Regression

what is regression?

re·gres·sion
/rəˈgreSH(ə)n/

noun
noun: **regression**; plural noun: **regressions**

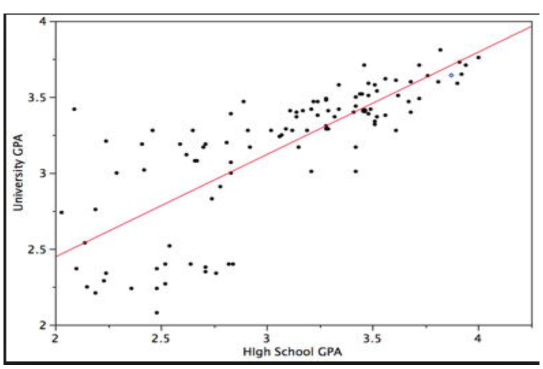
1. a return to a former or less developed state.
   - a return to an earlier stage of life or a supposed previous life, especially through hypnosis or mental illness, or as a means of escaping present anxieties.
     "regression therapy"
   - a lessening of the severity of a disease or its symptoms.
     "he seemed able to produce a regression in this disease"
2. STATISTICS
   a measure of the relation between the mean value of one variable (e.g., output) and corresponding values of other variables (e.g., time and cost).

Kinds of Regression...

- Simple Linear regression
  - predict values of Y given values of X
  - figure out: y = mx + b $y = \beta_0 + \beta_1 x + \varepsilon,$
- Multiple Linear Regression
  - predict Y based on multiple x factors
  - compute: $y_i = \beta_0 1 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$
- Polynomial Regression
  - relationship not linear $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon.$
- Logistic Regression
  - outcome is a category, not a value
  - e.g. Pass/Fail, Win/Lose, Buy/Sell

Kinds of Regression...

- Simple Linear regression
  - predict values of Y given values of X



y = mx + b

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

Scenario

- You are a sales manager – want to increase sales
- You know there are many factors that influence sales:
  - weather
  - new products
  - social media

Which factors matter most?
Which can you ignore?
How much will sales go up or down?

Thinking Like a Quant

- dependent variable –
  - the factor you are trying to predict
- independent variables
  - factors you suspect have an impact on the dependent variable

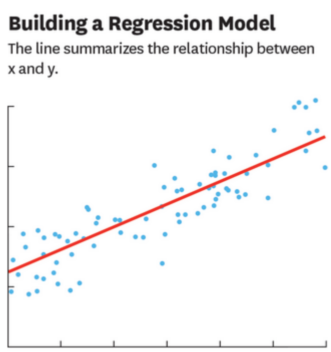
Gather the data and plot it

Y Axis:
dependent
variable

Y Axis:
Independent
variable

Gather the data and plot it

Y Axis:
dependent
variable



The red line is the best explanation of the relationship between the independent variable and the dependent variable

$Y = 200 + 5X + \text{error term}$

Error term reflects the fact that your equation is only an estimate of the true state of affairs.

$Y = 200 + 5X$

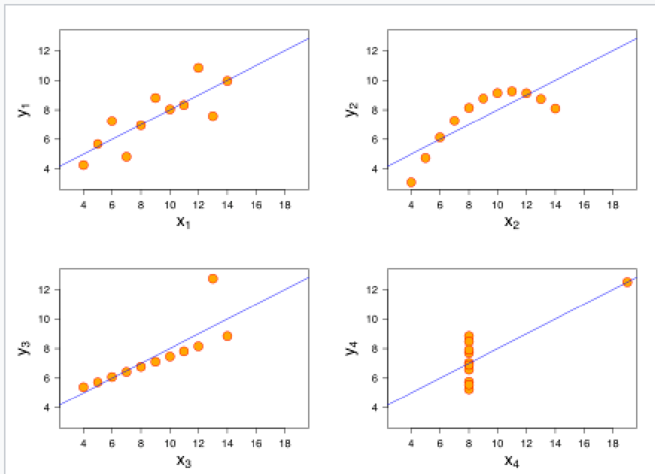
Two Approaches to Linear Regression

- Closed form equation
  – directly compute the model parameters using Least Squares Regression
- Iterative Approach
  – use gradient descent (GD) to tweak parameters and converge to the model parameters

Which is better?

