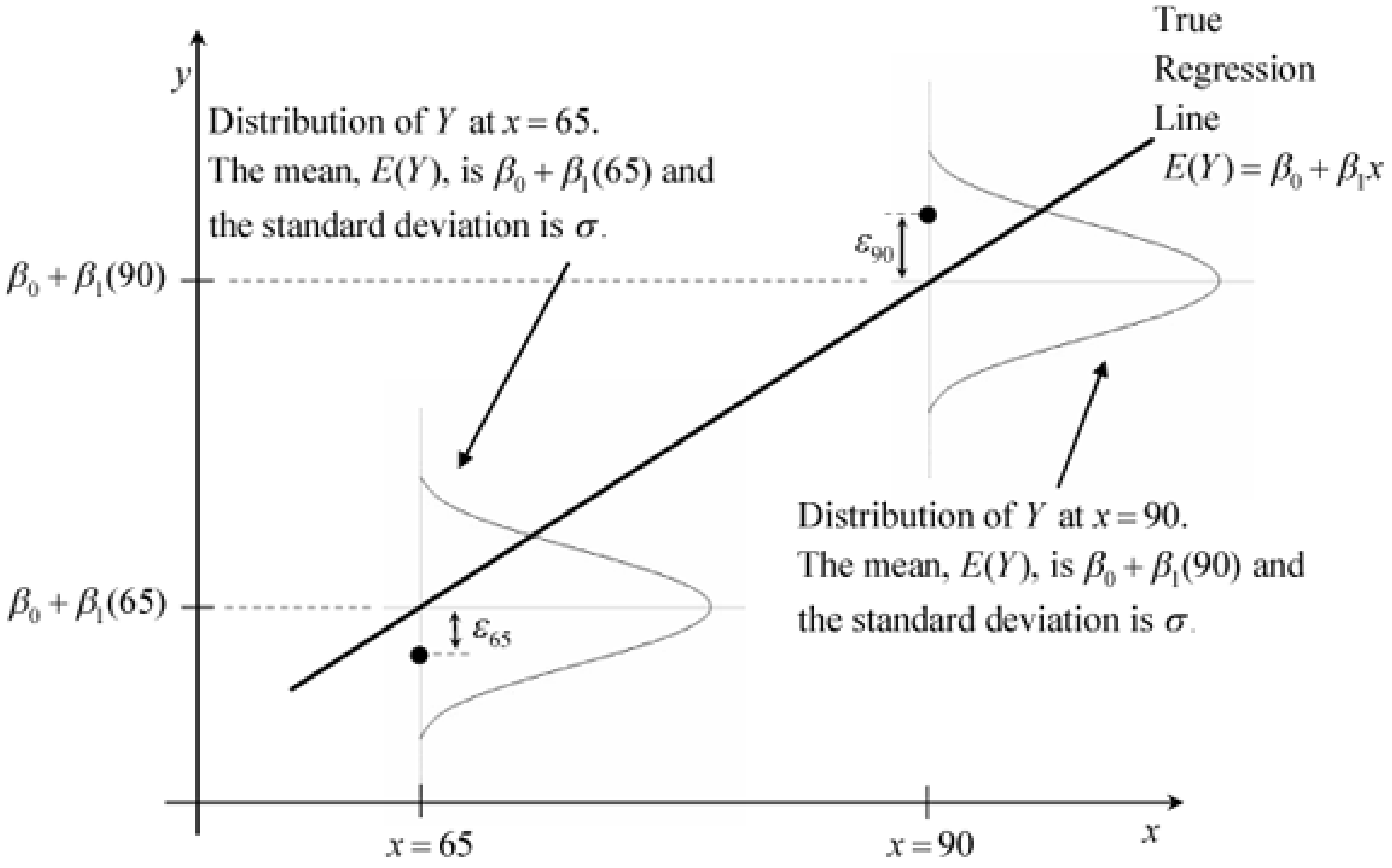
- Use estimated coefficients (“model”) to make predictions

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

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Linear Regression : Intuition

