Linear Regression

upGrad
knowledgehut

Learning Objectives

- Explain covariance and correlation
- Explore simple linear regression
- Discuss slope, intercept and interpretation
- Describe OLS method and assumptions
- Explain multiple linear regression
- Explore beta coefficient
- Explore beta coefficient
- Describe r2 and adjusted r2 and demonstrate it
- Explore linear regression diagnostic plots

Covariance and Correlation

Covariance

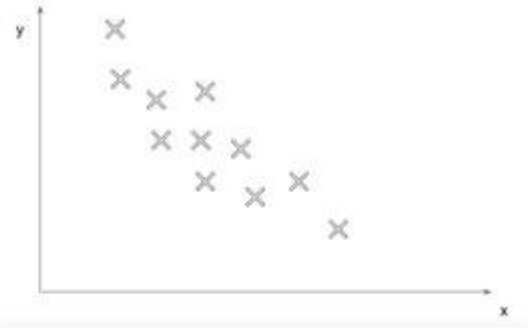
- Covariance is a statistical tool used to determine the relationship between the movements of two random variables.

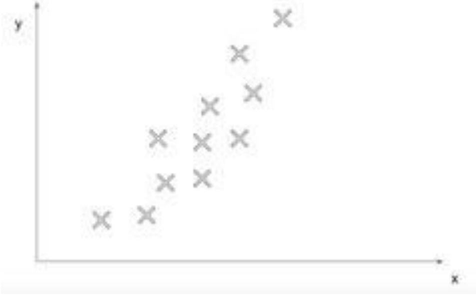
- When two stocks tend to move together, they are seen as having a positive covariance; when they move inversely, the covariance is negative.

Covariance

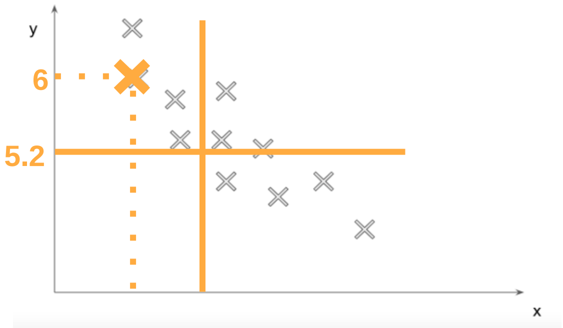
The expected value (or mean) of the product of the variances of two random variables, X and Y, is known as covariance.

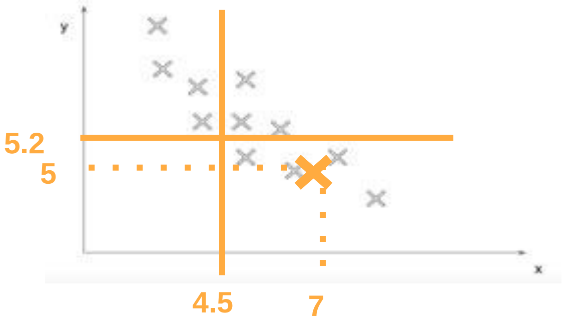
$$\mathrm{cov}(X,Y) = \mathrm{E}\left[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\right]$$

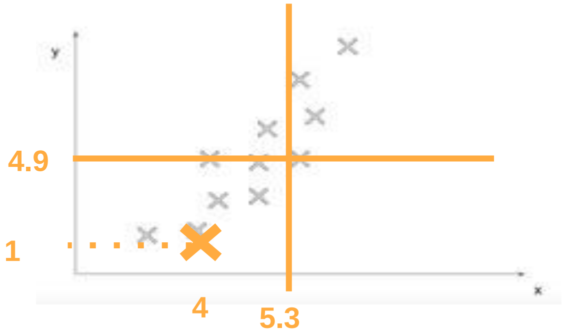


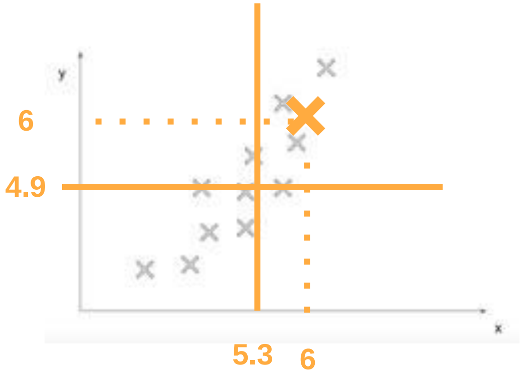


Covariance Examples









Correlation

The variance is divided by the standard deviation of x multiplied by the standard deviation of y.