Simple eh?

- Just compute values for y = mx+b



The data sets in the Anscombe's quartet are designed to have approximately the same linear regression line (as well as nearly identical means, standard deviations, and correlations) but are graphically very different. This illustrates the pitfalls of relying solely on a fitted model to understand the relationship between variables.

Use linear regression??

- Before attempting to fit a linear model, first determine whether or not there is a relationship between the variables of interest.
- Does not necessarily imply that one variable *causes* the other (for example, higher SAT scores do not *cause* higher college grades), but that there is some significant association between the two variables.
- One way: run a correlation where values near +1 or -1 indicate a relationship
- Another: Look at **Covariance**
  - **positive value : they vary together**
  - **negative value: they vary inversely**
  - **zero (or near) : no relationship**

Covariance

Covariance

- A descriptive measure of the linear association between two variables
  - positive value – direct or increasing relationship
  - negative – decreasing relationship
  - No comment about strength of relationship, only direction
  - correlation – measures the strength

Covariance formulae

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

Sample Covariance

$$\sigma_{xy} = \frac{\Sigma(x_i - \mu_x)(y_i - \mu_y)}{N}$$

Population Covariance

Are you measuring via a sample or the total population?

Example: Covariance example using sample covariance

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

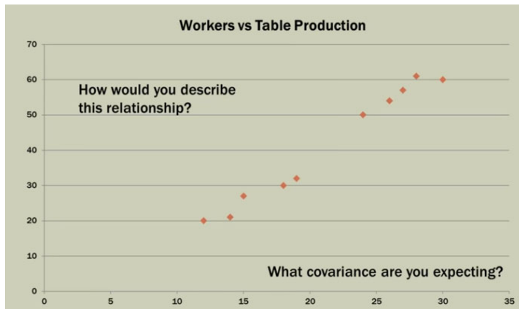
Sample Covariance

RISING HILLS MANUFACTURING

Rising Hills Manufacturing wishes to study the relationship between the number of workers, x, and number of tables produced, y, in its plant.

To do so it obtained 10 samples, each one hour in length, from the production floor.

x	y
12	20
30	60
15	27
24	50
14	21
18	30
28	61
26	54
19	32
27	57
$\bar{x} = 21.3$	$\bar{y} = 41.2$

