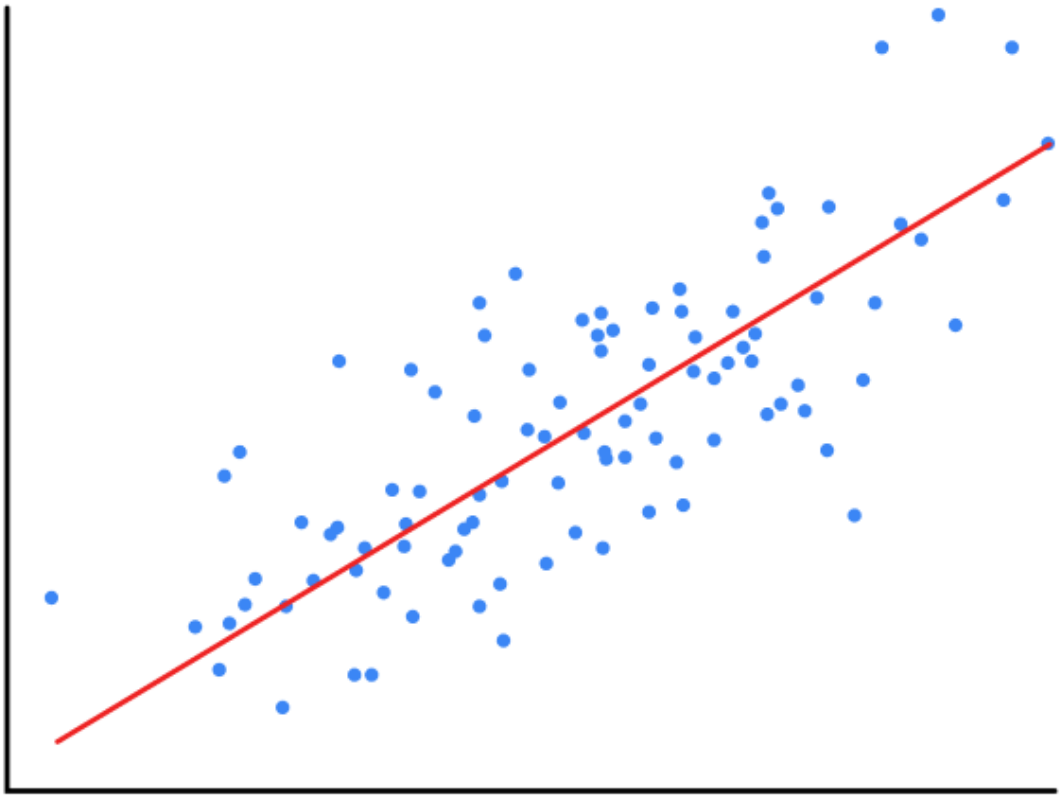
Linear Regression Comprehensive Cheat Sheet (with Examples!)

The purpose of this cheat sheet is to provide an overview of Linear Regression. When I first learned about linear regression in my Data Science bootcamp at Flatiron School, I heavily struggled with the material. But after spending countless nights understanding the material and writing down my notes, I decided to share my notes with everyone for those who are struggling or would like a refresher on the material.

In this cheat sheet, I have provided simple textbook definitions and will provide examples of each important concept. As we dive into what Linear Regression is and understand the concepts, I hope this can help you on your journey to becoming a Data Scientist like how it has helped me.



Example of what Linear Regression is and Line of Best Fit (Red)

What is Linear Regression?

Statistical method that helps estimate the strength and direction of the relationship between two (or more) variables.

Simple Linear Regression uses a single feature (one independent variable) to model a linear relationship with a target (one dependent variable) by fitting the best straight line to describe the relationship.

Multiple Linear Regression uses more than one feature to predict a target variable by fitting the best linear relationship.

In linear regression, your primary objective is to optimize your predictor variables in hopes of predicting your target variable as accurately as possible.

What is Linear Regression used for?

