Forecasting Methods / Métodos de Previsão

Week 2- Simple Linear Regression

ISCTE - IUL, Gestão, Econ, Fin, Contab.

Diana Aldea Mendes

diana.mendes@iscte.pt

February 10, 2011

Regression models

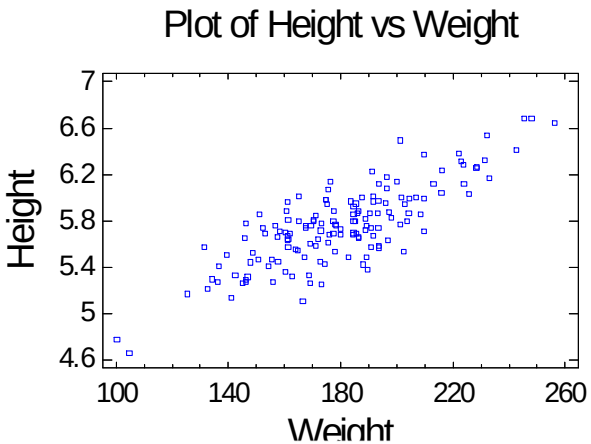
- Consider two variables x and y
- **Main objective**: establish a relationship between variables
- **Regression analysis** is used to predict the value of one variable (the dependent variable) on the basis of other variables (the independent variables) (the effect between variables it is considered **causal**)
- If we are interested only in determining whether a relationship exists, we employ **correlation analysis**
- **Pearson correlation coefficient (**measures the relative strength of the linear relationship between two variables**)**, $-1 < R < 1$

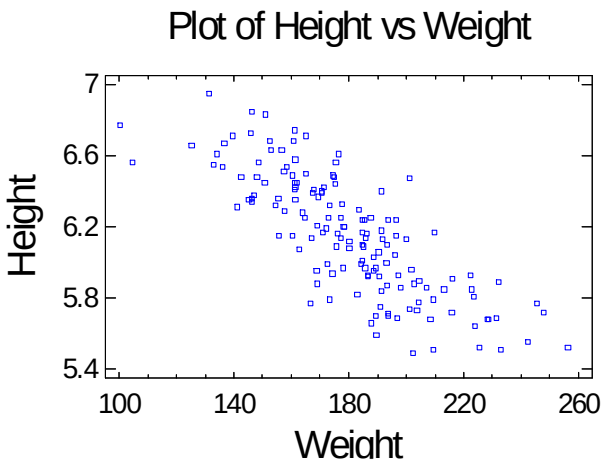
$$R = \frac{\sum_{t=1}^{T} (x_t - \overline{x})(y_t - \overline{y})}{\sqrt{\sum_{t=1}^{T} (x_t - \overline{x})^2} \sqrt{\sum_{t=1}^{T} (y_t - \overline{y})^2}}$$

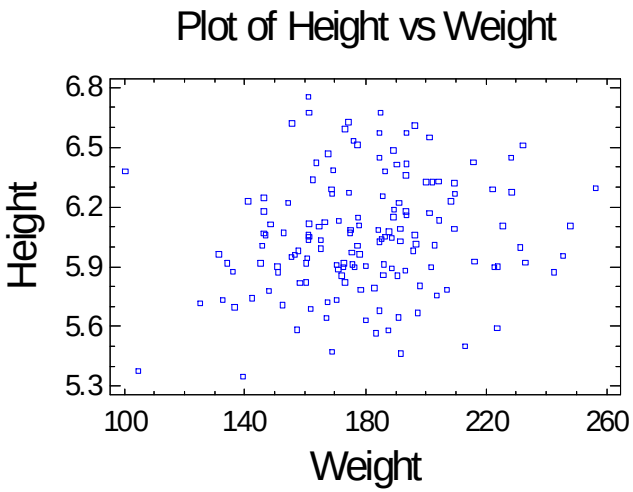
  - If the correlation coefficient is close to $+1(-1)$ that means you have a strong positive (negative) (linear) relationship.
  - If the correlation coefficient is close to 0 that means you have no correlation.