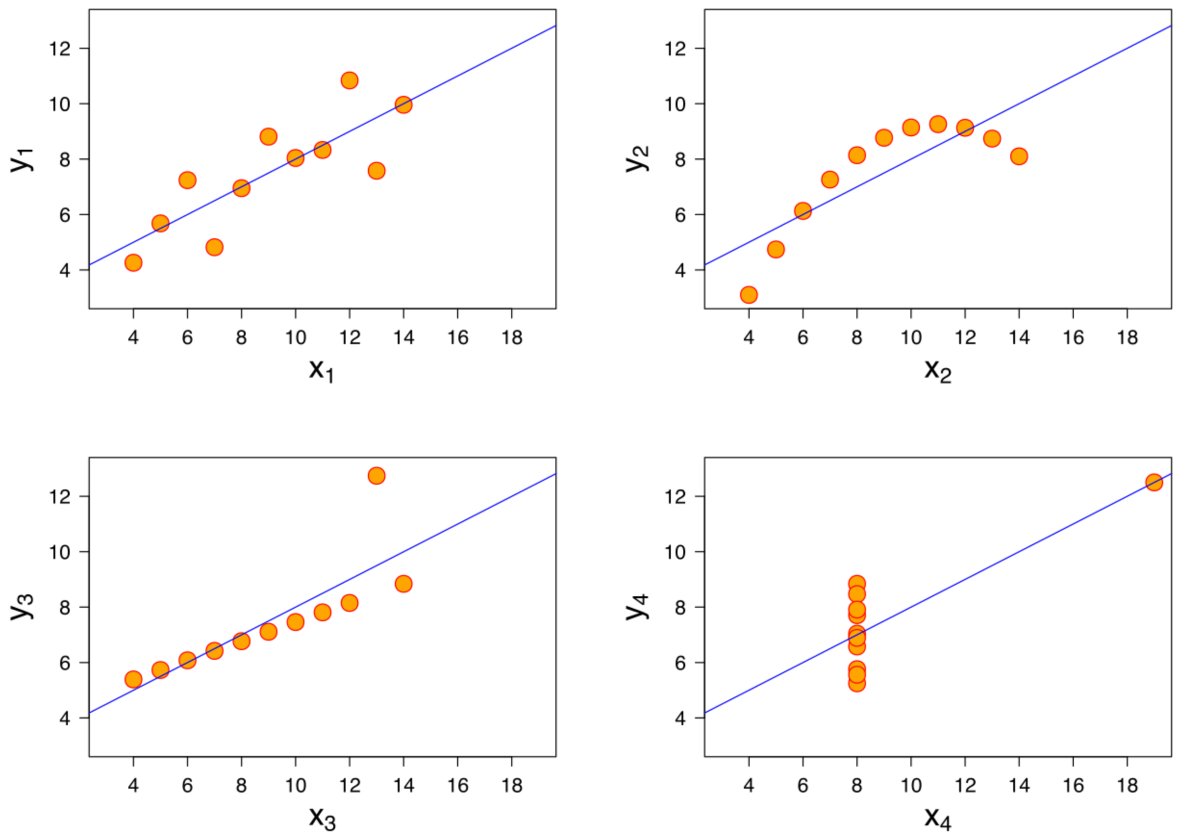
$$y = \beta_0 + \beta_1 x + \epsilon$$
$$\epsilon \sim N(0, \sigma^2)$$



