Linear Regression (PHYS3600ID)

Outline

- What is linear regression
- How do we make a linear regression?
- Assessing the model
- Demo in Python

What is regression?

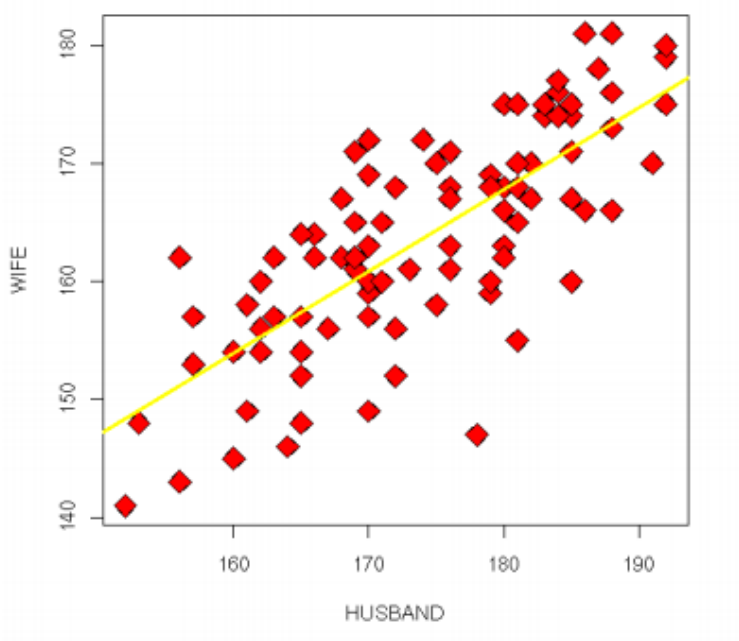
*In statistics, **regression analysis** includes any techniques for modeling and analyzing several variables, when the focus is on the **relationship** between a **dependent variable** and one or more **independent variables**.*

- Regression is an approach/method/algorithm used in finding the relationship between explanatory variables (or independent variables) and our variable of interest (or dependent variable).

Scatter Plots and Correlation

- A scatter plot (or scatter diagram) is used to show the relationship between two variables
- Correlation analysis is used to measure strength of the association (linear relationship) between two variables
  - Only concerned with strength of the relationship
  - **No causal effect is implied**

Height data: are wife/husband heights correlated?

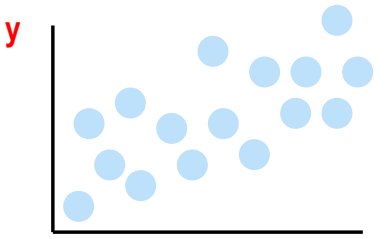


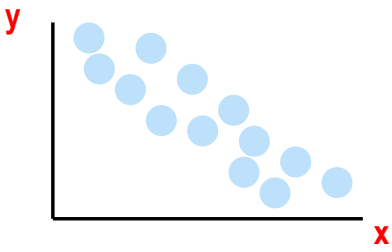
Regression vs Decision Trees

- Decision trees:
  - no assumption on variable dependence (target variable vs features);
  - simply seek to isolate concentrations of cases with like-valued target measurements
- Regressions are parametric models:
  - assume a specific association structure between input features and target
  - predict cases using a mathematical equation involving values of the input variables.