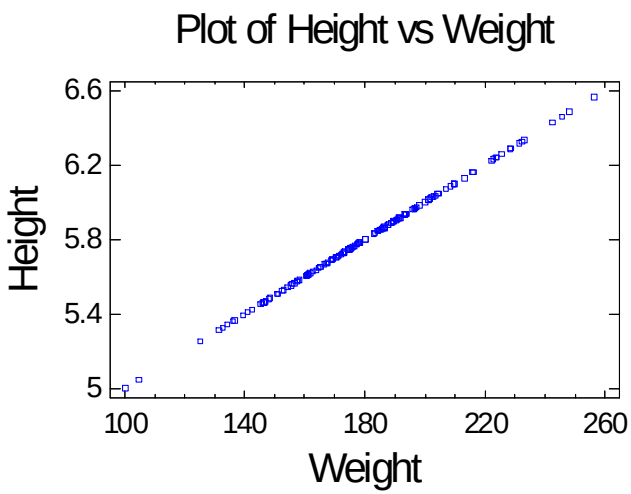
Regression models

- Correlation: scatter plot



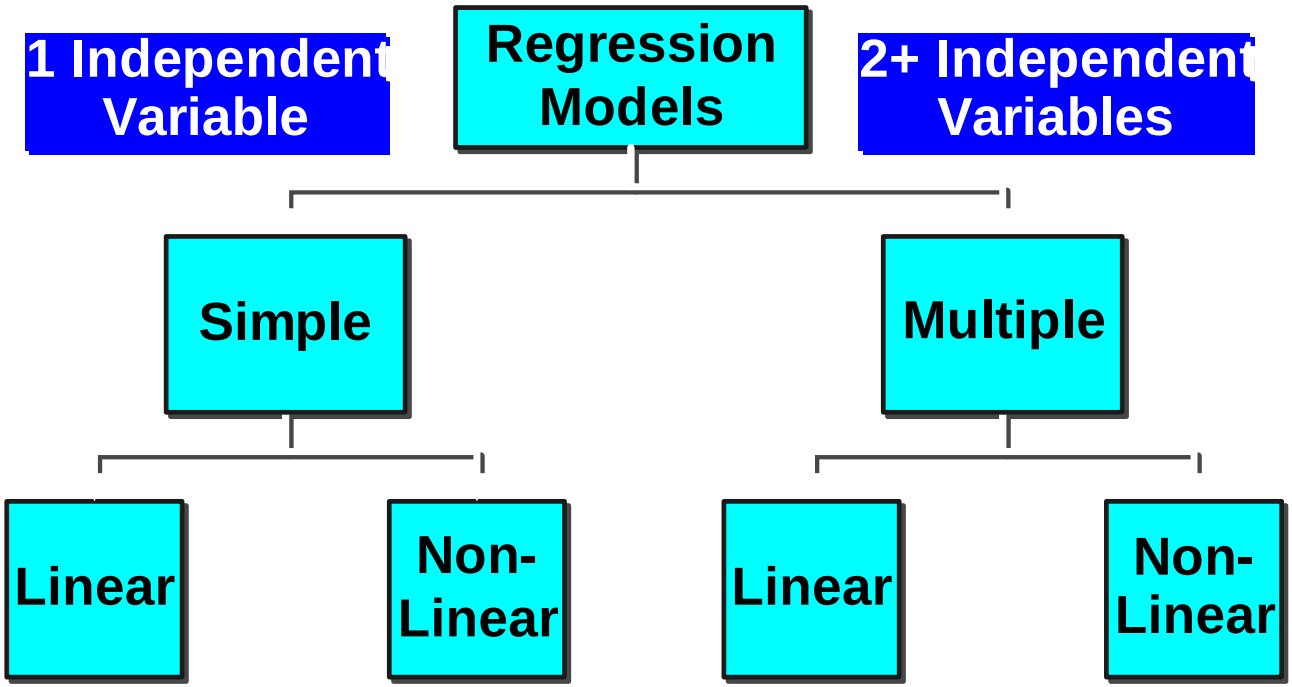






Regression models

- Specific statistical methods for finding the "line of best fit" for one dependent numerical variable based on one or more independent variables.