http://reliawiki.org/index.php/Simple_Linear_Regression_Analysis

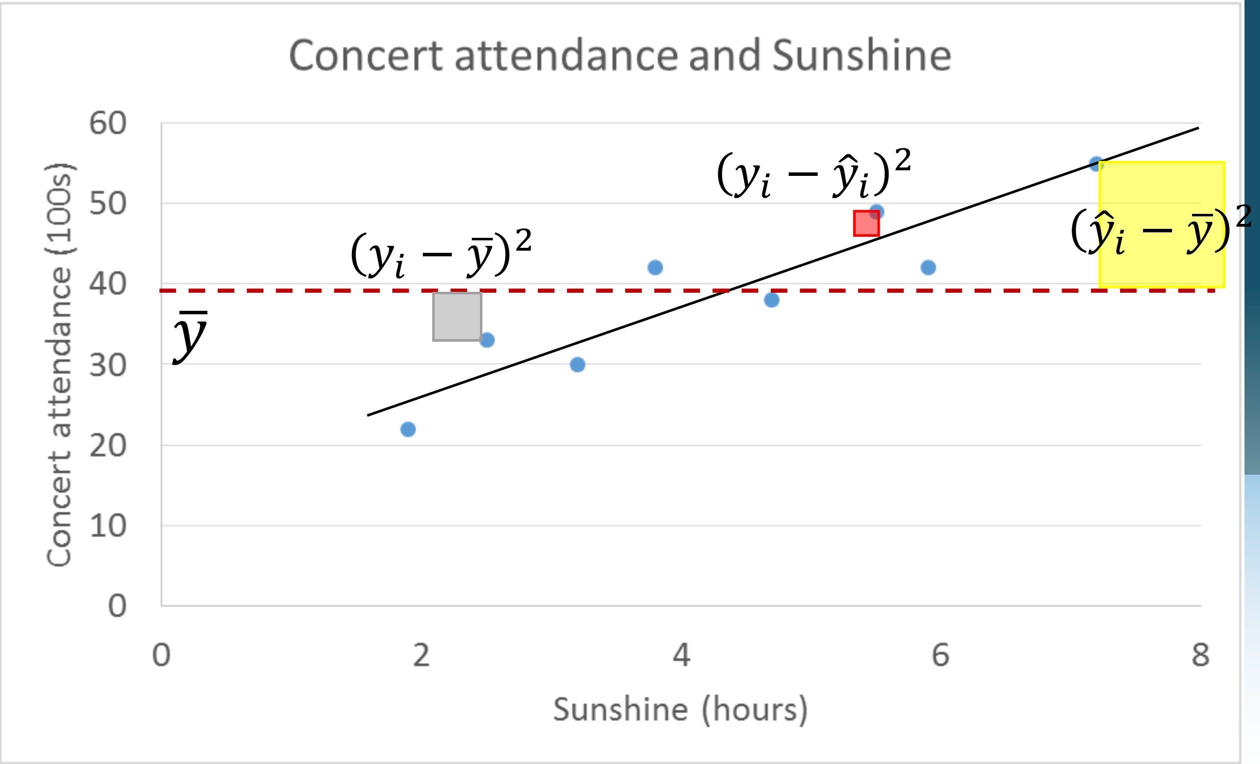
How good is your line?

- Among all possible lines, LR selects one that minimize the RSS
  - Is this good enough?
  - Visual comparison
  - Quantification



How good is your line? : Quantify.

- How good is your line / fit / model?
  - What would be the best line?
  - RSS = 0 : Not always possible : over-specified problem 2 variables, n data points
- Goodness of line = RSS?
  - Depends on the units of y
  - What is big? What is small?
  - Interpretability? Model comparison?
- Coefficient of Determination R-sq (R^2)
  - Intuition: P(Y|X) should have low variance
  - $TSS = \sum(y_i - \bar{y})^2$
  - $ESS = \sum(\hat{y}_i - \bar{y})^2$
  - $RSS = \sum(y_i - \hat{y}_i)^2$
  - $TSS = ESS + RSS$
  - $R^2 = \frac{ESS}{TSS} = \frac{TSS-RSS}{TSS} = 1 - \frac{RSS}{TSS}$
  - = Square of the pearson correlation (for simple LR)

