

 July 21 - 24, 2021

 Garki Hospital Abuja

 Resource Person

Research Methodology Boot Camp

with Epi Info Training

Dr. Adamu Onu

MBBS, FWACP (FM)

MS Epidemiology & Biostatistics

PhD Public Health (Epidemiology)

Target Audience

Clinical Researchers, Post-Part 1 Residents, and Others

Important Information

- Limited slots are available on a first come, first served basis
- Laptop running Windows 10 required
- Organized as morning lecture sessions and afternoon hands on coaching sessions

For further details contact

Email: epimetrix@gmail.com

Phone: +234 803 474 9930



Highlights

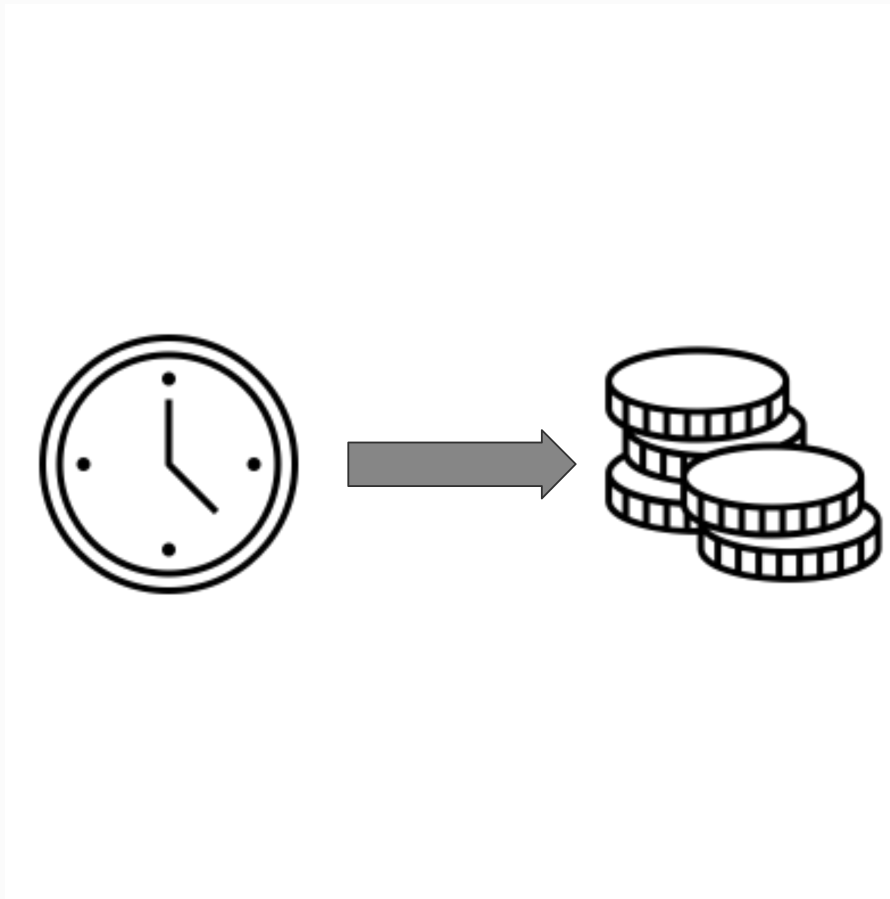
- Research Methodology
- Research Design
- Data Management
- Sample Size Calculations
- Test Statistics
- Interpretation of Results
- Report Writing
- Hands-on training sessions
- Statistical consulting sessions