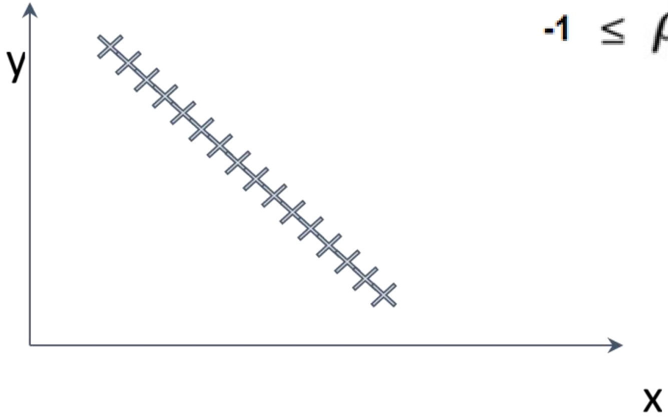
$$\operatorname{cov}(X,Y) = \mathrm{E}\left[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\right]$$

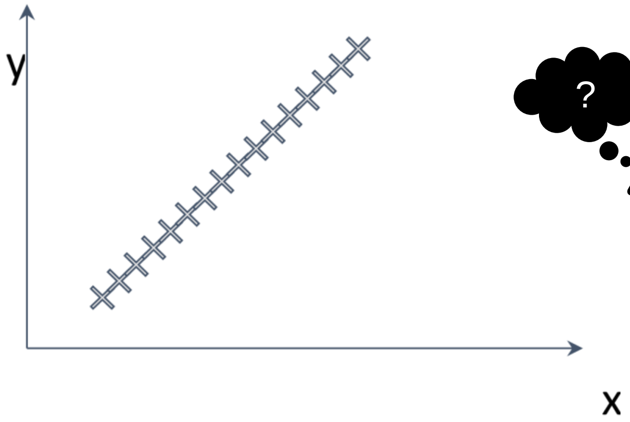
$$\operatorname{corr}(X,Y) = \frac{\operatorname{cov}(\underline{X},\underline{Y})}{\underline{\sigma_X}\,\underline{\sigma_Y}}$$

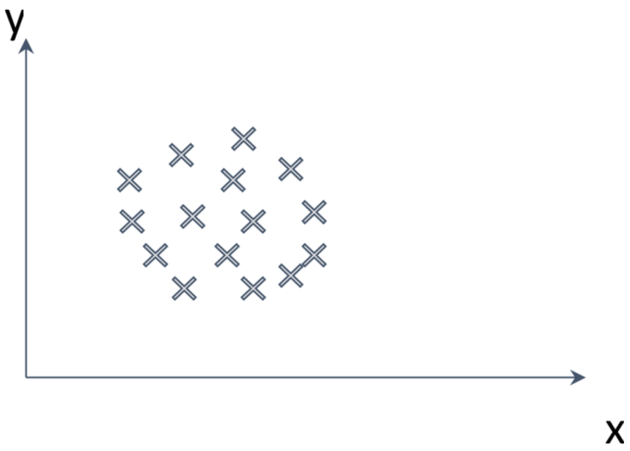
$$\rho_{X,Y} = \frac{\operatorname{cov}(\underline{X},\underline{Y})}{\underline{\sigma_X}\,\underline{\sigma_Y}}$$

Correlation Coefficient

$-1 \leq \rho_{X,Y} \leq 1$



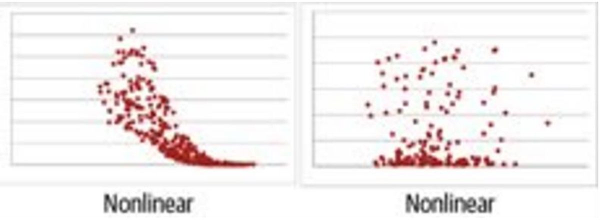




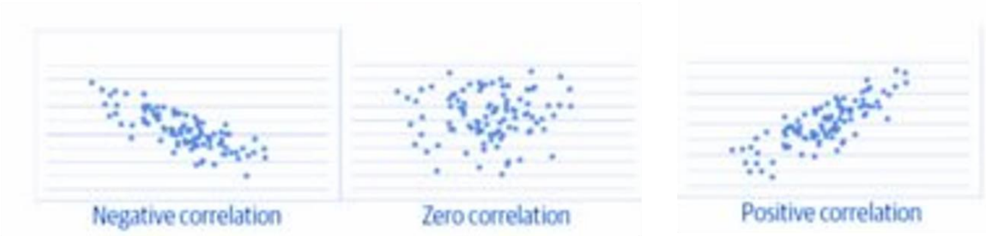
Simple Linear Regression

Pearson Correlation

Measures the strength of the linear relationship between two variables.