Regression analysis is a strong tool and has many use cases such as the following:

1. Identify the strength of the effect that the independent variable(s) have a on a dependent variable

2. Forecast effects or impacts of changes

3. Predicts trends and future values

Linear Regression Key Components

Straight Line Equation: y = mx + b

Dependent Variable (y): variable that is being estimated and predicted, also known as target

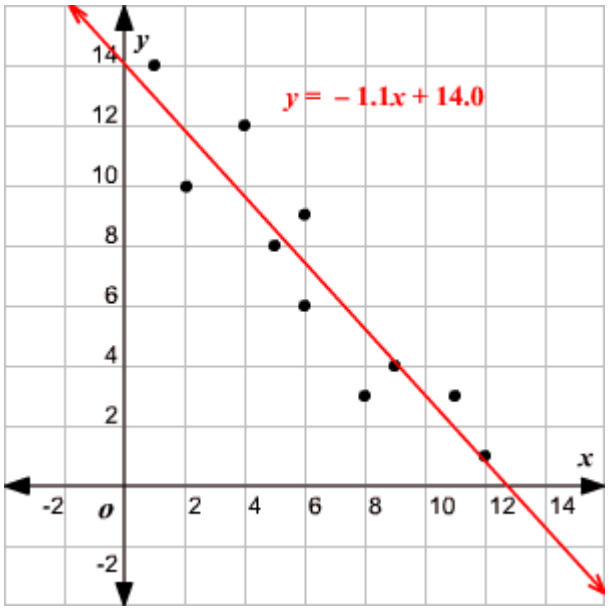
Independent Variable (x): input variable, also known as predictors or features

Coefficient: is a numerical constant, also known as parameter

Slope (m) : determines the angle of the line

Intercept (b): constant determining the value of y when x is 0

The challenge for regression analysis is to fit a line, out of an infinite number of lines that best describe the data.



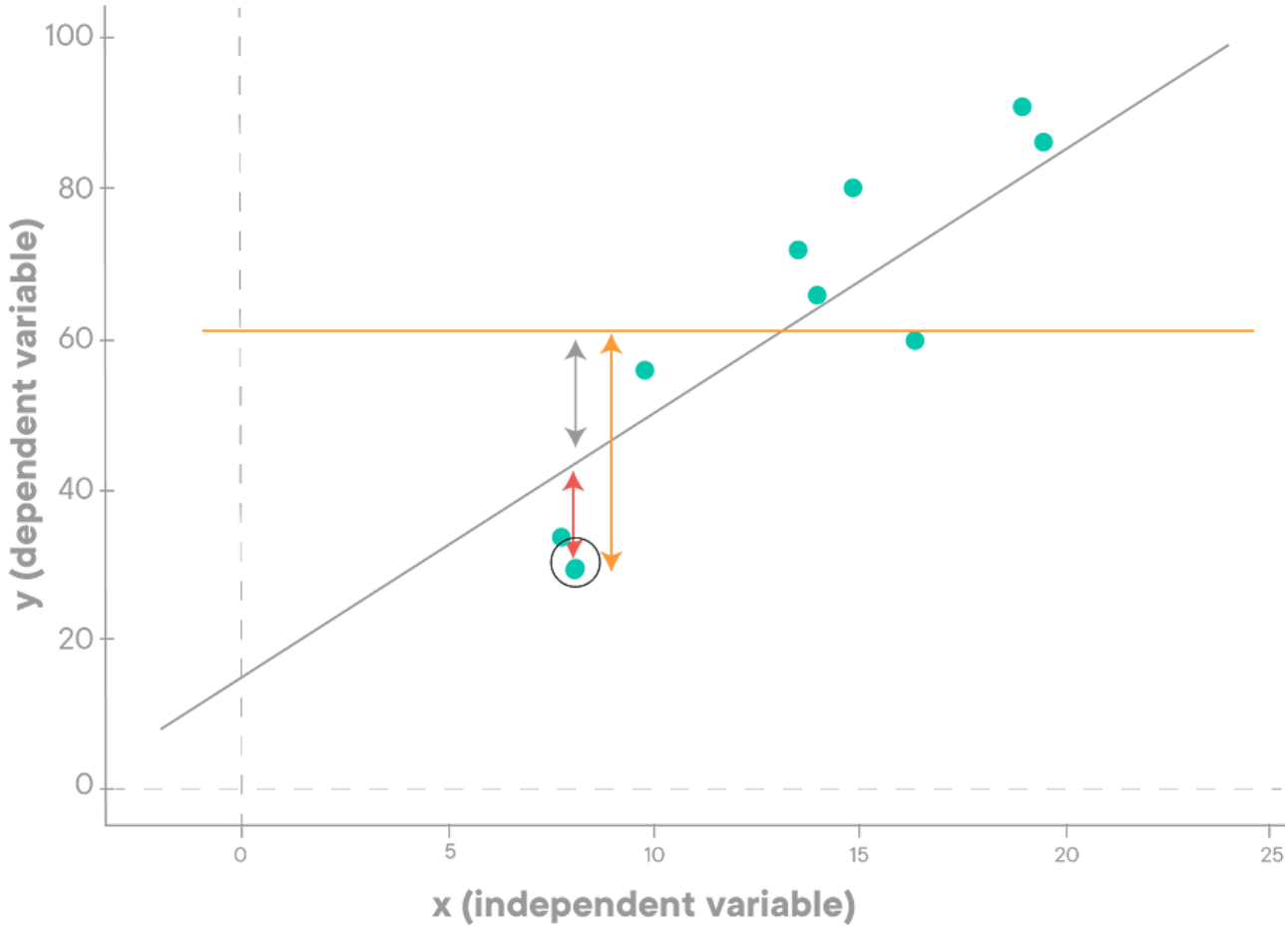
- For this example to the right, the intercept (b) starts at 14.0.
- The number -1.1 is the coefficient used to multiple the independent variable, x.
- For each unit change in x, we can calculate for y.