Linear Regression

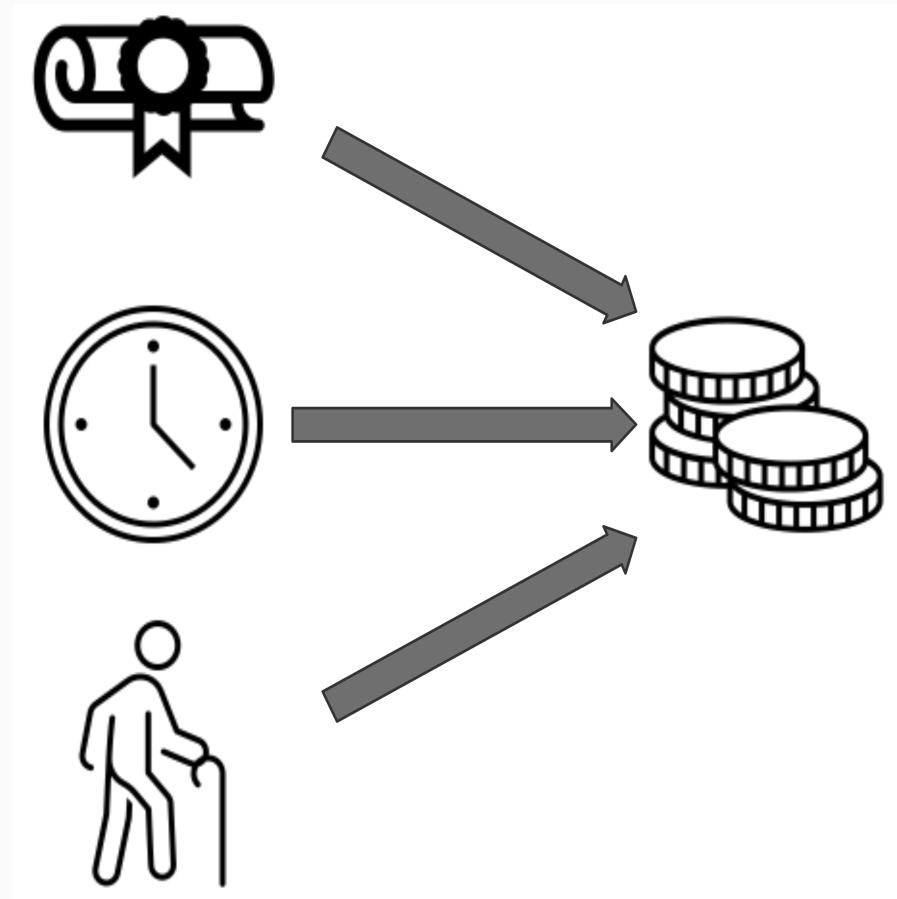
Linear regression

- Regression analysis is used to create a **model** that describes the relationship between a dependent variable and one or more independent variables.
- Depending on whether there are one or more independent variables, a distinction is made between **simple** and **multiple** linear regression analysis.

Simple linear regression



Multiple regression



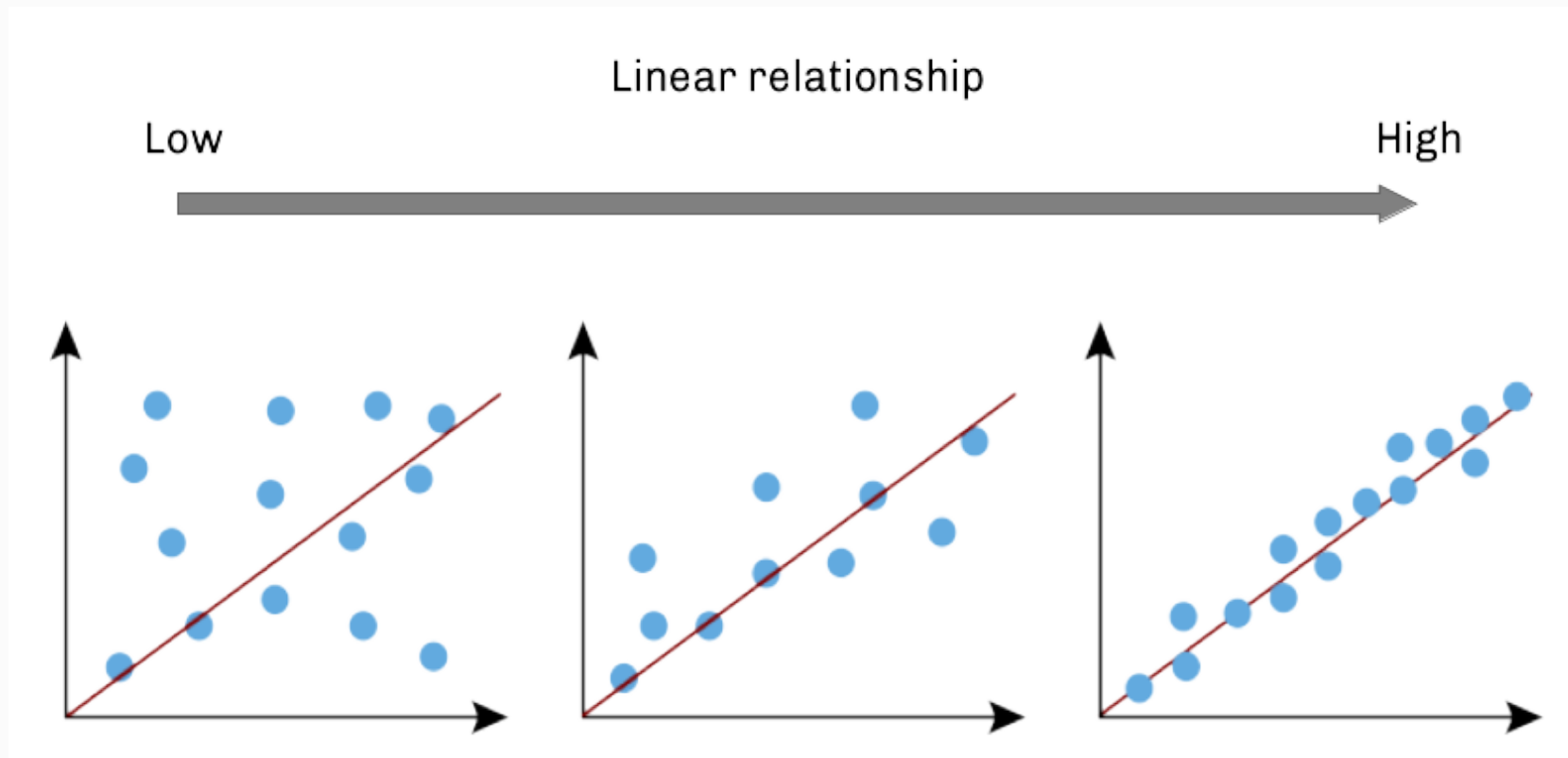
- In **simple linear regression**, the aim is to examine the influence of an independent variable on one dependent variable.
- In **multiple linear regression**, the influence of several independent variables on one dependent variable is analyzed.