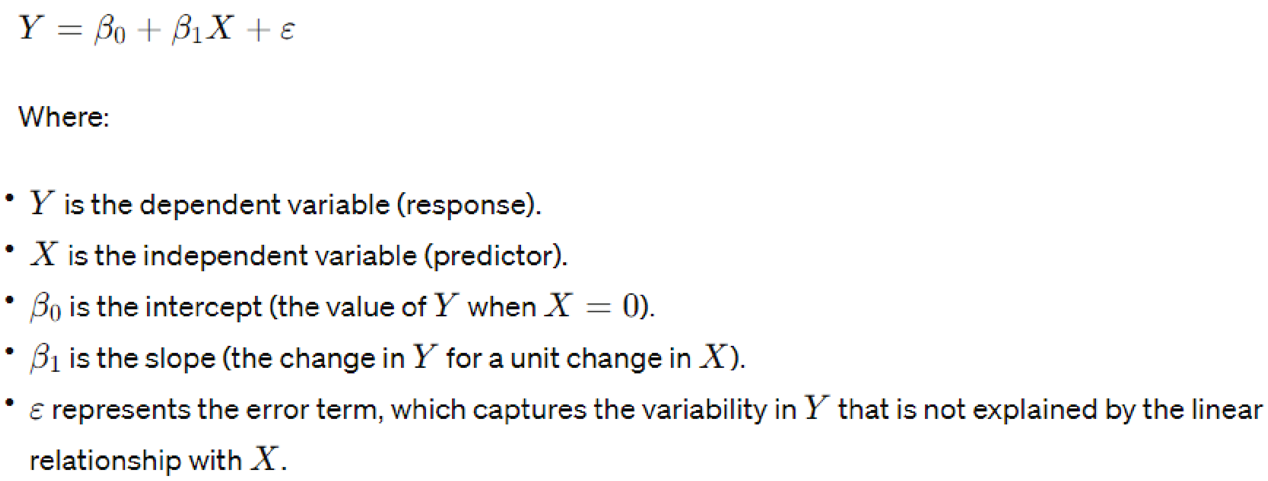
✗



✓

Simple Linear Regression

- Simple linear regression is a statistical method used to model the relationship between two continuous variables.
- It assumes that there is a linear relationship between the independent variable (predictor) and the dependent variable (response).
- In simple linear regression, we aim to fit a straight line to the data that best represents this relationship.



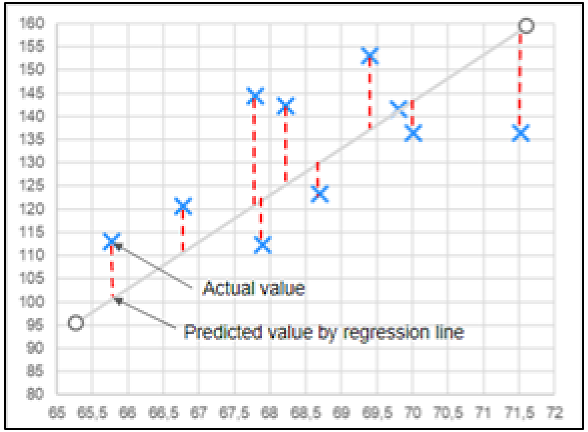
$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Where:

- Y is the dependent variable (response).
- X is the independent variable (predictor).
- β_0 is the intercept (the value of Y when $X = 0$).
- β_1 is the slope (the change in Y for a unit change in X).
- ε represents the error term, which captures the variability in Y that is not explained by the linear relationship with X.

Best Regression Line

- Problem with this measure: Errors will equal out (some residuals are positive, some are negative)
- Solution: Square the errors
- **Calculate the sum of squares error (SSE)**



Assumptions

Linearity

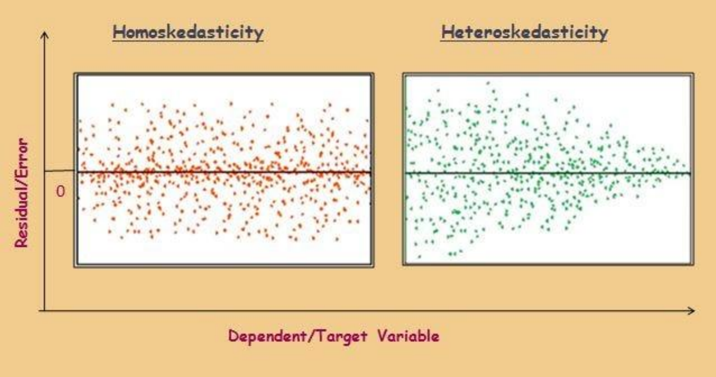
The relationship between X and the mean of Y is linear.