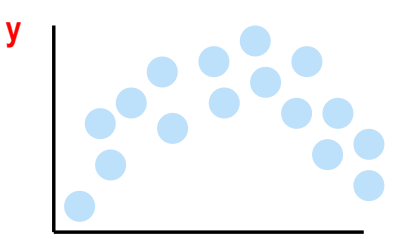
Scatter Plot Examples

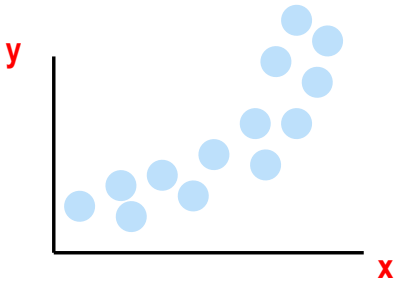
Linear relationships





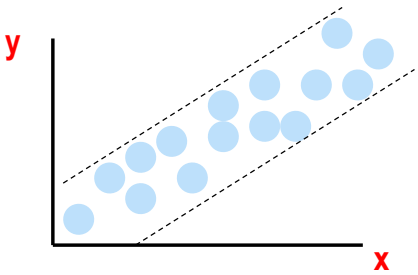
Curvilinear relationships

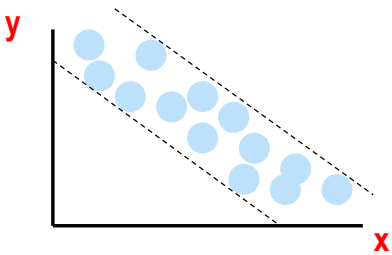




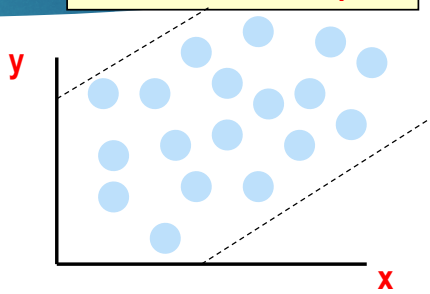
Scatter Plot Examples (continued)

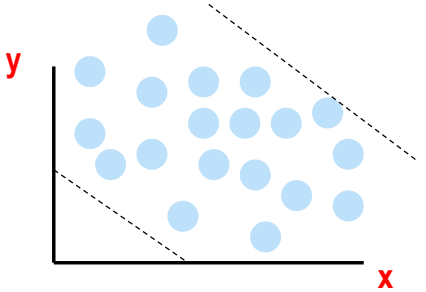
Strong relationships



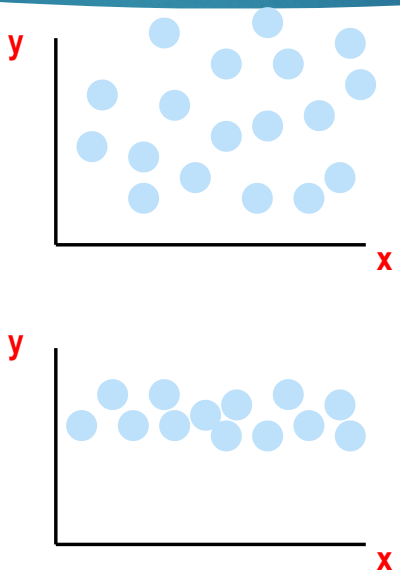


Weak relationships





Scatter Plot Examples (continued)



What is linear regression?

- It's a particular case of regression in which we hypothesize that the relationship between independent variable(s) and the outcome is LINEAR.
- We are interested in finding out how much our variable of interest (Y), will increase (or decrease) with a change in each of our explanatory/indep. variables (X_1, X_2, …)