Simple linear regression

- The goal of simple linear regression is to predict the value of a dependent variable based on an independent variable.
- The greater the proportion of the dependent variable's variance that can be explained by the independent variable, the more accurate is the prediction.
- Visually, the relationship between the variables can be shown in a scatter plot.

Simple linear regression

- The greater the linear relationship between the dependent and independent variables, the more the data points lie on a straight line.



Simple linear regression

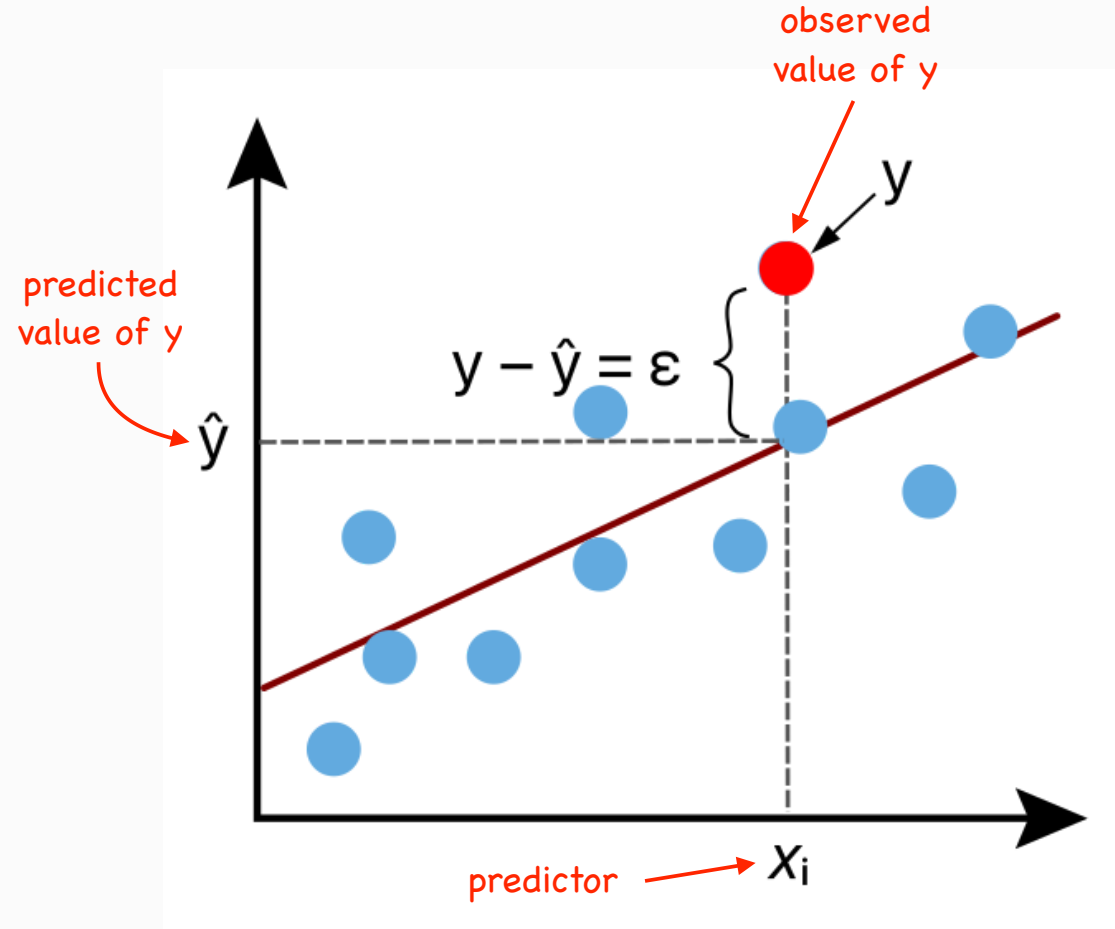
- The straight line which best describes the linear relationship between the dependent and independent variable.
- To determine this straight line, linear regression uses the **method of least squares**.