Homogeneity (scaling): $f(ax) = af(x)$

Additivity: $f(x_1 + x_2) = f(x_1) + f(x_2)$

Homoscedasticity

The variance of residual is the same for any value of X

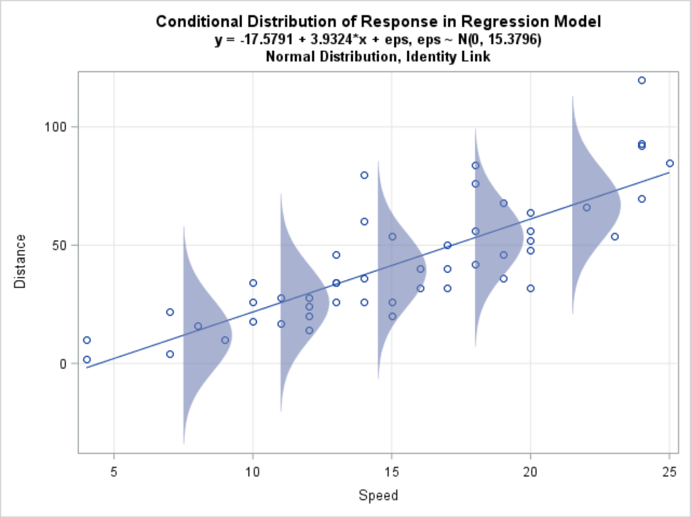


Independence

Observations are independent of each other.

Normality

For any fixed value of X, Y is normally distributed.



Slope and Intercept, Interpretation

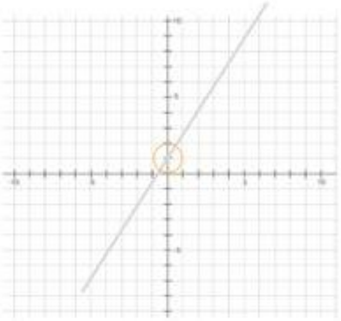
Simple Regression

$$\boxed{Y} = \boxed{\beta_0} + \beta_1 \times \boxed{X} + \epsilon$$

Dependent variable / Response / Output

Intercept / Constant

Independent variable / Regressor / Predictor / Input



$$\boxed{Y} = \boxed{\beta_0} + \boxed{\beta_1} \times \boxed{X} + \epsilon$$

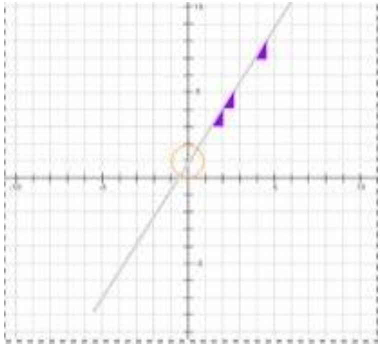
Dependent variable / Response / Output

Intercept / Constant

Coefficient (Slope)

Independent variable / Regressor / Predictor / Input

Y = 1 + 1.5 * X



$$\boxed{Y} = \boxed{\beta_0} + \boxed{\beta_1} \times \boxed{X} + \boxed{\epsilon}$$

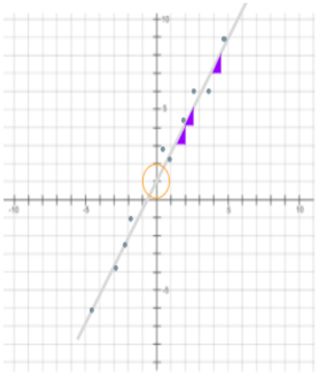
Dependent variable / Response / Output

Intercept / Constant

Coefficient (Slope)

Independent variable / Regressor / Predictor / Input

Error term



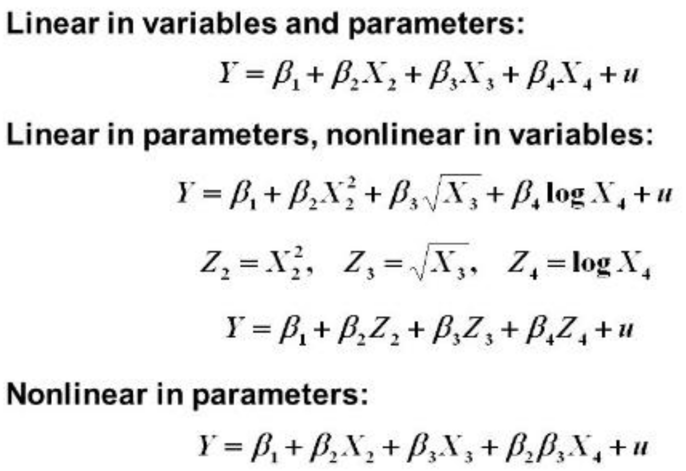
OLS Method and Assumptions

Ordinary Least Square Method

OLS	Used by most regression implementations to minimize the residuals.
	The lower the sum of squared residuals, the less difference between the actual and predicted values, = the better the regression equation is at making estimates.
	Generalized least squares, Maximum likelihood estimation, Bayesian regression.