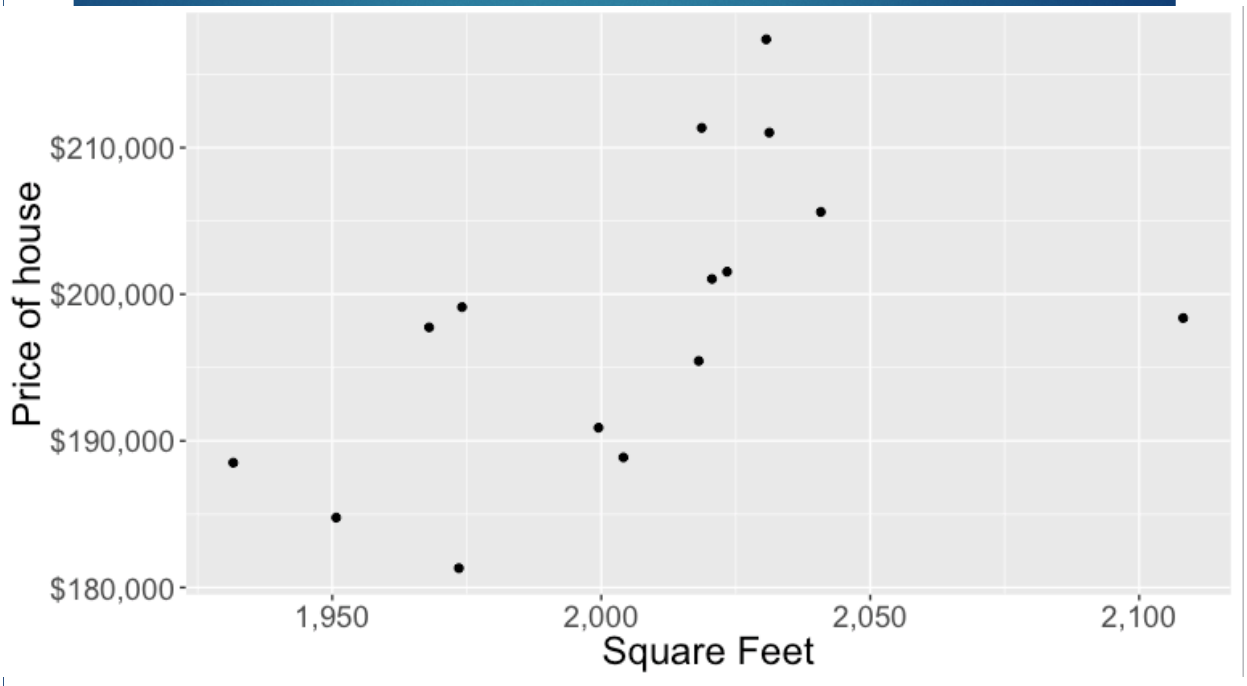
11

Example = House prices

- What are some factors that influence house prices and how much do they influence them?
  - Price (Y) related to:
    - Number of bedrooms (X_1)
    - Number of bathrooms (X_2)
    - Square footage (X_3)
    - Lot size (X_4)
    - Etc.

12

Creating a Linear Regression



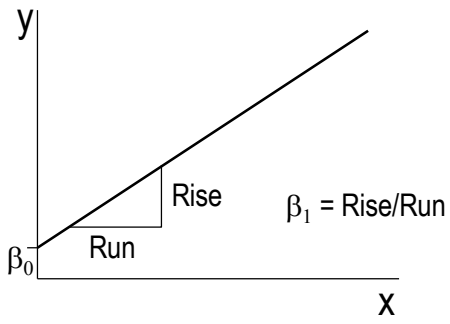
Drawing a line

- Slope intercept

$$y = \beta_0 + \beta_1 x$$

β_0 = y-intercept

β_1 = slope of the line



Linear Regression Analysis

- Analysis of the strength of the ***linear relationship*** between predictor (independent) variables and outcome (dependent/criterion) variables.
- In two dimensions (one predictor, one outcome variable) data can be plotted on a scatter diagram.

$$E(y) = \beta_0 + \beta_1 (x)$$

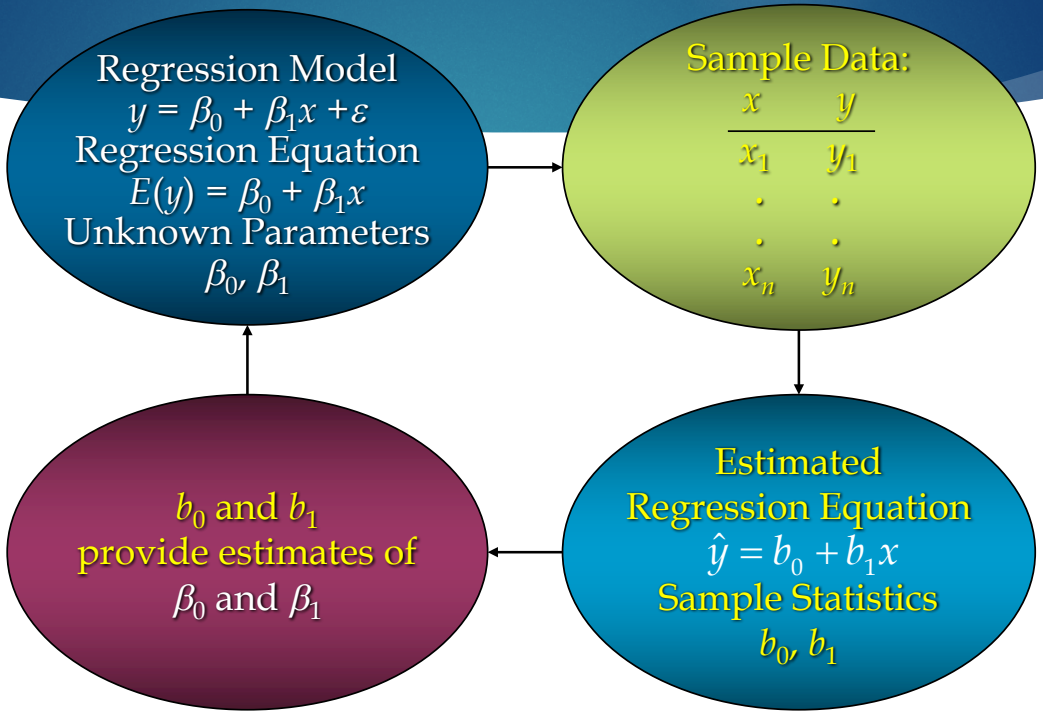
Expected value of y (outcome)