Simple linear regression

$$\hat{y} = b \cdot x + a$$

$$y = b \cdot x + a + \epsilon$$

- $\hat{y} \Rightarrow$ Estimated dependent variable
- $b \Rightarrow$ slope
- $x \Rightarrow$ independent variable
- $a \Rightarrow$ y intercept
- $\epsilon \Rightarrow$ error term



Error term

- If all points were exactly on one straight line, the estimate would be perfect.
- This is almost certainly not the case
- Therefore, a straight line must be found, which is as close as possible to the individual data points.
- The goal is to keep the error in the estimation as small as possible so that the distance between the estimated value and the true value is as small as possible.
- This distance or error is called the "residual" and is abbreviated as "e" (error).

OLS

- When calculating the regression line, an attempt is made to determine the regression coefficients (a and b) so that the sum of the squared residuals is minimal.
- The least-squares line, or the estimated regression line is the line $y = b \cdot x + a$ which minimizes the sum of the squared distances of the sample points from the line:

$$RSS = \sum_{i=1}^n (y_i - a - bx_i)^2$$

Estimation of the least-squares line

The coefficients of the regression line $y = bx + a$ are given by:

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

$$a = \frac{\sum_{i=1}^n y_i - b \sum_{i=1}^n x_i}{n}$$

Estimstion of the least-squares line

The coefficients of the regression line $y = bx + a$ are given by:

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

corrected sum of cross products

$$a = \frac{\sum_{i=1}^n y_i - b \sum_{i=1}^n x_i}{n}$$

Estimstion of the least-squares line

The coefficients of the regression line $y = bx + a$ are given by:

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

corrected sum of squares for x

$$a = \frac{\sum_{i=1}^n y_i - b \sum_{i=1}^n x_i}{n}$$

Coefficient of determination

- The coefficient of determination R^2 indicates how well the regression model predicts or explains the dependent variable.
- R^2 indicates how large the portion of the variance is that can be explained by the independent variables.
- The more variance can be explained, the better the regression model is.

$$R^2 = \frac{s_{\hat{y}}^2}{s_y^2}$$

variance of predicted values

variance of observed values

Multiple linear regression

- Multiple linear regression allows more than 1 independent variables to be considered.
- The goal is to estimate a variable based on several other variables.
- The variable to be estimated is called the dependent variable.
- The variables that are used for the prediction are called independent variables (predictors).