Assumptions

The linear regression model is “linear in parameters”



Linear in variables and parameters:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + u$$

Linear in parameters, nonlinear in variables:

$$Y = \beta_1 + \beta_2 X_2^2 + \beta_3 \sqrt{X_3} + \beta_4 \log X_4 + u$$

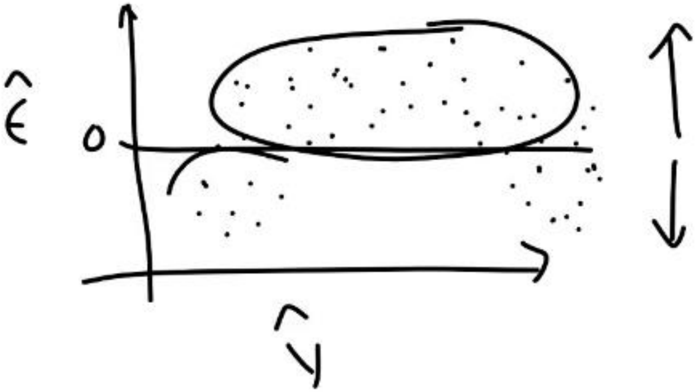
$$Z_2 = X_2^2, \quad Z_3 = \sqrt{X_3}, \quad Z_4 = \log X_4$$

$$Y = \beta_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 Z_4 + u$$

Nonlinear in parameters:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_2 \beta_3 X_4 + u$$

Zero mean of the error term

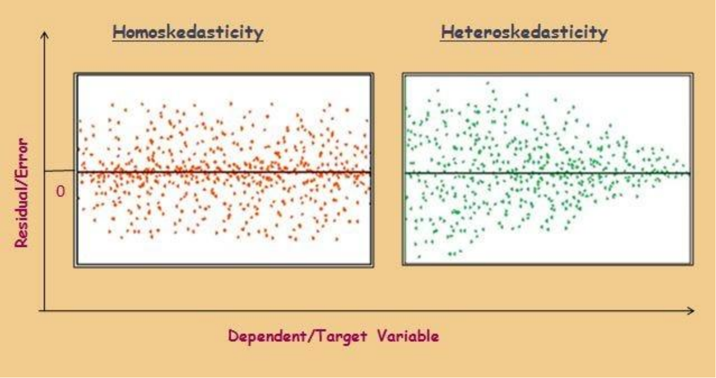


Assumptions

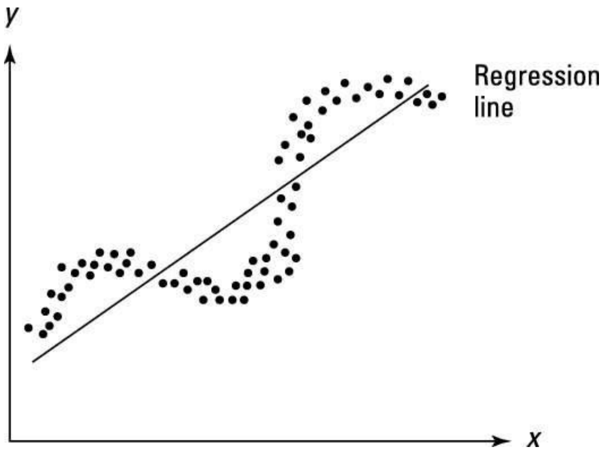
Exogeneity of the input variables

It means that the predictor variables are not affected by the random errors in the model.