Intercept Term

coefficient

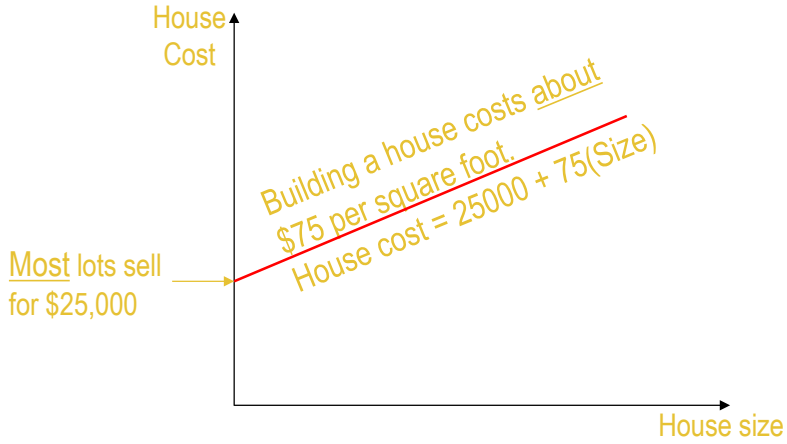
Predictor variable

Estimation Process



The Model

The model has a deterministic and a probabilistic component



The Model

However, house cost vary even among same size houses.

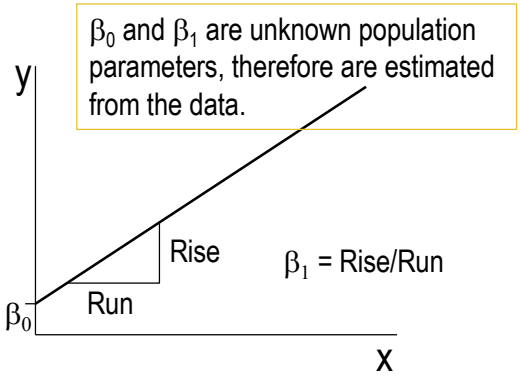
Since cost behave unpredictably, we add a random component.

The Model

- The first order linear model

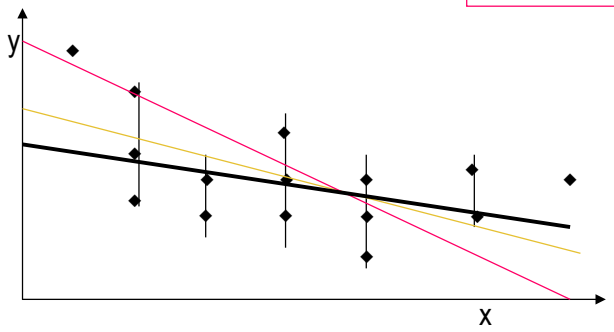
$$y = \beta_0 + \beta_1 x + \varepsilon$$

y = dependent variable
x = independent variable
β_0 = y-intercept
β_1 = slope of the line
ε = error variable

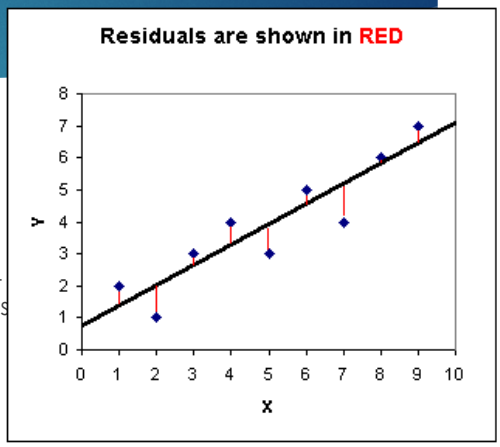


Finding the regression line (estimating coefficients)

Question: What should be considered a good line?



Least Squares Regression Line



- The **Least Squares Regression Line** is the line which minimizes the sum of the square or the error of the data points. It is an averaging line of the data.
- The **actual data points** (x,y) are the **blue** dots.
- The **Least Squares Regression Line** of the dependent (y) variable based on the independent (x) variable is shown in **black**.
- The **errors (residuals)** are the vertical distances between the observed values of y and the predictions of the "line of best fit," which are shown in **red**

Residual = Observed Value - Predicted Value

(residuals should be normally distributed around 0)