Multiple linear regression

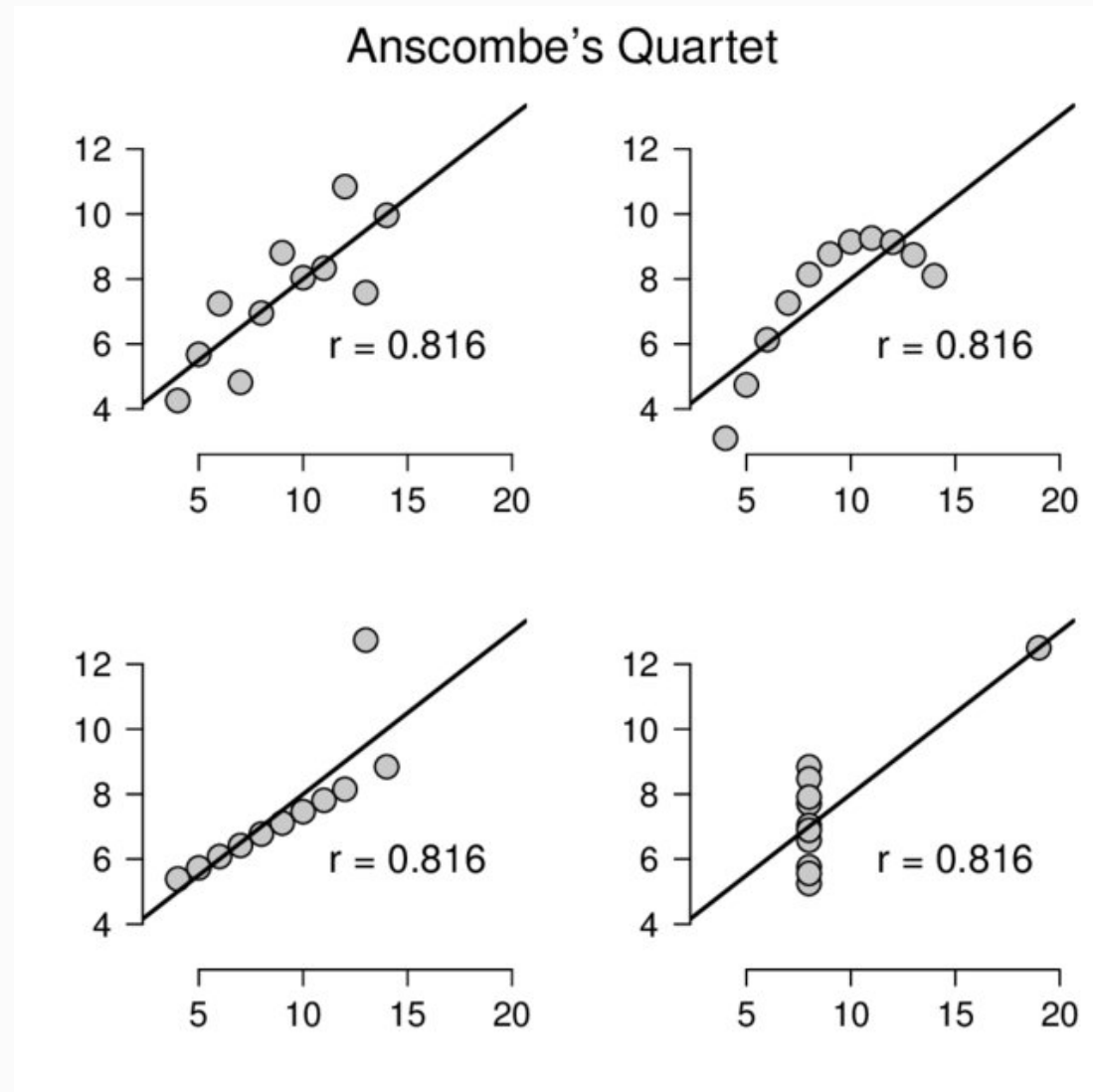
$$\hat{y} = b_1 \cdot x_1 + b_2 \cdot x_2 + \cdots b_n \cdot x_n + a$$

- The coefficients are interpreted similarly to the linear regression equation.
- If all independent variables are 0, the resulting value is a .
- If an independent variable changes by 1 unit, the coefficient indicates by how much the dependent variable changes.
 - So if the independent variable x_j increases by 1 unit, the dependent variable y increases by b_j .