How to determine the line of best fit?

R-Squared (Coefficient of Determination): statistical measure that is used to assess the goodness of fit of a regression model

- It uses a baseline model that finds the mean of the dependent variable (y) and compares it with the regression line (yellow line below)

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

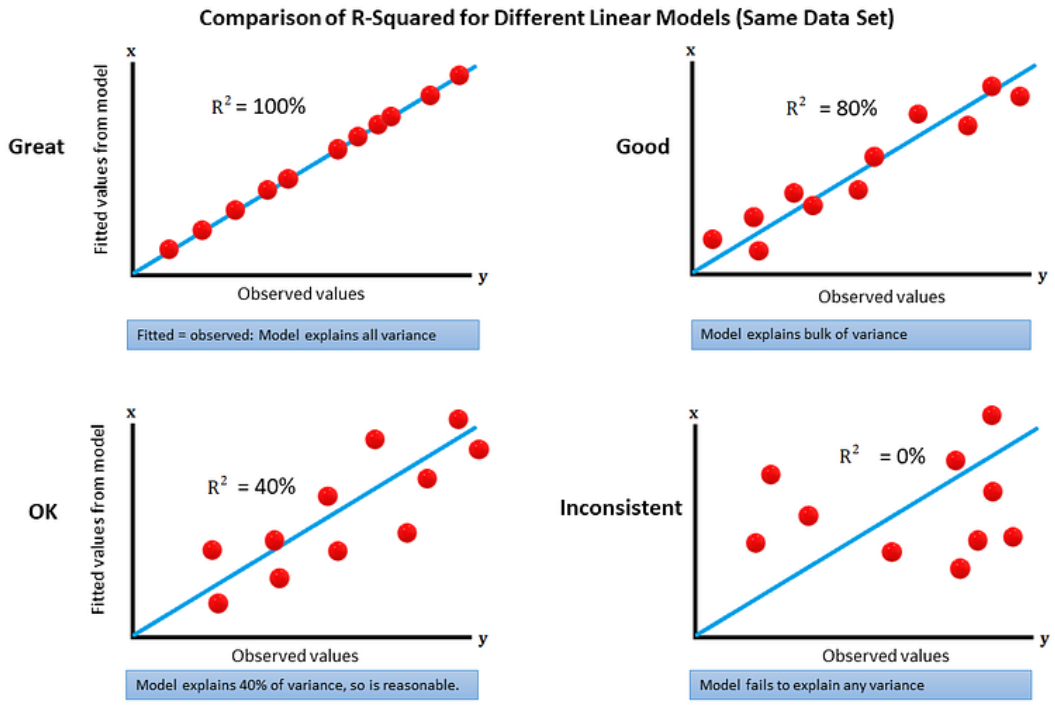


Red:

Residual Sum of Squared Errors (RES) : also known as SSE and RSS, is the sum of squared difference between y and predicted y (red arrow)

Total Sum of Squared Errors (TOT): also known as TSS, is the sum of squared difference between y and predicted y (orange arrow)

R-Squared can take a value between 0 and 1 where values closer to 0 represents a poor fit and values closer to 1 represent an (almost) perfect fit



85% of the variations in dependent variable y are explained by the independent variable in our model.