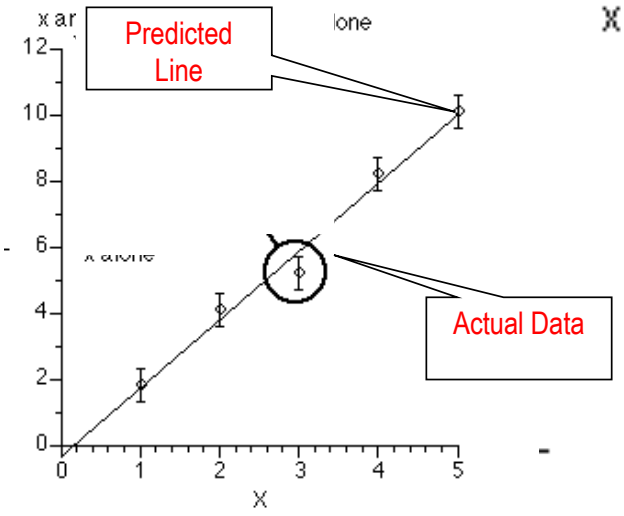
Otherwise, assumptions about linear regression do not hold.

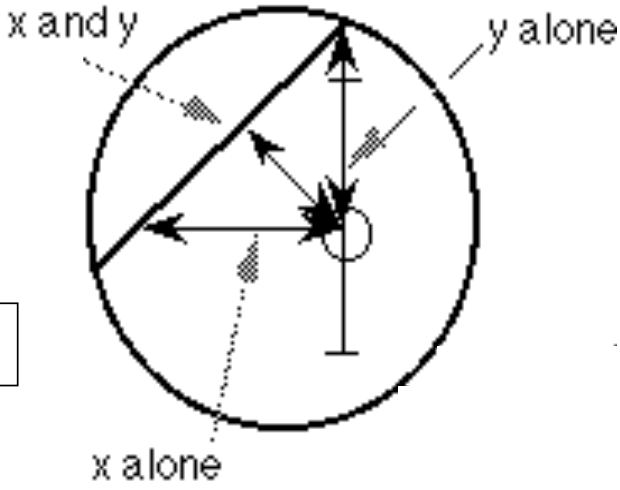
The Least Squares (Regression) Line

A good line is one that minimizes the sum of squared differences between the points and the line.

Least Squares Estimation Procedure

- The sum of the vertical deviations (y axis) of the points from the line is minimal.

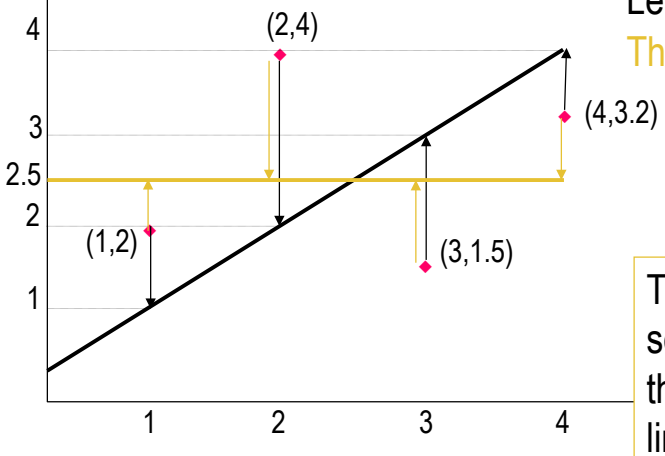




The Least Squares (Regression) Line

Sum of squared differences = $(2 - 1)^2 + (4 - 2)^2 + (1.5 - 3)^2 + (3.2 - 4)^2 = 6.89$

Sum of squared differences = $(2 - 2.5)^2 + (4 - 2.5)^2 + (1.5 - 2.5)^2 + (3.2 - 2.5)^2 = 3.99$



Let us compare two lines

The second line is horizontal

The smaller the sum of squared differences the better the fit of the line to the data.

The Estimated Coefficients

To calculate the estimates of the slope and intercept of the least squares line , use the formulas:

$$b_1 = \frac{SS_{xy}}{SS_{xx}}$$

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

$$SS_{xy} = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}$$

$$SS_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = (n-1)s_x^2$$

To extended to multiple independent variables and betas, use matrix multiplication:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Example: Finding the least squares line

- A car dealer wants to find the relationship between the odometer reading and the selling price of used cars.
- A random sample of 100 cars is selected, and the data recorded.
- Find the regression line.

Car	Odometer	Price
1	37388	14636
2	44758	14122
3	45833	14016
4	30862	15590
5	31705	15568
6	34010	14718
.	.	.
.	Independent variable x	Dependent variable y
.		