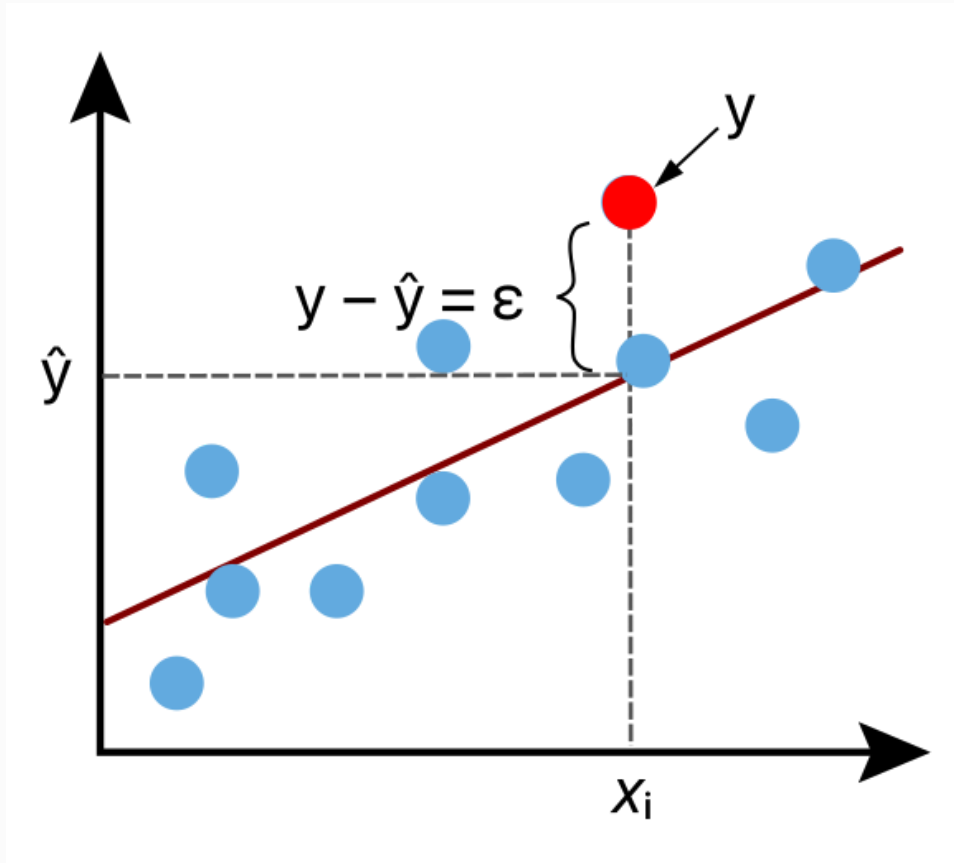
Assumptions of Linear Regression

- In order to interpret the results of the regression analysis meaningfully, certain conditions must be met.
 - **Linearity**: There must be a linear relationship between the dependent and independent variables.
 - **Homoscedasticity**: The residuals must have a constant variance.
 - **Normality of residuals**: Normally distributed error
 - No **Multicollinearity**: No high correlation between the independent variables