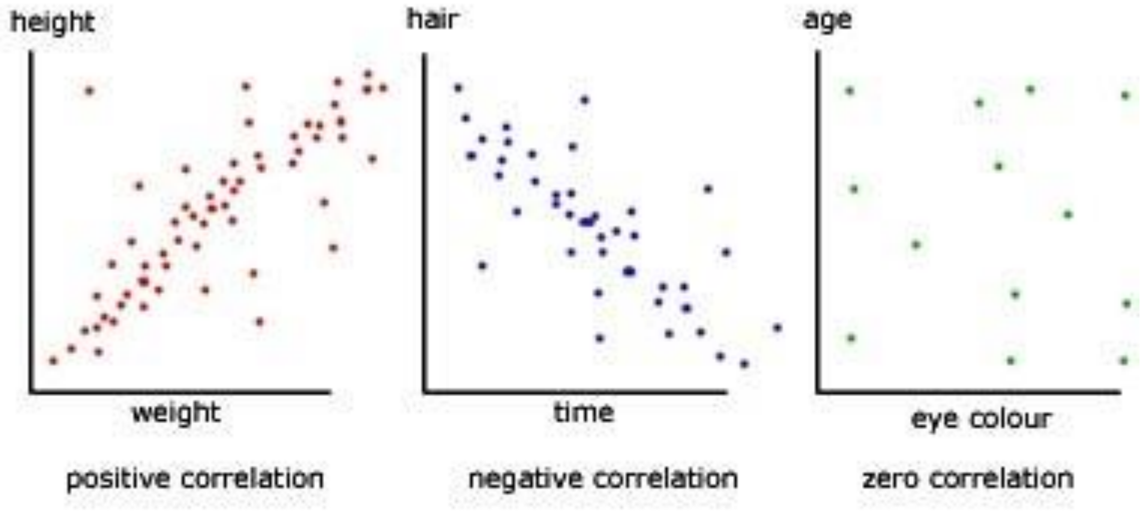


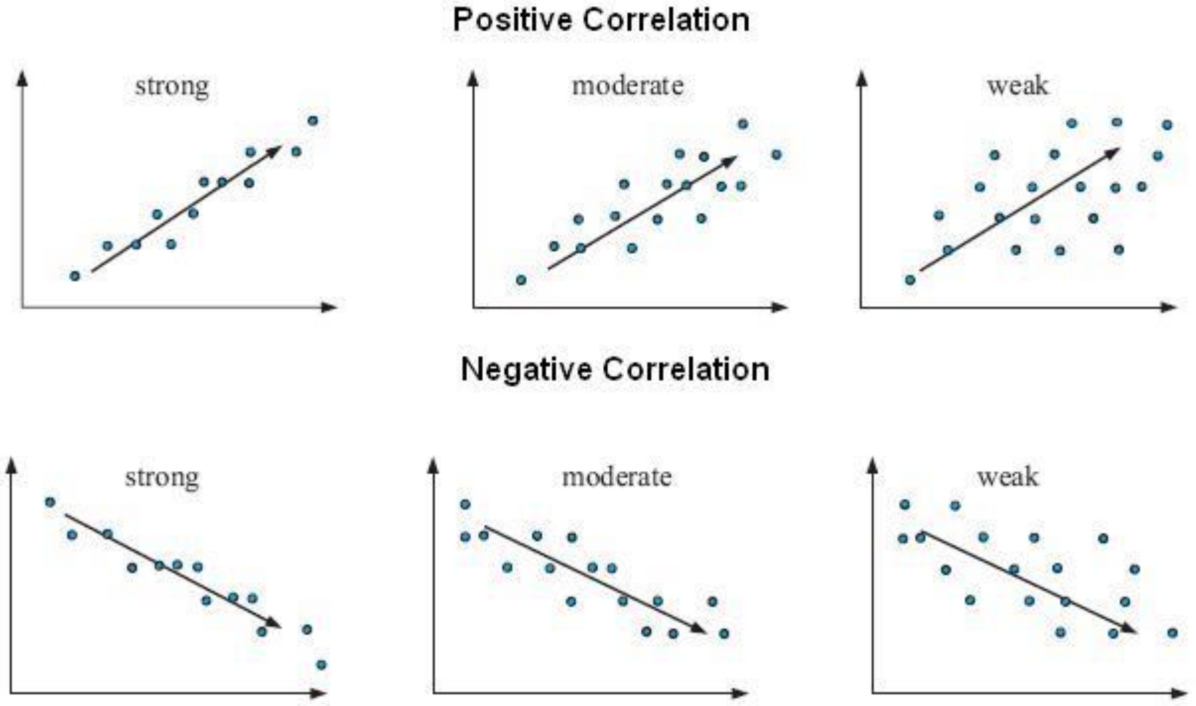
$$1 = \frac{ESS}{TSS} + \frac{RSS}{TSS} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} + \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$$

Coefficient of Determination : Correlation

- Coefficient of Determination R-sq (R^2)
  - $1 - \frac{SSE}{SST} = R^2$
  - = Square of the pearson correlation (for simple LR)

$$r = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2} \sqrt{\sum_i (y_i - \overline{y})^2}}$$





LR: Statistics?