Linear Regression Line

- Solution
  - Solving by hand: Calculate a number of statistics

$\bar{x} = 36{,}009.45;$ $\quad SS_{xx} = \sum x_i^2 - \dfrac{\left(\sum x_i\right)^2}{n} = 43,528,690$

$\bar{y} = 14{,}822.823;$ $\quad SS_{xy} = \sum (x_i y_i) - \dfrac{\sum x_i \sum y_i}{n} = -2,712,511$

where n = 100.

$$b_1 = \frac{SS_{xy}}{(n-1)s_x^2} = \frac{-2,712,511}{43,528,690} = -.06232$$

$$b_0 = \bar{y} - b_1\bar{x} = 14,822.82 - (-.06232)(36,009.45) = 17,067$$

$$\hat{y} = b_0 + b_1 x = 17{,}067 - .0623x$$

27

Simple Linear Regression

- Regression analysis gives us very useful information about the relation between two variables.
- Among other things, it will tell us, for a certain confidence level, **the effect of a change in the explanatory variable on the explained variable.**
- If there is a well established relation between number of policemen in an area and the crime rate, it would prove very useful to know, for example, what would be the impact on the crime rate of hiring a new policeman.

28

Assessing the Model

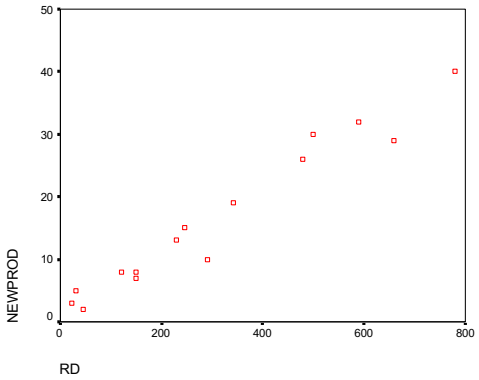
- The least squares method will produce a regression line whether or not there is a linear relationship between x and y.
- Consequently, it is important to assess **how well the linear model fits the data.**
- Several methods are used to assess the model.

Example: Investment in R&D and introduction of new products

- How does investment in R&D affects the number of new products developed? We can postulate the following relation:

  *# of new products = α + β*Investment in R&D + u*

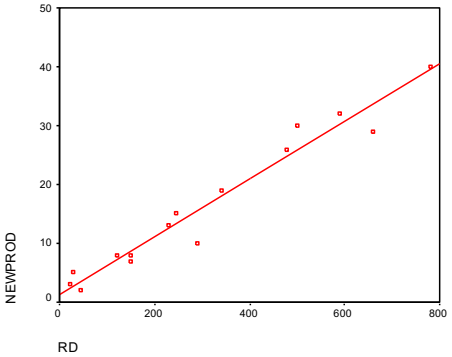
  Let us look at the scatter plot:



- Both the explained and the explanatory variables are not bounded (just by zero on the left side). But they can take any positive number.

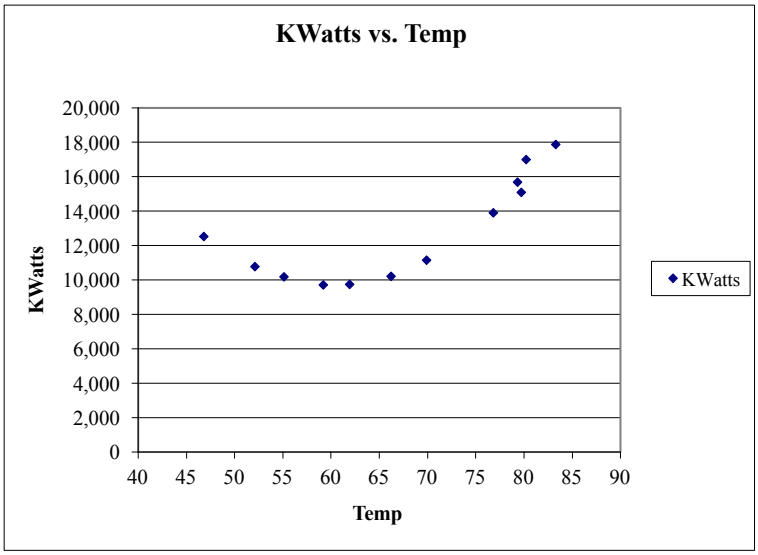
Example: Investment in R+D and introduction of new products

- It makes sense to assume a linear relation between X and Y in this case.
- The estimate for β = 0.049
- This tells us that in order to increase the number of new products in one unit, we need to invest a little bit more than 20 monetary units in R&D.
- If a company invests 1000 in R&D, we would predict this company to develop around 49 new products