Regression models

- Steps to do in a regression model
  - Graphical analysis (scatter plot) of the sample points (x, y), in order to decide if exists a linear relation between (x, y)
  - Use the sample to estimate the unknown parameters
  - Check Residual properties (residual analysis)
  - Test Reliability and Validity of the model (statistical evaluation to asses the "goodness of fit" of the model)
  - The validate model can be used for forecasting

Linear regression

- **Simple linear regression model** (1 dependent variable (we observe this) y and 1 independent variable (we provide this) x)

$$y = \underbrace{\beta_0 + \beta_1 x}_{\text{deterministic}} + \underbrace{\varepsilon}_{\text{random error}}$$

- The y variable is assumed to be random or "stochastic" and the x variables is assumed to have fixed ("non-stochastic") values in repeated samples.
- Some alternative names for the y and x variables:
  - y : regressand, explained variable, effect variable, endogenous variable
  - x : regressors, causal variables, explanatory variable, exogenous variable
- β_0 and β_1 = regression **coefficients** (parameters to be estimated)
- β_0 = **intercept** ($\cap$ with y), β_1 = **slope**
- ε = normal random variable (disturbance term), zero mean and constant variance σ^2, $\varepsilon \sim N\left(0, \sigma^2\right)$

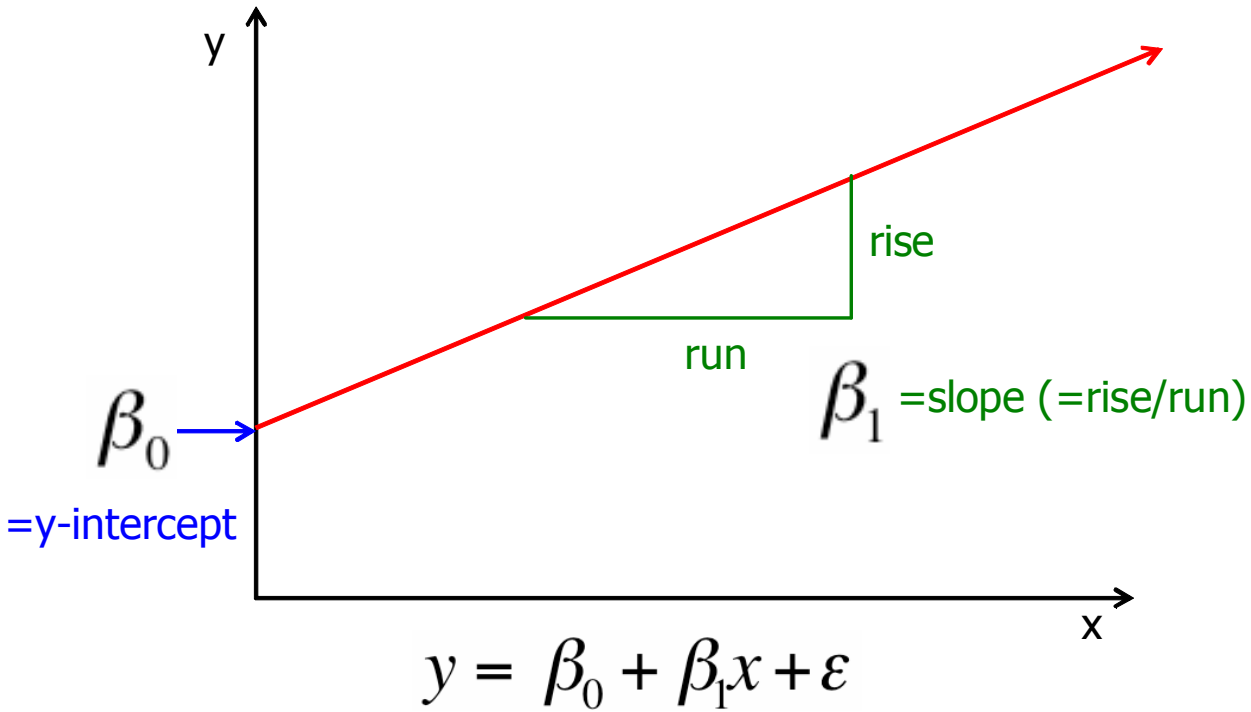
Linear regression

- The disturbance term ε can capture a number of features:
  - We always leave out some determinants of y
  - There may be errors in the measurement of y that cannot be modelled.
  - Random outside influences on y which we cannot model
  - non-linearity
  - random nature of human behavior

Linear regression

Meaning of β_0 and β_1

$\beta_1 > 0$ [positive slope] $\beta_1 < 0$ [negative slope]