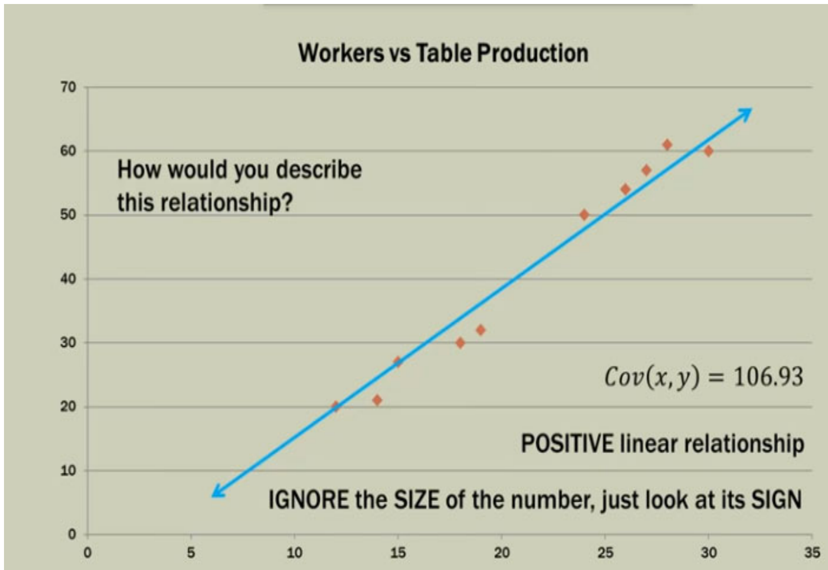


x	y	$(x_i - \bar{x})(y_i - \bar{y})$
12	20	197.16
30	60	163.56
15	27	89.46
24	50	23.76
14	21	147.46
18	30	36.96
28	61	132.66
26	54	60.16
19	32	21.16
27	57	90.06
$\bar{x} = 21.3$	$\bar{y} = 41.2$	$\Sigma = 962.4$

$$Cov(x,y) = s_{xy} = \frac{962.4}{n-1}$$

$$\frac{962.4}{9}$$

$$Cov(x,y) = 106.93$$



Exercise

- Write Python code

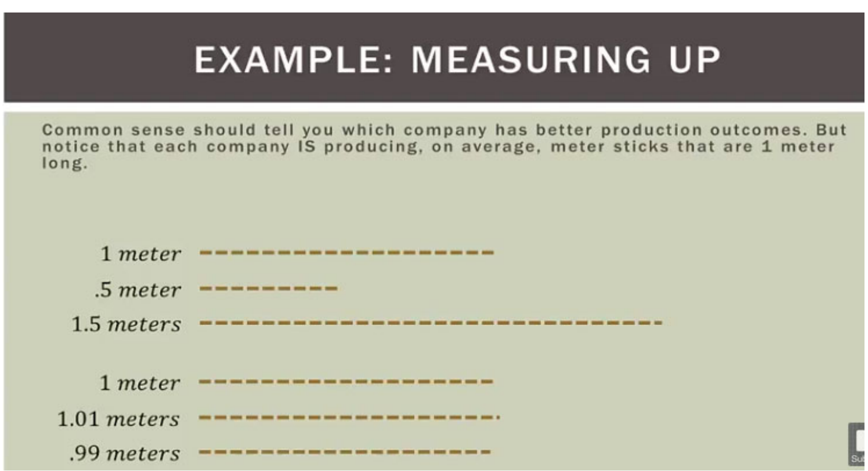
def covariance(...):

- Make it flexible so you can compute either sample or population covariance

Variance

Variance

- Variance measures SPREAD of a data set over the values



Common sense should tell you which company has better production outcomes. But notice that each company IS producing, on average, meter sticks that are 1 meter long.

What is the difference? VARIATION

1 *meter* -------------------

.5 *meter* ---------

1.5 *meters* ------------------------------

$$\bar{x} = \frac{1 + 0.5 + 1.5}{3} = 1\ meter$$

1 *meter* -------------------

1.01 *meters* -------------------

.99 *meters* -------------------

$$\bar{x} = \frac{1 + 1.01 + 0.99}{3} = 1\ meter$$

Variance Example

EXAMPLE: ENGINE CYLINDERS

When a standard car, truck, or similar engine is made the cylinders must be “bored” from a block of metal. The pistons must fit inside the cylinder VERY precisely. So yes the cylinders must be the correct diameter…but they ALSO must have a VARIANCE near zero.

Stock market

EXAMPLE: STOCK RETURNS AS %

Ford Motor Company vs. Whole Foods Markets

Calculation of variance

Variance and its square-root, the standard deviation, are both measures of the spread or variability in data. Two or more data sets could have the same mean, but very different variances.

Population Variance

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$$

Sample Variance

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1}$$

Exercise

- Write a Python function:
- def variance(..):
- That can handle both sample and population variance

Variance and its square-root, the standard deviation, are both measures of the spread or variability in data. Two or more data sets could have the same mean, but very different variances.