Homoscedasticity

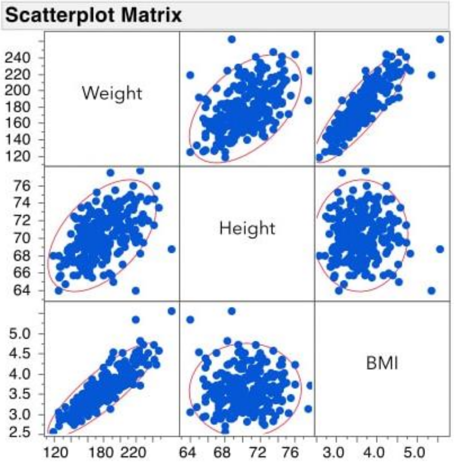


No autocorrelation

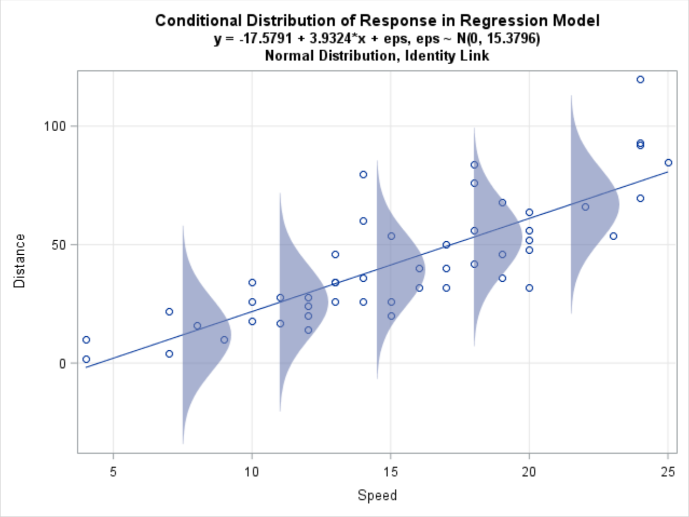


Assumptions

No Multicollinearity



The error term is normally distributed



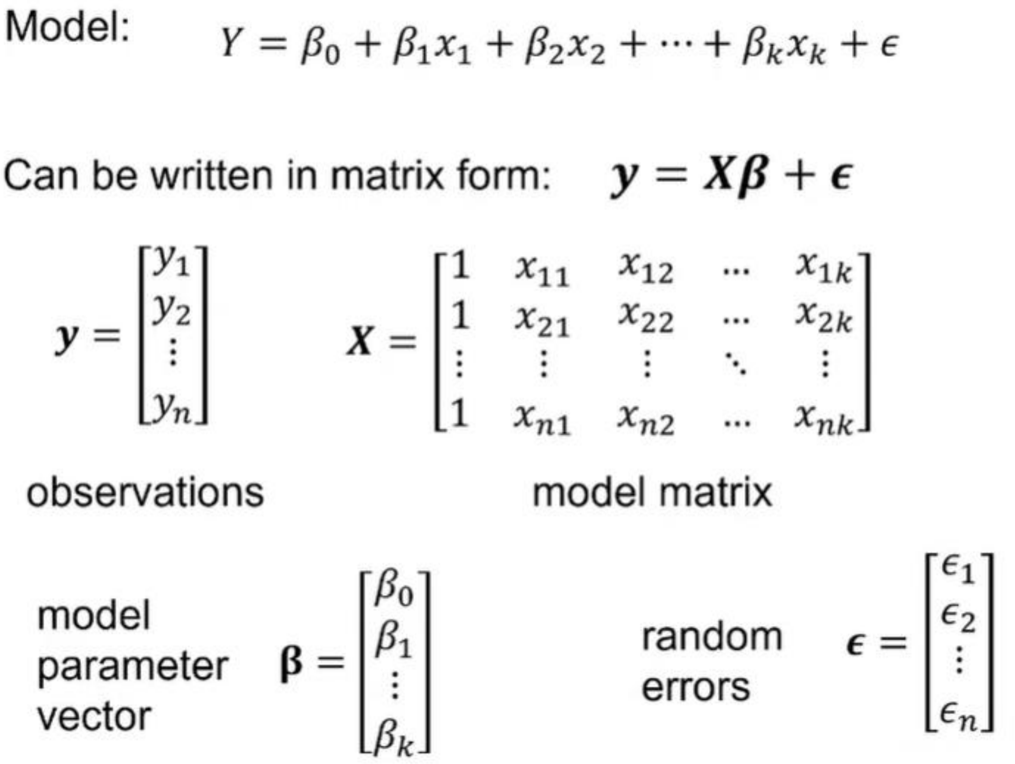
Statsmodel

The models module of scipy.stats was originally written by Jonathan Taylor. For some time, it was part of scipy but was later removed.

- Statsmodels is a Python library that provides classes and functions for the estimation of many different statistical models, tests, and data exploration.

- **Regression Analysis**: Statsmodels allows you to perform various types of regression analysis, including ordinary least squares (OLS).

Linear Regression in Matrix Form



Model: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$

Can be written in matrix form: $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}$$