Data : Sample or Population

- Different lines for different samples of the data
  - Estimated parameters depend on the data set

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- Prediction (Model)
  - For a given x, predict y : use given (sample) data to establish a relationship
  - For a given x, predict y : use given (sample) data to build a model
  - Use given (sample) data to build a model which can be applied on population (future data points)
  - A regression line provides a point estimate from a sample.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- Estimated parameters
  - Are sample statistics
  - Are random variables
  - Will create a sampling distribution

Inferential Statistics on model parameters

- Sampling Distribution of model parameters
  - Standard Error (s.d. of the sampling distribution)

$$SE(\hat{\beta}_1) = \sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \qquad SE(\hat{\beta}_0) = \sqrt{\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}$$

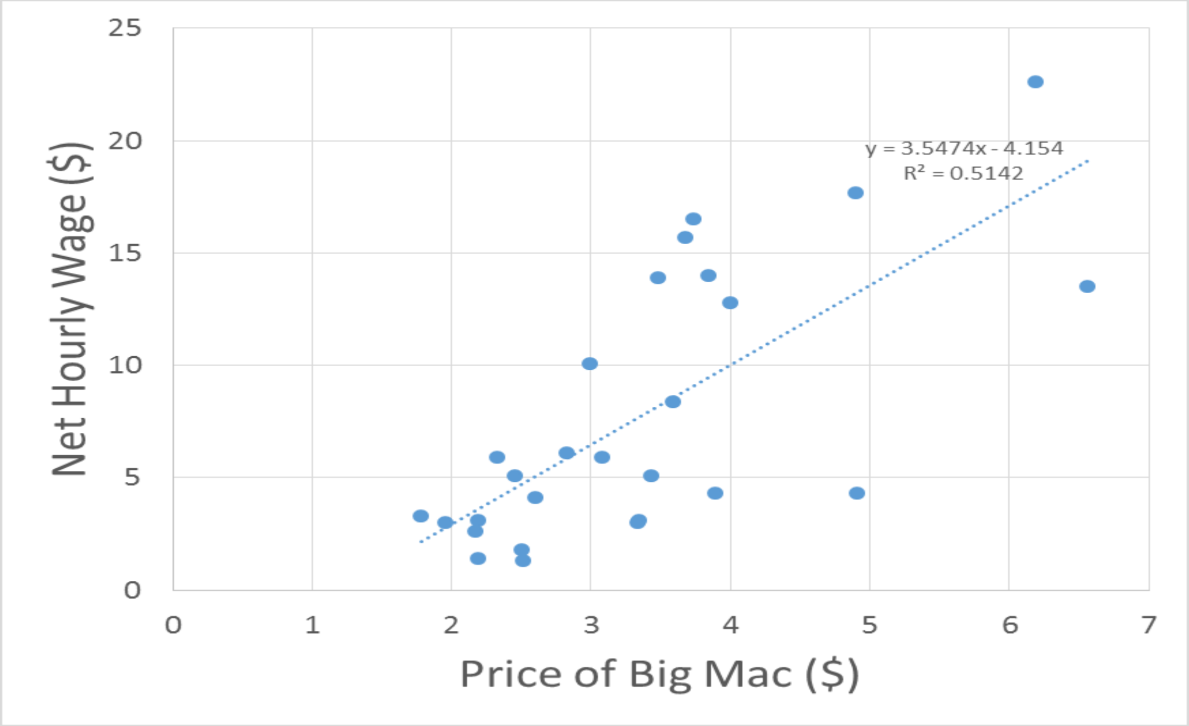
- Variance of the population
  - Unknown
  - Estimate (Residual Standard Error)
  - Assume large enough sample

