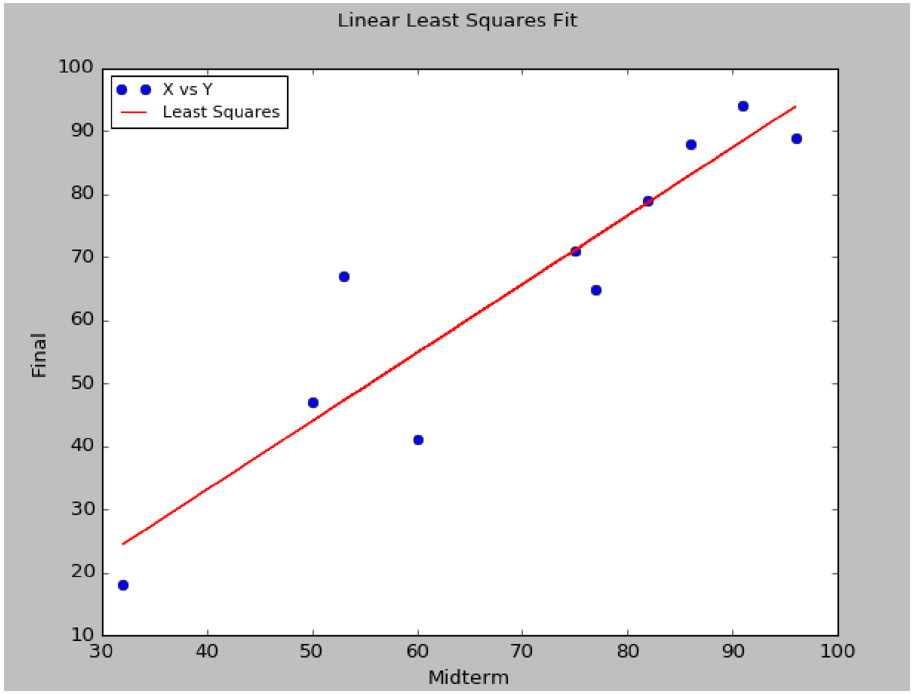
Linear Regression

- Engineers and Scientists are often called upon to predict the outcome of a certain event based on their past experiences
- Several practical applications involve a large number of variables that influence the analysis of possible outcomes
  - But understanding this technique with very limited number of variables is still very useful
- Statisticians like to split these variables into two specific categories:
  - Dependent variable that can be ***predicted***
  - Independent variable that influence or provide explanation of the outcome
  - If the relationship between the independent and dependent variable can be modeled as a straight line, we call this technique linear regression

Regression Analysis

Student	Midterm	Final
1	50	47
2	60	41
3	75	71
4	77	65
5	86	88
6	96	89
7	32	18
8	82	79
9	53	67
10	91	94
11	80	?



Regression Analysis

- Regression is a way of describing how one variable, the outcome (y), is numerically related to predictor variables (x)
- Regression analysis is the study of the relationship between the dependent and independent variables
  - Select a model to link the variables
  - Fit this model to the data
  - If the model fits well it can then be used to predict the value of the dependent variable

Several models are available for linear regression, for example:
Least Squares Approximation, Pearson Correlation Coefficient

Regression Analysis

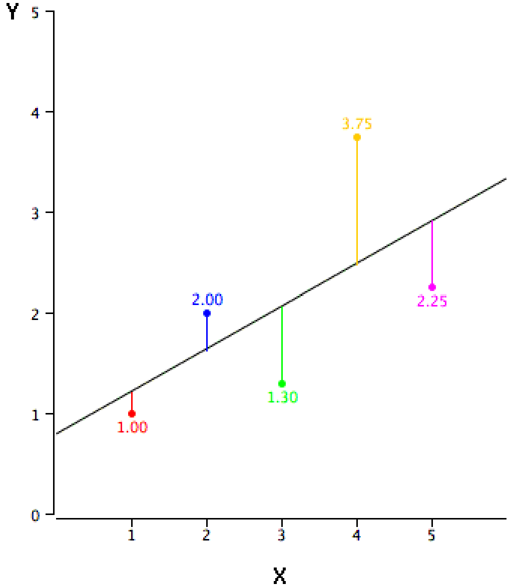
- In the simplest form of regression analysis, a single dependent variable y is related to a single independent variable x, and the relationship is modeled as a straight line

  $$y = mx + b$$

  - where *m* is the slope and *b* is the intercept
  - *m* and *b* are called linear regression coefficients
  - When we fit the model to the data we seek the line that is the biggest abstraction for our collection of *(x, y)* data points
  - Statistic is called Correlation Coefficient
  - Measures the strength of the linear relationship between y and x