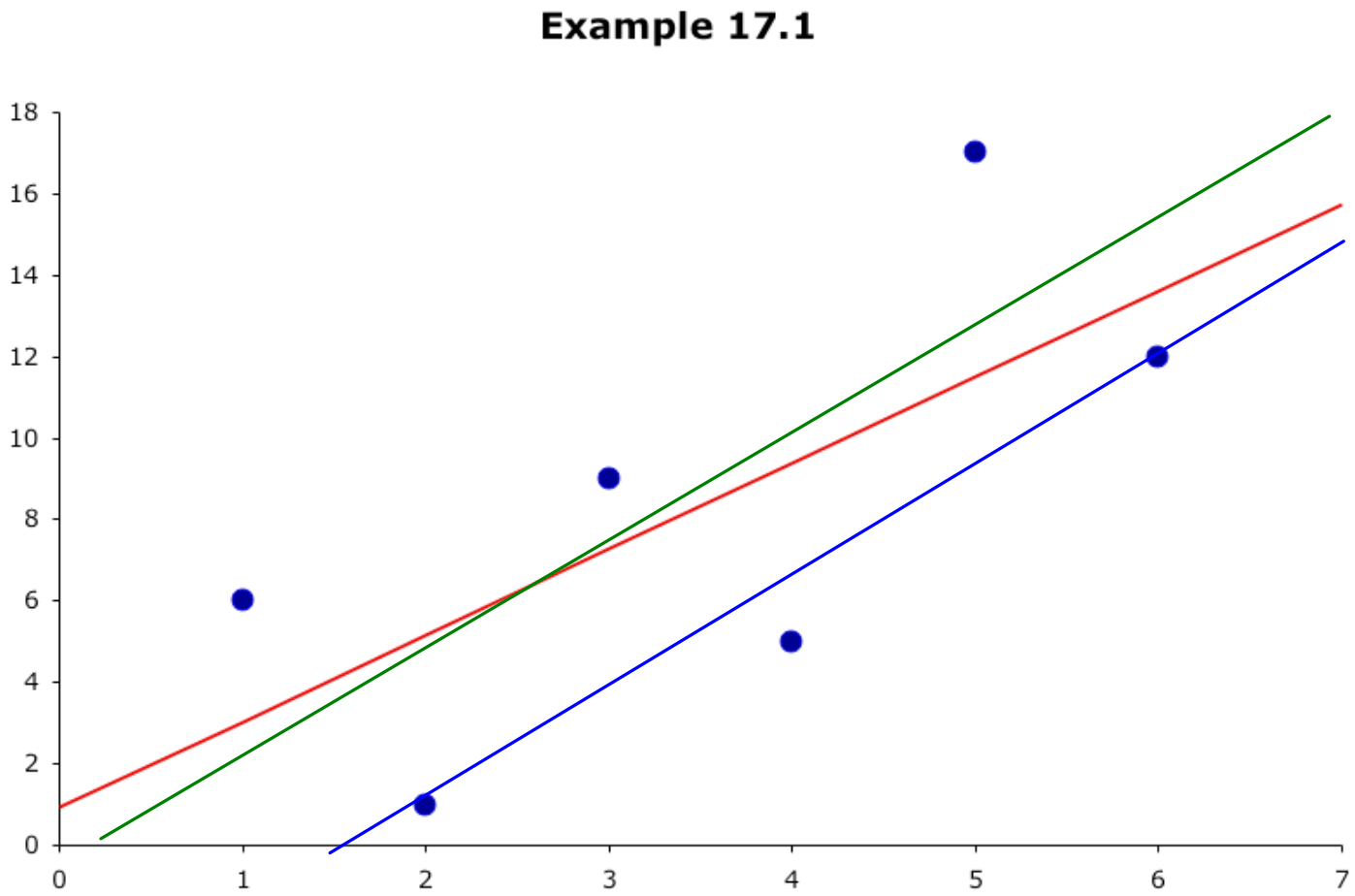


Linear regression

- Basic idea of regression is to estimate the population parameters from a sample (inference)
- The population is the total collection of all objects or people to be studied
- A sample is a selection of just some items from the population
- We also want to know how "good" our estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ are
- In order to use ordinary least square method (the best), we need a model which is linear in the parameters (β_0 and β_1). It does not necessarily have to be linear in the variables (y and x).
- **Estimators** are the formulae used to calculate the coefficients and **estimates** are the actual numerical values for the coefficients.