Linearity



Homoscedasticity

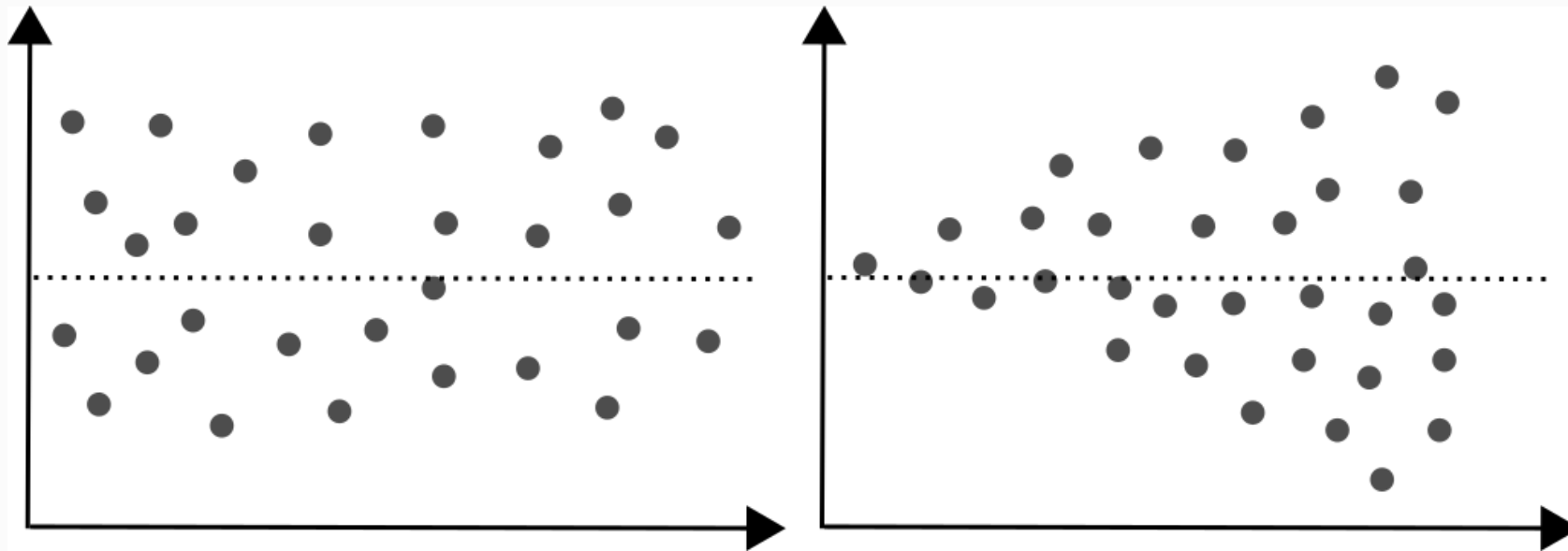


$$y = b_1 \cdot x_1 + b_2 \cdot x_2 + \dots + b_n \cdot x_n + a + \epsilon$$

The regression model never exactly predicts the dependent variable, there is always an error.

The error must have a constant variance over the predicted range.

- The dependent variable is plotted on the x-axis and the error on the y-axis.
- When homoscedasticity is present the error should scatter evenly over the entire range.




Normality of residuals

- ε must be normally distributed.
- This is examined graphically using the QQ plot
- Alternatively the Kolmogorov-Smirnov test or the Shapiro-Wilk test can be used.
 - If the p value > 0.05 , there is no deviation of the data from the normal distribution

Multicollinearity

- Multicollinearity means that two or more independent variables are strongly correlated with one another.
- The problem with multicollinearity is that the effects of each independent variable cannot be clearly separated from one another.
- Given $\hat{y} = b_1x_1 + b_2x_2 + \dots + b_nx_n + a$
 - $\hat{x}_1 = b_2x_2 + \dots + b_nx_n + a$, and $\hat{x}_2 = b_2x_2 + \dots + b_nx_n$
- A high correlation between x_1 and x_2 , makes it difficult to determine b_1 and b_2 – the regression model becomes unstable

variance inflation factor


$$VIF = \frac{1}{1 - R^2}, \quad \text{warning: } VIF > 10$$

Significance testing

1. Significance test for the whole regression model
2. Significance test for the regression coefficients

Significance test for the regression model

- The null hypothesis is that the R^2 in the population is zero, $H^0 : \rho^2 = 0$
- To confirm or reject the null hypothesis, the following F-test is calculated:

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - k - 1}{k}$$

- The numerator df are k (number of predictors in the regression) and the denominator df are $n - k - 1$, where n = number of data points

Significance test for the regression coefficients

- Which variables have a significant contribution to the prediction of the dependent variable?
- This is determined by checking whether the regression coefficients also differ from zero in the population.

$$t = \frac{b_i}{s_{b_i}}$$

- where b_i is the i th regression coefficient and s_{b_i} is the standard error of b_i .
- This test statistic is t-distributed with the *df* of $n - k - 1$.

Example linear regression

weight	height	age	gender
79	1.80	35	male
69	1.68	39	male
73	1.82	25	male
95	1.70	60	male
82	1.87	27	male
55	1.55	18	female
69	1.50	89	female
71	1.78	42	female
64	1.67	16	female
69	1.64	52	female