Population Variance

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$$

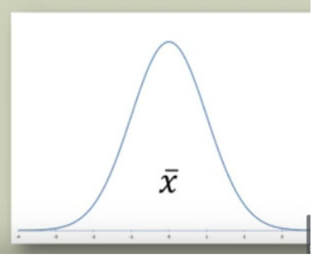
Sample Variance

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

Sampling from a population

When we take many samples of the same size from a population and then find the sample means, $\bar{x}$, those sample means follow the normal curve when placed in their own distribution.

The Sampling Distribution of $\bar{x}$



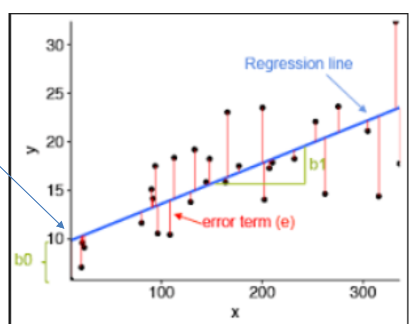
Let's Compute the Regression Line using

Closed Form Equation

Either way, we need to determine the best values for m and b in the line:

$$y = mx + b$$

b is y-intercept (also sometimes called the bias)



m is slope of the line

Simple Linear Regression Equation (Prediction Line)

The simple linear regression equation provides an estimate of the population regression line

Estimated (or predicted) Y value for observation i

Estimate of the regression intercept

Estimate of the regression slope

Value of X for observation i

$$\hat{Y}_i = b_0 + b_1 X_i$$

Regression Formula:

Y= a + bX

where slope of trend line is calculated as:

$$b_1 = \frac{\Sigma (x - \bar{x}) * (y - \bar{y})}{\Sigma (x - \bar{x})^2}$$

and the intercept is computed as:

$$b_0 = y - (b_1 * X)$$

Look familiar?

Computing Regression

Regression Formula:

Y= a + bX

where slope of trend line is calculated as:

$$b_1 = \frac{\Sigma (x - \bar{x}) * (y - \bar{y})}{\Sigma (x - \bar{x})^2}$$

Covariance(X,Y)

Variance(X,Y)

and the intercept is computed as:

$$b_0 = y - (b_1 * X)$$

also can think of as ‘a’ – the intercept

Exercise:

- x = [95,85,80,70,60]
- y = [85,95,70,65,70]

- Compute Slope and Y-Intercept
- y = a + bx
- b = covariance(X,Y) / variance(X,Y)
  - for sample population

Expected Answer: y = 26.78 + 0.6438 x

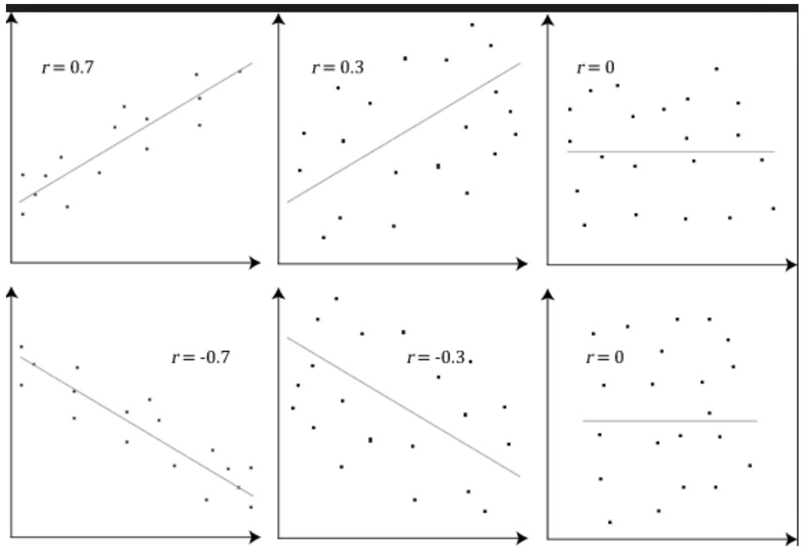
Pearson Correlation

- Covariance shows in what direction two variables are related

  but NOT how strong

- Pearson Correlation (r) shows how strong
- 1 = VERY Strong
- 0 = not related
- -1 = VERY Strong with inverse relationship