Linear regression

- Which line has the best "fit" to the data?



Linear regression

- So how do we determine what β_0 and β_1 are the best (or best line to fit the data)?
- Choose β_0 and β_1 so that the (vertical) distances from the data points to the fitted lines are minimized (so that the line fits the data as closely as possible): The most common method used to fit a line to the data is known as OLS (ordinary least squares).
- What we actually do is take each distance

$$y_t - \hat{y}_t = u_t = y_t - \widehat{\beta}_0 - \widehat{\beta}_1 x_t$$

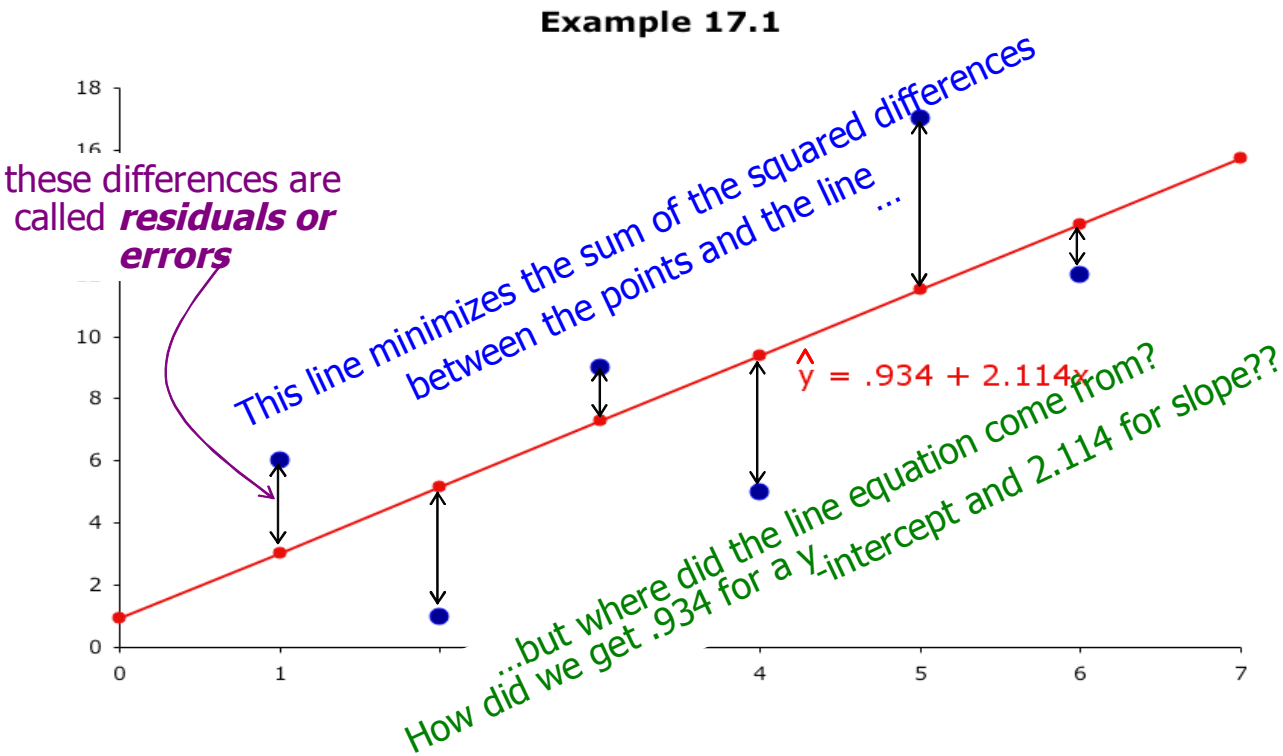
and square it and minimize the total sum of the squares (hence least squares).

$$\min(SSE) = \min \sum_{t=1}^{T} u_t^2 = \min \underbrace{\sum_{t=1}^{T} \left(y_t - \widehat{\beta}_0 - \widehat{\beta}_1 x_t \right)^2}_{SSR = \textbf{risk } r}$$