Example: Kwatts vs. Temp

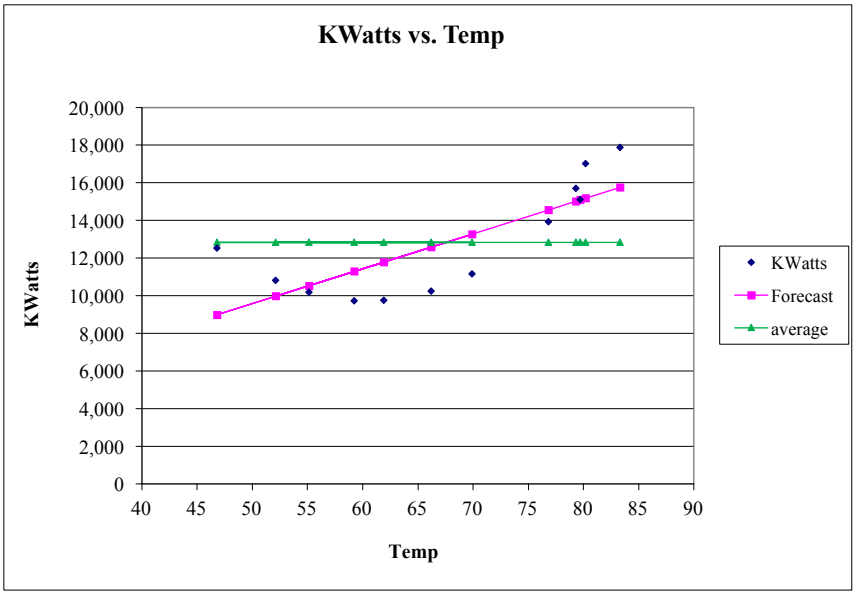
Temp	Kwatts
59.2	9,730
61.9	9,750
55.1	10,180
66.2	10,230
52.1	10,800
69.9	11,160
46.8	12,530
76.8	13,910
79.7	15,110
79.3	15,690
80.2	17,020
83.3	17,880



Linear Regression Results

Let X = Temp, Y = Kwatts

Y = 319.04 + 185.27 X



Coefficient of Determination

- How "strong" is relationship between predictor & outcome? (Fraction of observed variance of outcome variable explained by the predictor variables).
- Relationship Among SST, SSR, SSE

SST = SSR + SSE

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

where:

SST = total sum of squares

SSR = sum of squares due to regression

SSE = sum of squares due to error

Coefficient of Determination (R^2)

(How well independent variable explains dependent variable)

$R^2 = SSR/SST$

where:

SSR = sum of squares due to regression

SST = total sum of squares

Kwatts vs. Temp Example

	df	SS
Regression	1	58784708.31
Residual	10	38696916.69
Total	11	97481625

$R^2 = 0.603033734$

Coefficient of Determination

- It measures the ***strength of the linear relationship***.
- It's bounded between 0 and 1

$$R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2}, \text{ or } R^2 = 1 - \frac{SSE}{\sum (y_i - \bar{y})^2}$$

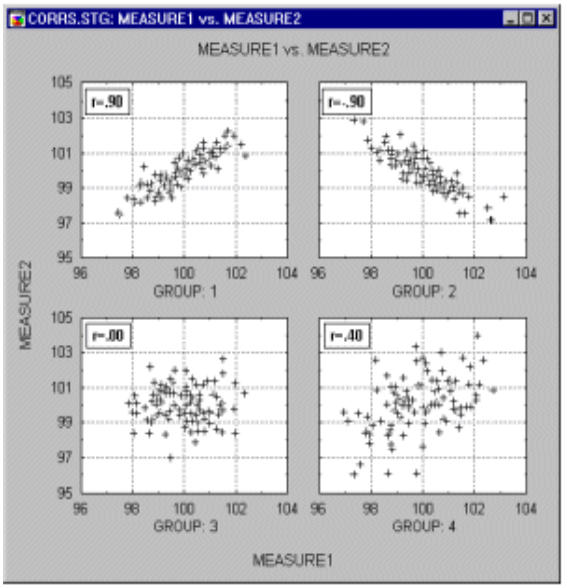
Coefficient of Determination

- If R^2 has a value of 0.6, this means 60% of the variation in the auction selling prices (y) is explained by your regression model. The remaining 40% is ***unexplained***, i.e. due to error.
- Unlike the value of a test statistic, the ***coefficient of determination*** does **not** have a ***critical value*** that enables us to draw conclusions.
- In general the higher the value of R^2, the ***better*** the model fits the data.
- $R^2 = 1$: Perfect match between the line and the data points.
- $R^2 = 0$: There is no (linear) relationship between x and y.