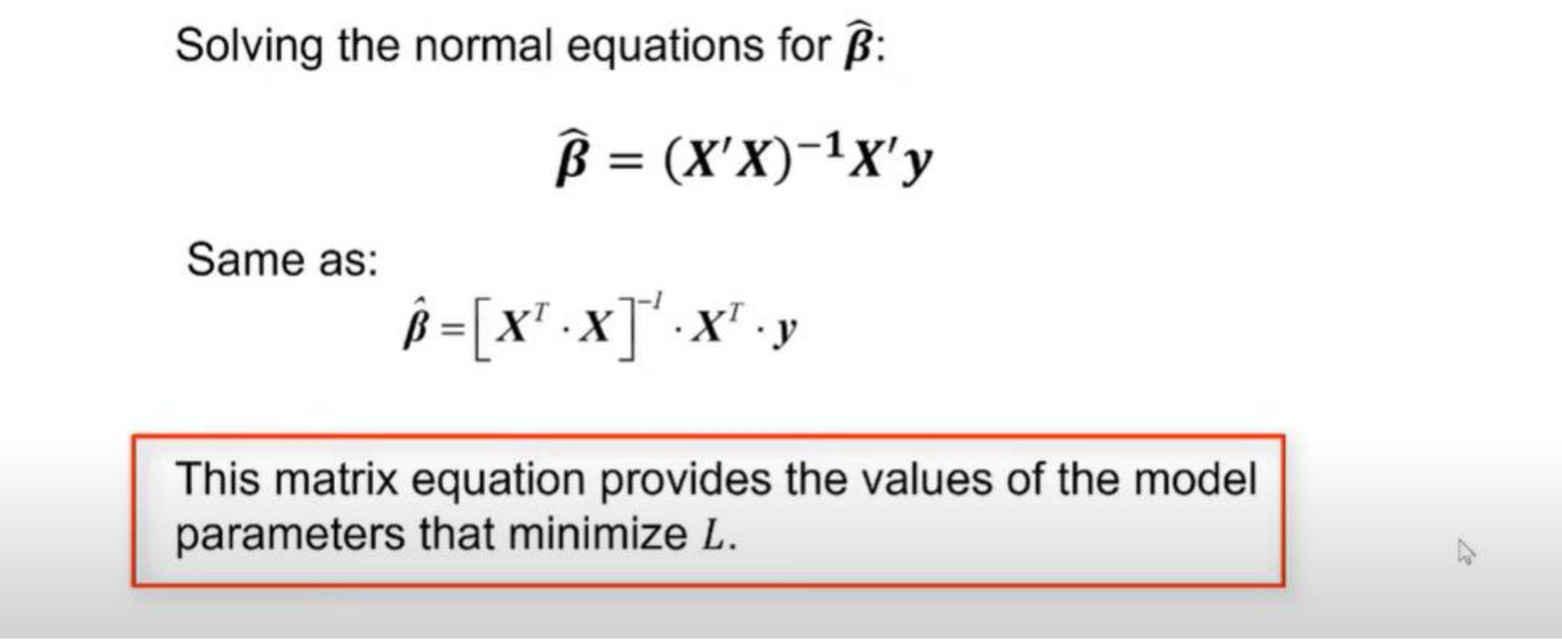
observations

model matrix

model parameter vector $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$

random errors $\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$

Linear Regression in Matrix Form



Solving the normal equations for $\widehat{\boldsymbol{\beta}}$:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

Same as:

$$\hat{\beta} = \left[X^T \cdot X\right]^{-1} \cdot X^T \cdot y$$

This matrix equation provides the values of the model parameters that minimize L.

Multiple Linear Regression

Multiple Linear Regression and Fitted Values

Multiple linear regression is a regression model that estimates the relationship between **two or more** input variables to an output variable.

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p + e$$

Fitted (predicted) values are often called *Y_hat* or *beta_hat*.

$$\boxed{\widehat{Y}_i} = \boxed{\hat{b}_0} + \boxed{\hat{b}_1} X_{1,i} + \boxed{\hat{b}_2} X_{2,i} + \ldots + \boxed{\hat{b}_p} X_{p,i}$$

Polynomial Linear Regression

1. $\boldsymbol{y} = \beta_o + \beta_1 \boldsymbol{x_1} + \beta_1 \sqrt{\boldsymbol{x_1}}$

2. $\boldsymbol{y} = \beta_o + \beta_1 \boldsymbol{x_1} + \beta_2 \boldsymbol{x_1}^2 \mathbf{+}\ \beta_3 \boldsymbol{x_1}^3$

Beta Coefficients

House Price Example

Small coefficient?

Large coefficient?

b_1 Living Area (sqm)	b_2 Land area (sqm)	$\cdots$ No. of Bedrooms	b_p Rating	Y House Price
222	870	6	7	305,195
349	1872	4	10	1,091,868
191	2418	4	8	773,273
…	…	…	…	…

R2 and Adjusted R2

R2

R-Squared a.k.a:

- “Coefficient of determination”
- R^2 or r^2
- “R Square"

$$R^2 = 1 - \frac{RSS}{TSS}$$

Measures “the goodness of fit”
Between 0 and 1.
1: Perfect fit (all points are accounted for by the model)
0: No fit at all (no predictive value, the model can’t explain any of the variability)

RSS… sum of squares of residuals
TSS… total sum of square

Adjusted R2

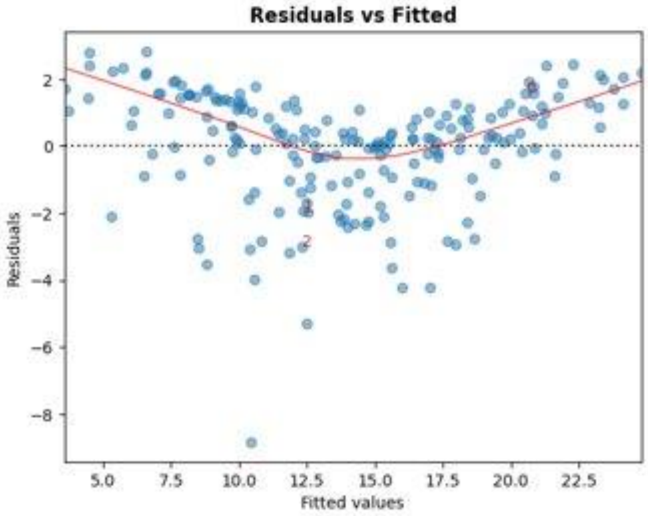
- Problem with R-Squared: The more variables you add, the better the R-Squared will probably be.
- Models with different number of parameters are not really comparable wrt. R-squared.
- Adjusted R-Squared tries to consider this.
- Adjusted R-Squared is always lower or al than R-Squared .