Linear regression

- we use the following notation:
  - y_t and x_t, $t = 1, ..., T$ denote the actual data points (variables y and x)
  - $\hat{y}_t$ $t = 1, ..., T$ denote the fitted value from the regression line
  - $\hat{\beta}_0$ and $\hat{\beta}_1$ denote the estimated coefficients
  - $u_t = y_t - \hat{y}_t$ denote the regression residuals (estimate of the error term, is the difference between the fitted line and the sample point)
  - $L(y, \widehat{y}) = (y - \widehat{y})$ - loss function

Linear regression



- Note that the values of the residuals are not the same as the values of the disturbance term. The diagram now shows the true unknown relationship as well as the fitted line.

Linear regression

- Ordinary least square = Optimization problem
- Purpose: estimate the coefficients β_0 and β_1
- First order conditions

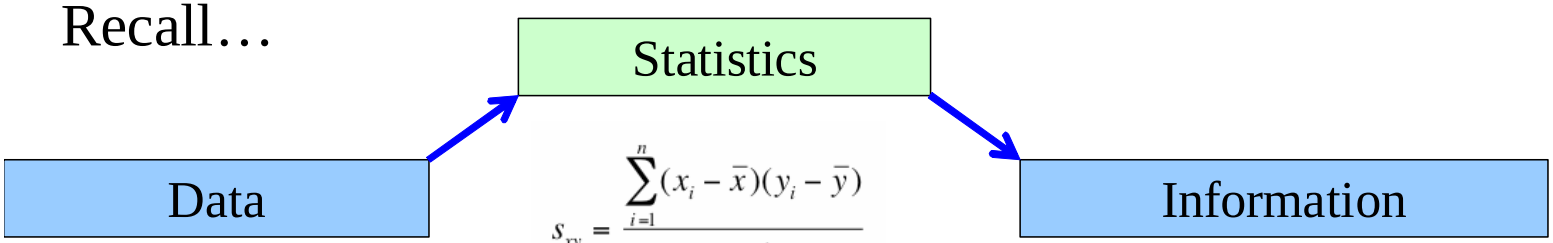
$$\begin{cases} \dfrac{\partial r}{\partial \widehat{\beta}_0} = -2\sum_{t=1}^{T}\left(y_t - \widehat{\beta}_0 - \widehat{\beta}_1 x_t\right) = 0 \\ \dfrac{\partial r}{\partial \widehat{\beta}_1} = -2\sum_{t=1}^{T} x_t\left(y_t - \widehat{\beta}_0 - \widehat{\beta}_1 x_t\right) = 0 \end{cases} \quad \begin{cases} \widehat{\beta}_1 = \dfrac{\sum_{t=1}^{T}(x_t - \bar{x})(y_t -} {\sum_{t=1}^{T}(x_t - \bar{x})^2} \\ \widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x} \end{cases}$$

$$\bar{x} = \frac{\sum_{t=1}^{T} x_t}{T},\ \bar{y} = \frac{\sum_{t=1}^{T} y_t}{T} \text{ (mean)}$$

- The **slope estimate** is the sample covariance between x and y divided by the sample variance of x
- If x and y are positively correlated, the slope will be positive
- If x and y are negatively correlated, the slope will be negative

Linear regression

Recall…



Data Points:

x	y
1	6
2	1
3	9
4	5
5	17
6	12

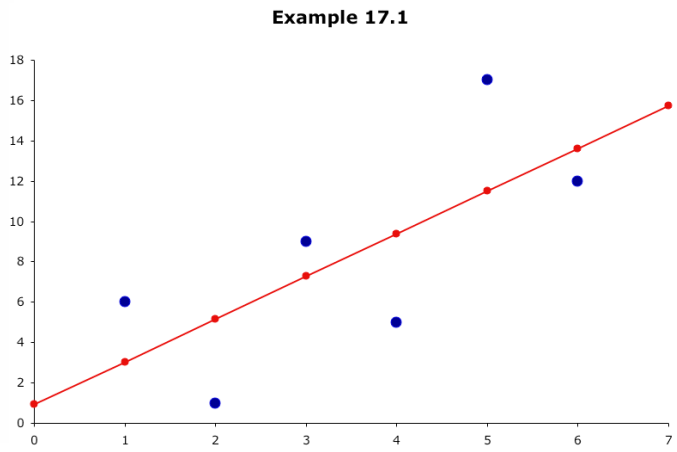
$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

$$b_1 = \frac{s_{xy}}{s_x^2}$$

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

$$b_0 = \bar{y} - b_1\bar{x}$$



$\hat{y} = .934 + 2.114x$

Linear regression

The Assumptions Underlying the Linear Regression Model (LRM)