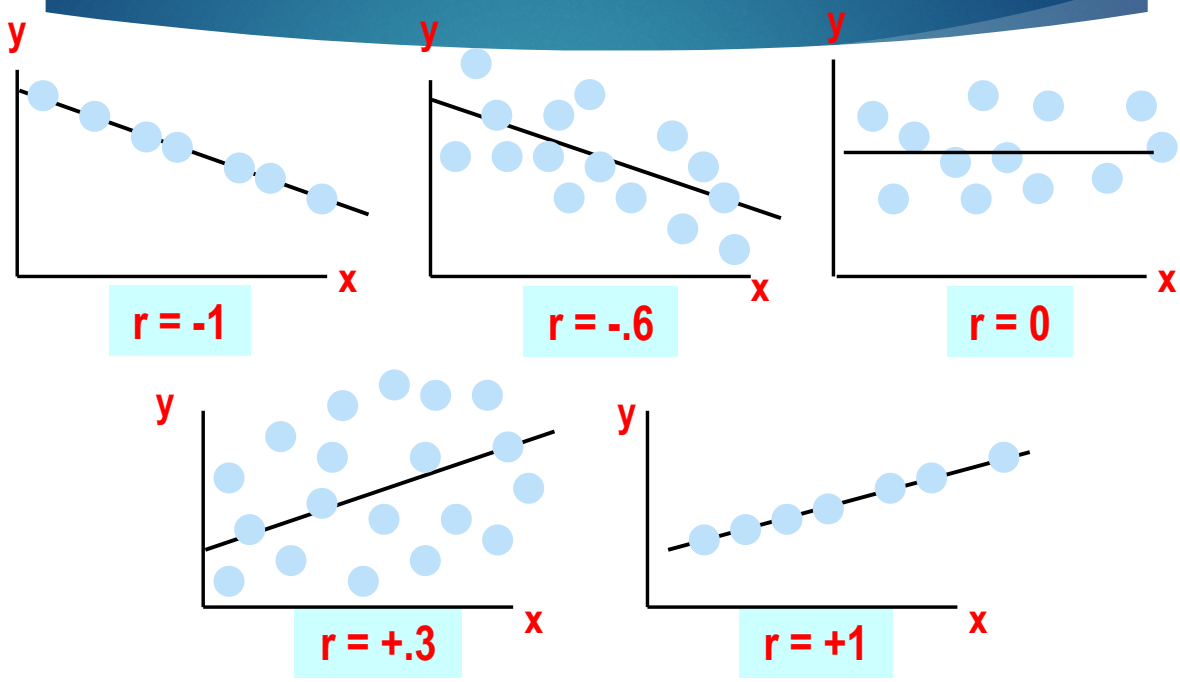
(Pearson) Correlation coefficient

It's a different quantity ρ, called the ***linear correlation coefficient***, which measures the strength and the direction of a linear relationship between two variables.

- Correlation coefficient (ρ: from -1 to 1)

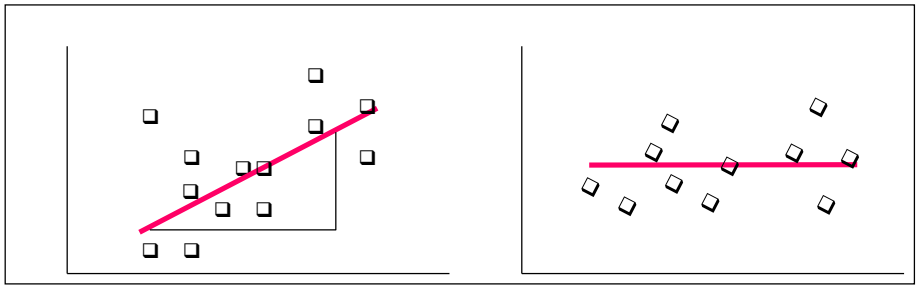


Examples of Approximate r (estimate of rho) Values



Sanity Check: Testing the slope

- When no linear relationship exists between two variables, the regression line should be horizontal.



Linear relationship.
Different inputs (x) yield different outputs (y).

The slope is not equal to zero

No linear relationship.
Different inputs (x) yield the same output (y).

The slope is equal (or close) to zero

Regression Diagnostics

- The three conditions required for the validity of the regression analysis are:
  - the error variable is normally distributed.
  - the error variance is constant for all values of x.
  - The errors are independent of each other.
- How can we diagnose violations of these conditions?

Residual Analysis

- Examining the residuals help detect violations of the required conditions.
  - Checking to see if residuals (actual value – predicted value) are normally distributed with mean 0.