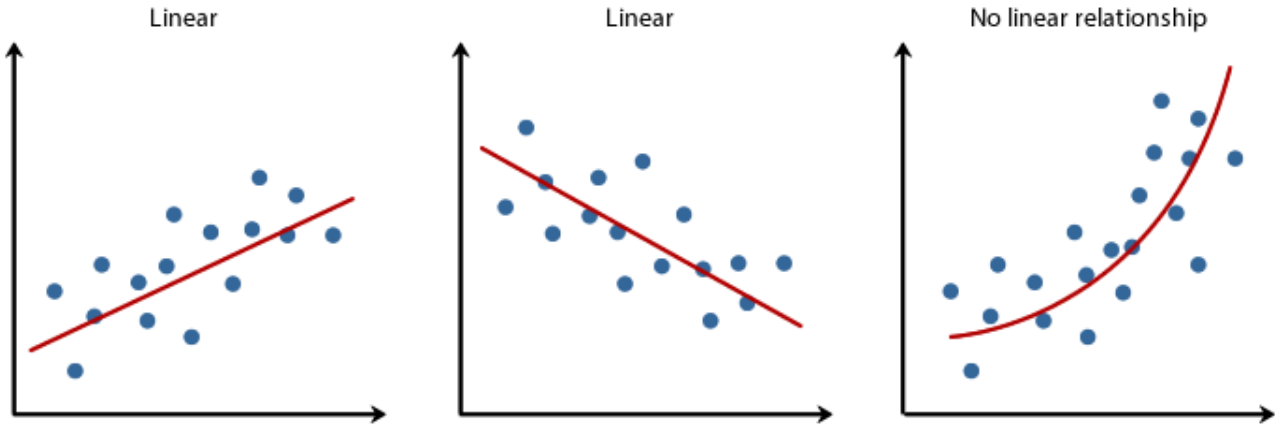
Assumptions of Linear Regression

There are four assumptions associated with a linear regression model. If these assumptions are violated, it may lead to biased or misleading results.

Linearity: relationship between independent variable(s) and dependent variable is linear

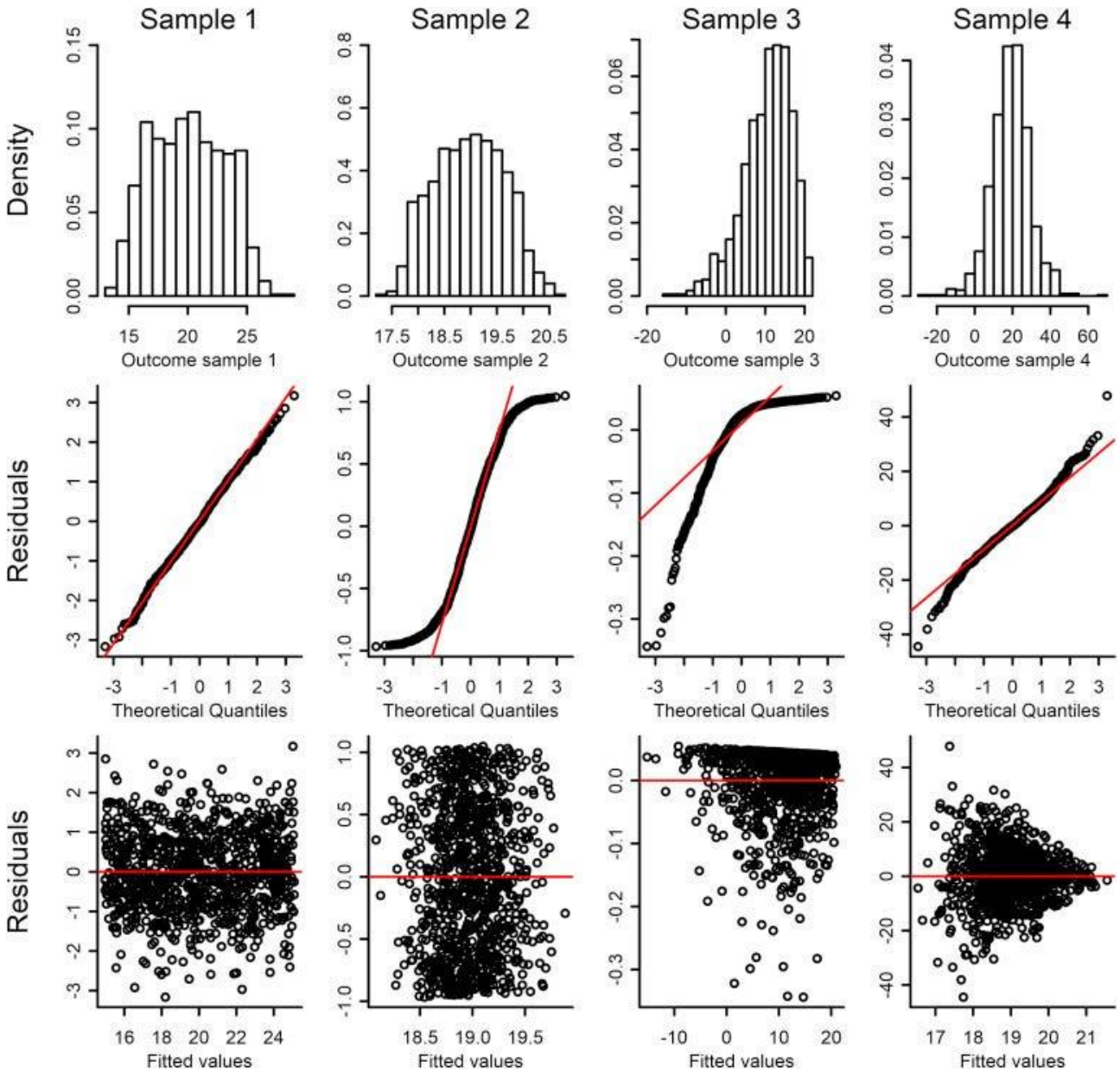
- if not respected, regression will underfit and will not accurately model the relationship between independent and dependent variables
- if there is no linear relationship, various methods can be used to make the relationship linear such as polynomial and