- We observe data for x_t, but since y_t also depends on u_t (ε_t), we must be specific about how the u_t (ε_t) are generated.
- We make the following set of assumptions about the u_t's (the unobservable error terms) for the regression methods to be valid :

1. $E(u_t) = 0$, The errors have zero mean
2. $Var(u_t) = \sigma^2, \forall x_t$, The variance of the errors is constant and finite over all values of x_t (homoscedasticity)
3. $Cov(u_i, u_j) = 0$ ($E(u_i, u_j) = 0$, $(i \neq j)$), The errors are statistically independent of one another
4. $Cov(u_t, x_t) = 0$ ($E(u_t, x_t) = 0$), No relationship between the error and corresponding x variable (alternative assumption: the x_t's are non-stochastic or fixed in repeated samples.)
5. u_t is normally distributed ($u \sim N(0, \sigma^2)$) (this assumption is required if we want to make inferences about the population parameters from the sample parameters)

Linear regression

The Assumptions Underlying the Linear Regression Model (LRM)

- If assumptions 1. through 4. hold, then the estimators determined by OLS are known as **Best Linear Unbiased Estimators** (BLUE).
  - "Estimator" - is an estimator of the true value of y.
  - "Linear" - is a linear estimator
  - "Unbiased" - On average, the actual value of the parameters will be equal to the true values.
  - "Best" - means that the OLS estimator has minimum variance among the class of linear unbiased estimators. The Gauss-Markov theorem proves that the OLS estimator is best.

Linear regression

The Assumptions Underlying the Linear Regression Model (LRM)

- **Consistent :** The least squares estimators are consistent. That is, the estimates will converge to their true values as the sample size increases to infinity (Need the assumptions 2 and 4 to prove this).
- **Unbiased:** The least squares estimates are unbiased. That is $E\left(\widehat{\beta}\right) - \beta = 0$, thus on average the estimated value will be equal to the true values (to prove this we need assumption 1). Unbiasedness is a stronger condition than consistency.
- **Efficiency:** An estimator of parameter β is said to be efficient if it is unbiased and no other unbiased estimator has a smaller variance, i.e. $Var\left(\widehat{\beta}\right) < Var\left(\widetilde{\beta}\right)$. If the estimator is efficient, we are minimizing the probability that it is a long way off from the true value of β.