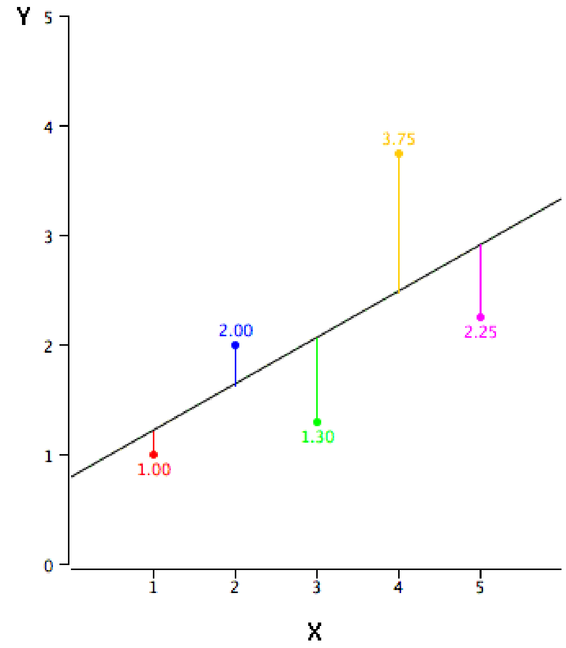
Method of Least Squares

- The method of **least squares** is a standard approach to the approximate solution of over-determined systems *(i.e., sets of equations in which there are more equations than unknowns)*
- "Least squares" means that the overall solution minimizes the sum of the squares of the errors made in the results of every single equation
  - The best fit in the least-squares sense minimizes the sum of squared residuals, a residual being the difference between an observed value and the fitted value provided by a model



Method of Least Squares

- The most important application is in data fitting
- To make a correct prediction of the dependent variable, we will need to use the "coefficient of determination"
  - Typically a number between 0 & 1 (closer it is to 1, the more certainty that dependency exists)



Regression using Least Squares

Least Squares Approximation

$$
\begin{aligned}
s_x &= \sum_{i=0}^{n-1} x_i \\
s_y &= \sum_{i=0}^{n-1} y_i \\
s_{xy} &= \sum_{i=0}^{n-1} x_i y_i \\
s_{xx} &= \sum_{i=0}^{n-1} x_i^2 \\
m &= (s_x s_y - s_{xy} n)/(s_x^2 - s_{xx} n) \\
b &= (s_y - s_x m)/n
\end{aligned}
$$

m and ***b*** are regression coefficients which are then used to find out new values of y (*new y is called f(x)*)

f(x) = mx + b

Sum of squared residuals (r):

$$\sum_{i=0}^{n-1} (y_i - f(x_i))^2$$

Total sum of squares (t):

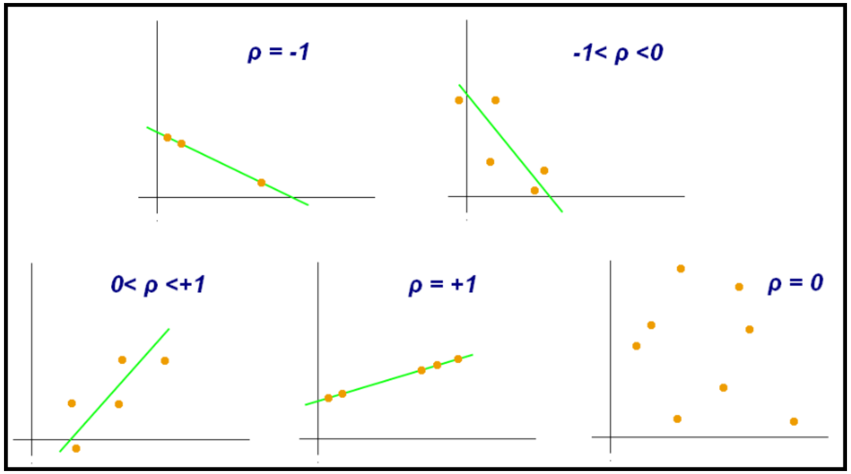
$$\sum_{i=0}^{n-1} (y_i - \bar{y})^2,$$

Coefficient of determination:

(between 0 & 1): 1-(r/t)

Regression using Pearson

- The **Pearson product-moment correlation coefficient** (sometimes referred to as **Pearson's *r***) is a measure of the linear dependence between two variables *X* and *Y*
  - Gives a value between +1 and −1 inclusive
    - 1 is total positive linear correlation
    - 0 is no linear correlation
    - −1 is total negative linear correlation



  - Measure of the strength of a linear association between two variables. It attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, *r*, indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit).

Regression using Pearson

Pearson Correlation Coefficient

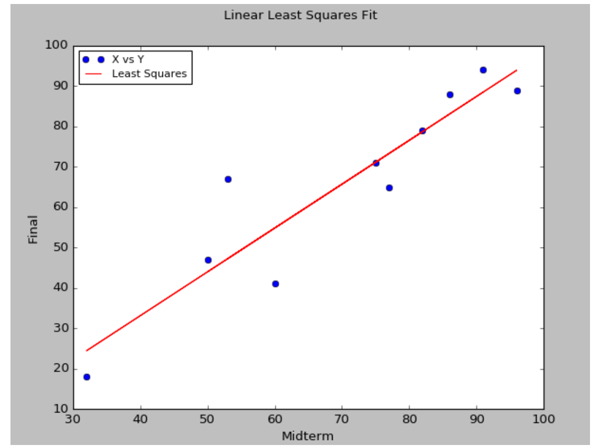
$$r = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sqrt{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} \sqrt{\sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}}}$$

OR

$$r = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \overline{y})^2}}$$

Example

Student	Midterm	Final
1	50	47
2	60	41
3	75	71
4	77	65
5	86	88
6	96	89
7	32	18
8	82	79
9	53	67
10	91	94
11	80	?




```
Least Squares Method
--------------------
Coefficients: m = 1.085155      b = -10.277906
Sum of Squared Residuals:   780.373418
Total Sum of Squares:  5282.900000
Coefficient of determination:     0.852283

Pearson Method
--------------
Pearson Correlation Coefficient:     0.923192
Predicted value of y (for x=80):    76.53
```