$$\text{Adjusted R}^2 = 1 - \frac{SS_{residuals} / (n-K)}{SS_{total} / (n-1)}$$

RSS… sum of squares of residuals
TSS… total sum of squares
n… number of observations
K… number of parameters

Linear Regression Diagnostics Plots

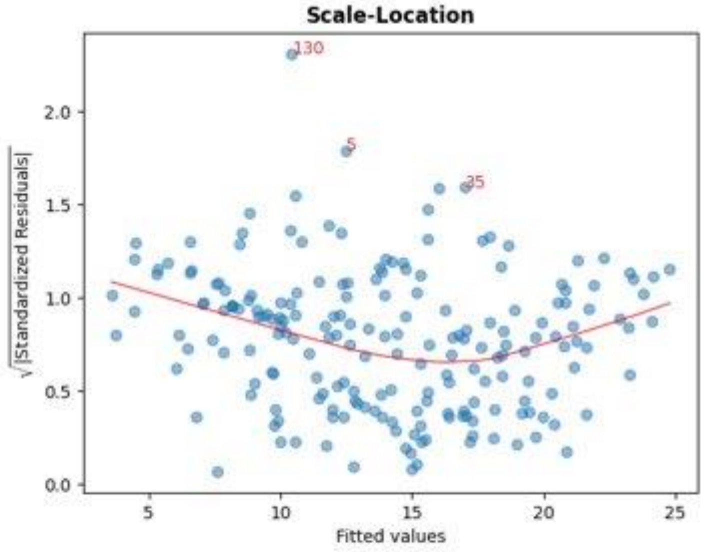
Residual vs. Fitted



Check for non-linearity: Are errors normally distributed, and the plot should not show any notable pattern.

The red line in the graph should (roughly) be horizontal around 0.

Scale-location

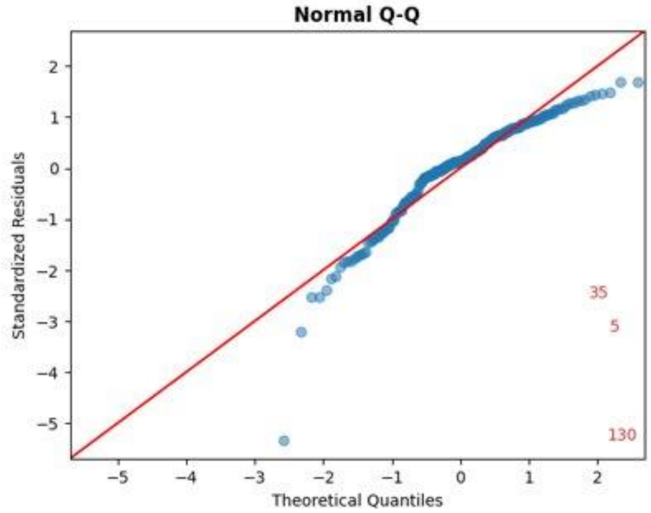


Scale-location (spread-location) plot shows if residuals are spread equally along fitted values.

Ideally: Horizontal line with equally spread points.

Here: Violation - variability of residuals increases with fitted values for Y (**heteroscedasticity**).

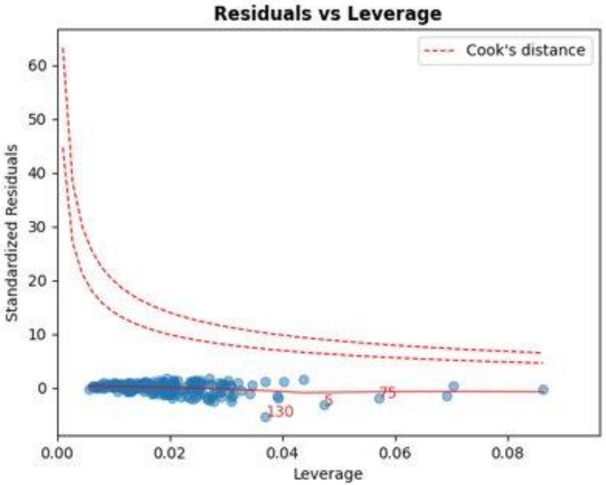
Normality of Residuals



Plots **quantiles** of observed data versus **quantiles** of an ideal distribution to check if residuals are normally distributed.

Points spread along the diagonal line suggest normality.

Normality of Residuals



Individual points can heavily influence the regression model - outliers have high leverage and can quickly become problematic.

Cook's distance shows the influence of each observation on the fitted response values.

Points falling outside the Cook’s distance curves are considered observations that can influential.

Try to keep points outside these curves as shown.

Thank you