Linear regression

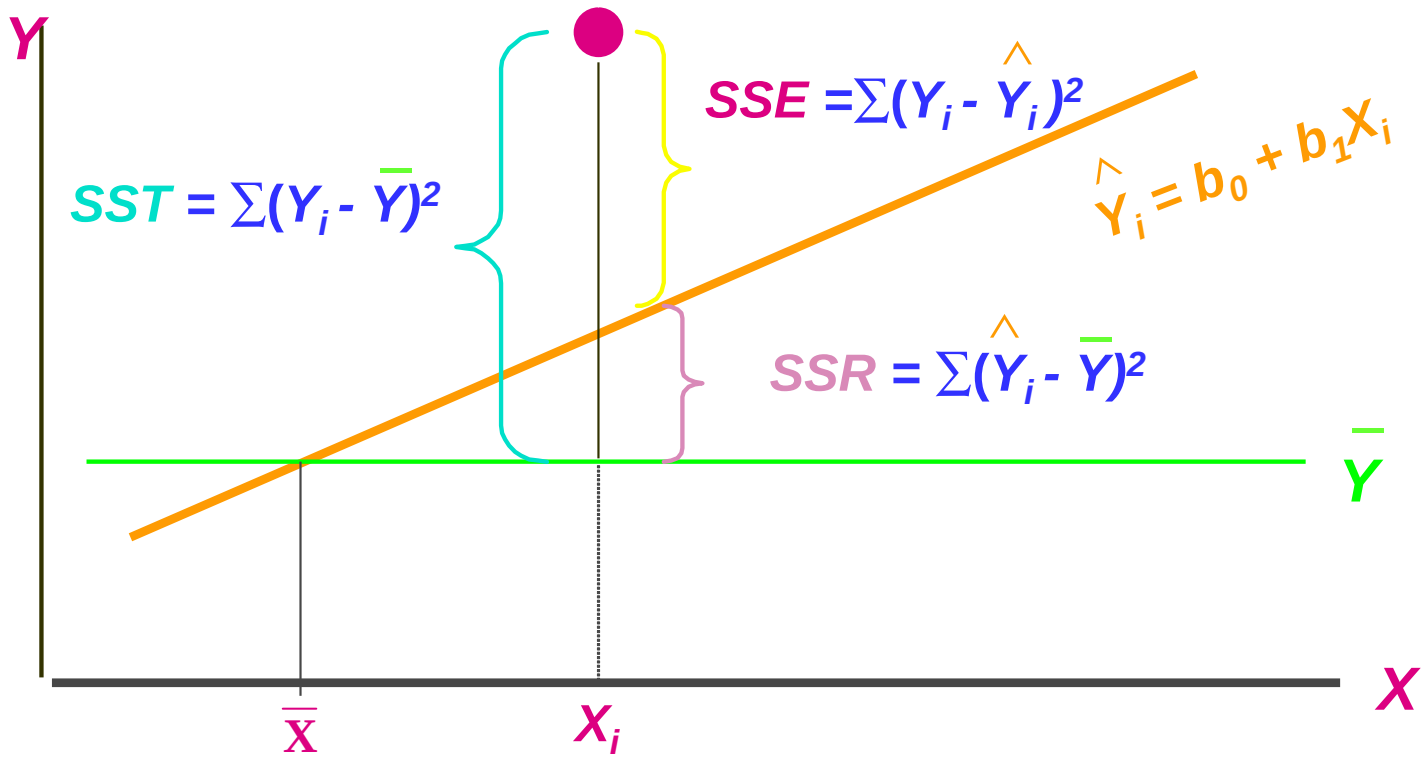
Assessing the Model

- In addition to determining the coefficients of the least squares line, we need to assess it to see how well it "fits" the data. They're based on the what is called sum of squares for errors (SSE).

$$
\begin{aligned}
SST &= \sum_{t=1}^{T} (y_t - \bar{y})^2 \text{ (total sum of squares)} \\
SSE &= \sum_{t=1}^{T} (\hat{y}_t - \bar{y})^2 \text{ (explained sum of squares)} \\
SSR &= \sum_{t=1}^{T} (u_t)^2 \text{ (residual sum of squares)}
\end{aligned}
$$

Linear regression

Assessing the Model



Linear regression

Assessing the Model

- How do we think about how well our sample regression line fits our sample data?
- Compute the fraction of the total sum of squares (SST) that is explained by the model, call this the **R-squared** of regression (Coefficient of Determination), to judge the adequacy of the regression model

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

- $0 \leq R^2 \leq 1$, represents the percent of the data that is the closest to the line of best fit.
- The higher the R^2, the more useful the model.
- For example, $R^2 = 0.850$, means that 85% of the total variation in y can be explained by the linear relationship between x and y (as described by the regression equation). The other 15% of the total variation in y remains unexplained.

Linear regression

Assessing the Model

- Another measure of how well the model fits the data is the Standard Error of the y estimate (measures the spread of the actual points around the fitted line)
- Since σ^2 is best estimated by s^2

$$s = \sqrt{\frac{SSE}{GL}} = \sqrt{\frac{\sum_{t=1}^{T}(y_t - \widehat{y}_t)^2}{T-2}} = \sqrt{SS_{yy} - \widehat{\beta}_1 SS_{xy}}$$

$$SS_{yy} = \sum_{t=1}^{T}(y_t - \overline{y})^2 = \sum_{t=1}^{T} y_t^2 - \frac{\left(\sum_{t=1}^{T} y_t\right)^2}{T}$$

$$SS_{xy} = \sum_{t=1}^{T} x_t y_t - \frac{\left(\sum_{t=1}^{T} x_t\right)\left(\sum_{t=1}^{T} y_t\right)}{T}$$

- Standard error for the β_1 and β_2 estimates