Pearson Correlation



compute pearson r via scipy learn

from scipy.stats.stats import pearsonr

correlation , pvalue = pearsonr(x,y)

Pearson Correlation Coefficient

- Pearson

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Note: the formula uses the population , not the sample calculation

where:

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

Exercise: show the SciPy and the formula yield the same pearson correlation

```
from scipy.stats.stats import pearsonr
correlation , pvalue = pearsonr(x,y)
```

Computing Regression

```
#easy breezy
from scipy.stats import linregress
linregress(x,y)
```

```
In [7]: #easy breezy
        from scipy.stats import linregress
        linregress(x,y)

Out[7]: LinregressResult(slope=0.6438356164383562, intercept=26.78082191780822, rvalue=0.6930525298193004, pvalue=0.194467490094009
        15, stderr=0.38664772840212874)
```

stderr : values < 3 are considered good.

Multiple Regression

(predicting based on multiple independent variables)

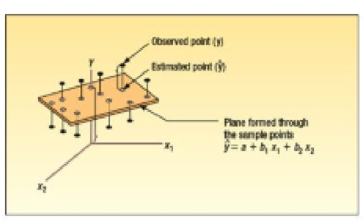
LO14-1 Use multiple regression analysis to describe and interpret a relationship between several independent variables and a dependent variable.

Multiple Regression Analysis

The general multiple regression equation with k independent variables is given by:

GENERAL MULTIPLE REGRESSION EQUATION $\hat{y} = a + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_kx_k$ **[14–1]**

- $X_1 \ldots X_k$ are the independent variables.
- a is the y-intercept
- b_1 is the net change in Y for each unit change in X_1 holding $X_2 \ldots X_k$ constant. It is called a partial regression coefficient or just a regression coefficient.
- Determining b_1, b_2, etc. is very tedious. A software package such as Excel or MINITAB is recommended.
- The least squares criterion is used to develop this equation.



Multiple Regression with n features

- y = $\theta_0 + \theta_1 X_1 + \theta_2 X_2 + \theta_3 X_3 \ldots \theta_n X_n$
- y : predicted value
- n : number of features
- x_j : feature j
- θ_j : the jth feature weight
- θ_0 : called the bias term – actually the y-intercept of the line

Multiple Regression

Given a data set $\{y_i,\, x_{i1}, \ldots, x_{ip}\}_{i=1}^n$ of *n* statistical units, a linear regression model assumes that the relationship between the dependent variable *y* and the *p*-vector of regressors **x** is linear. This relationship is modeled through a *disturbance term* or *error variable* ε — an unobserved random variable that adds "noise" to the linear relationship between the dependent variable and regressors. Thus the model takes the form

$$y_i = \beta_0 1 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^\mathsf{T} \boldsymbol\beta + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where $^\mathsf{T}$ denotes the transpose, so that $\mathbf{x}_i^\mathsf{T}\boldsymbol\beta$ is the inner product between vectors $\mathbf{x}_i$ and $\boldsymbol\beta$.

Often these *n* equations are stacked together and written in matrix notation as

$$\mathbf{y} = X\boldsymbol\beta + \boldsymbol\varepsilon,$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

$$X = \begin{pmatrix} \mathbf{x}_1^\mathsf{T} \\ \mathbf{x}_2^\mathsf{T} \\ \vdots \\ \mathbf{x}_n^\mathsf{T} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix},$$

$$\boldsymbol\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Note that to compute the regression line with multiple independent variables we need to compute the transpose of a matrix .

Exercise

- Write Python code to compute the transpose of a matrix
- Do not use built-in library function
  - m2 = m1.transpose()