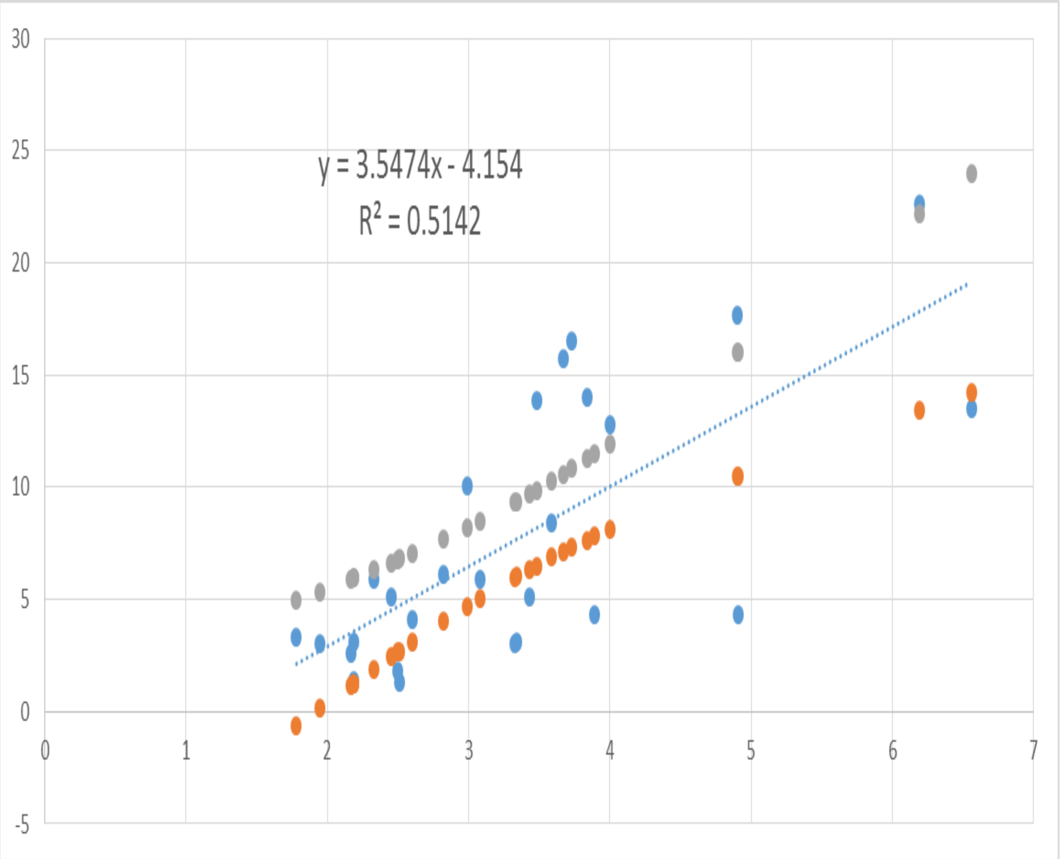
$$\hat{\sigma}^2 = RSE = \sqrt{\frac{RSS}{(n-2)}} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2}{n-2}}$$

- Confidence Interval
  - In which the true (population) parameters lie

$$95\%\ \text{C.I.}:\ \hat{\beta}_1 \pm 2SE(\hat{\beta}_1)$$
$$95\%\ \text{C.I.}:\ \hat{\beta}_0 \pm 2SE(\hat{\beta}_0)$$

Example (cont'd)





	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-4.154014573	2.447784673	-1.697050651	0.102104456	-9.195321476	0.88729233
Big Mac Price ($)	3.547427488	0.689658599	5.143744297	2.57053E-05	2.127049014	4.967805962

$$95\% \text{ C.I. : } \hat{\beta}_1 \pm 2SE(\hat{\beta}_1)$$

$$95\% \text{ C.I. : } \hat{\beta}_0 \pm 2SE(\hat{\beta}_0)$$

Q?

Praphul Chandra