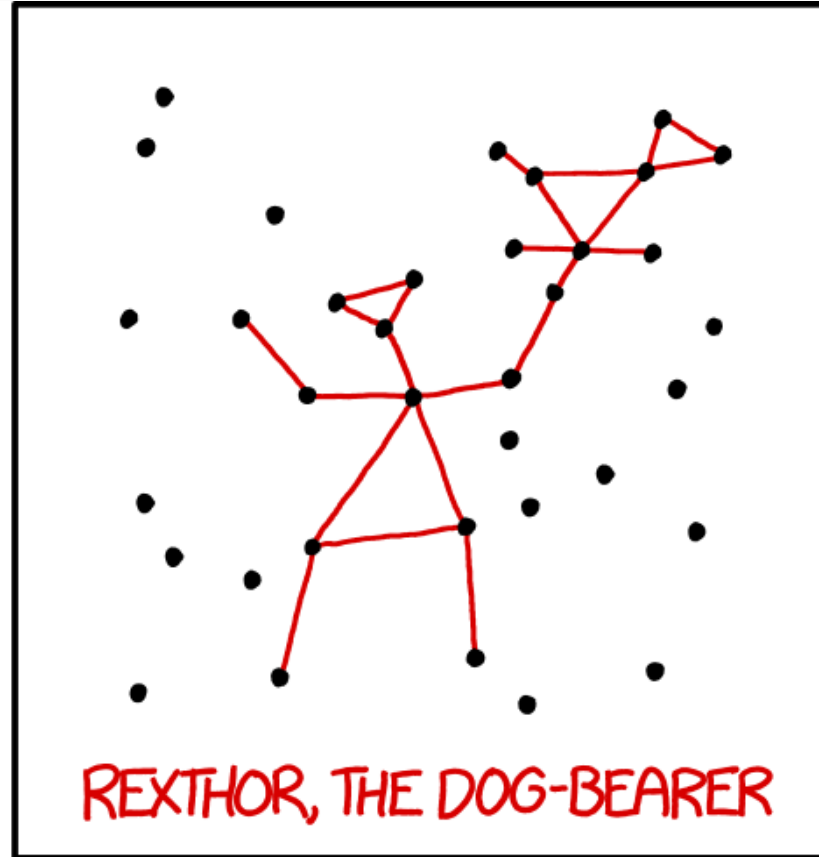
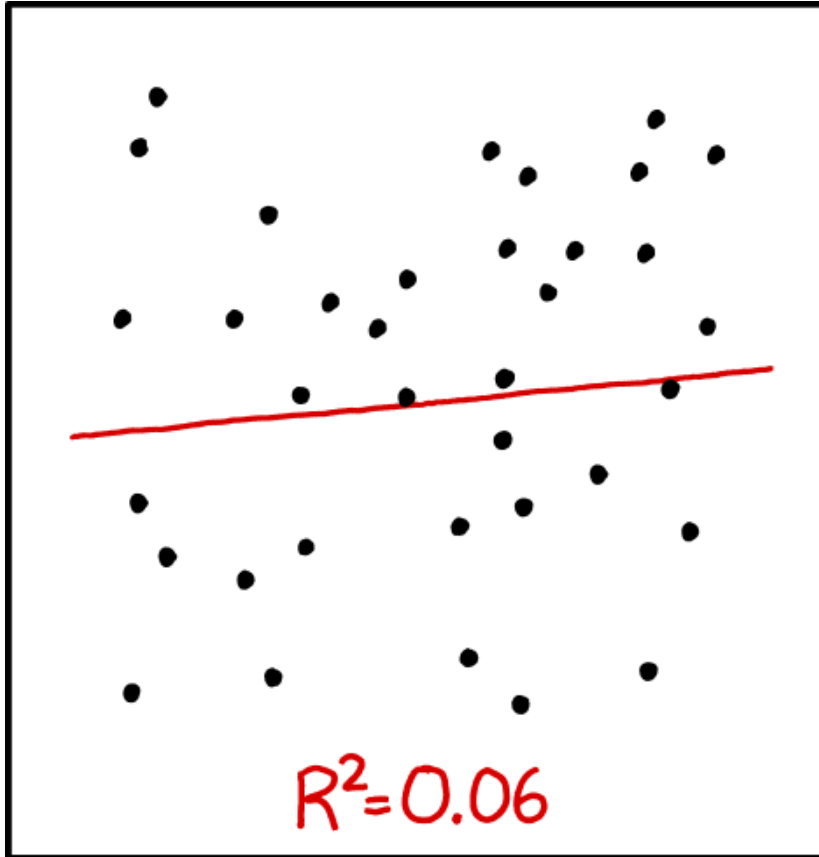
Model to predict body weight of a person

Model	<i>B</i>	β	<i>SE</i>	T	Sig.
(Constant)	-24.41		47.641	-0.512	0.627
height	47.379	0.520	27.628	1.715	0.137
age	0.297	0.607	0.114	2.602	0.041
male	8.992	0.434	5.603	1.592	0.162

- $R^2 = 0.754$; Standard error of the estimate = 6.587
- $\text{Weight} = 47.379 \times \text{Height} + 0.297 \times \text{Age} + 8.922 \times \text{is male} - 24.41$
- For the dichotomous variable gender, the coefficient is interpreted as the difference: a man weighs 8.922 kg more than a woman.

Standardized coefficients (β)

- The standardized coefficients beta (β) range between -1 and $+1$.
- The greater β is, the greater is the contribution of each independent variable to explain the dependent variable.
- In this regression analysis, the variable age has the greatest influence on the variable weight.



I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER
TO GUESS THE DIRECTION OF THE CORRELATION FROM THE
SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.