exponential transformations for both independent and dependent variables



Normality: model residuals should follow a normal distribution

- if distribution is not normal, regression results will be biased and it may highlight that there are outliers or other assumptions being violated
- correct the large outliers in the data and verify if the other assumptions are not being violated

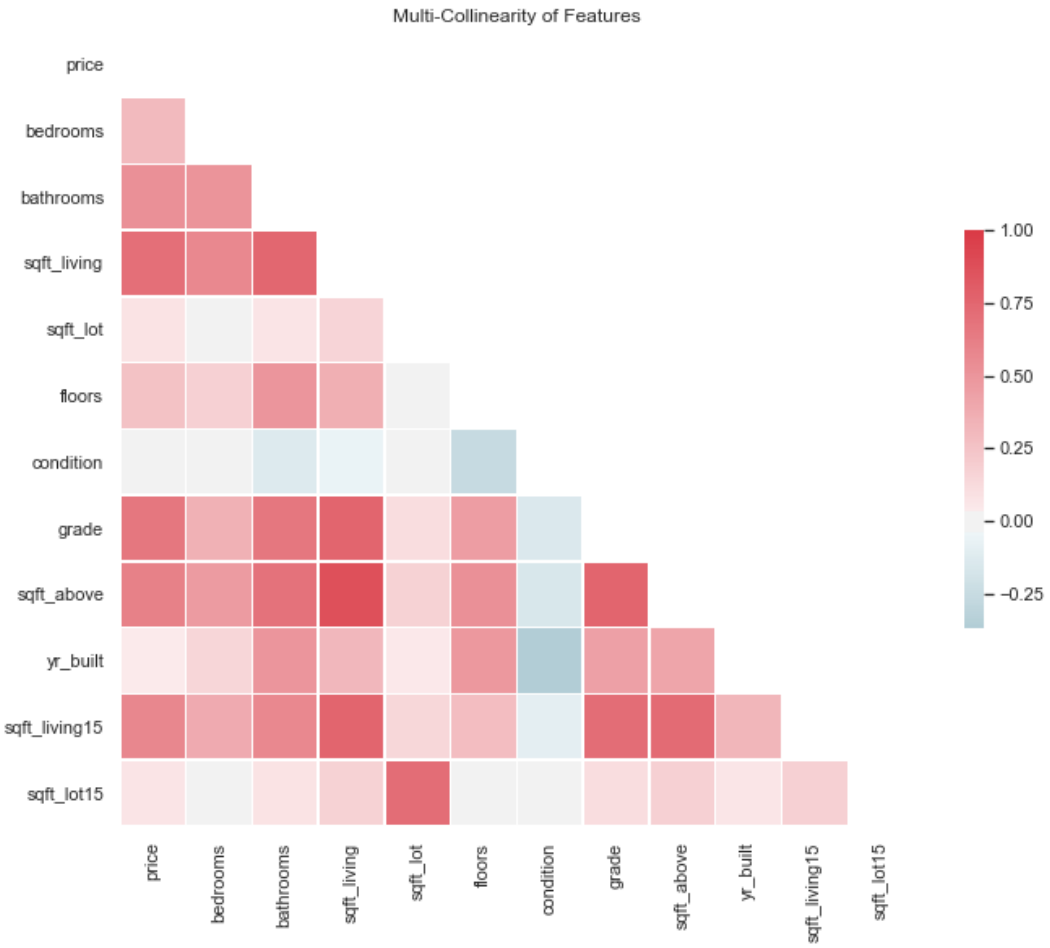


Sample 1 follows a normal distribution

Independence: each independent variable should be independent from other independent variables

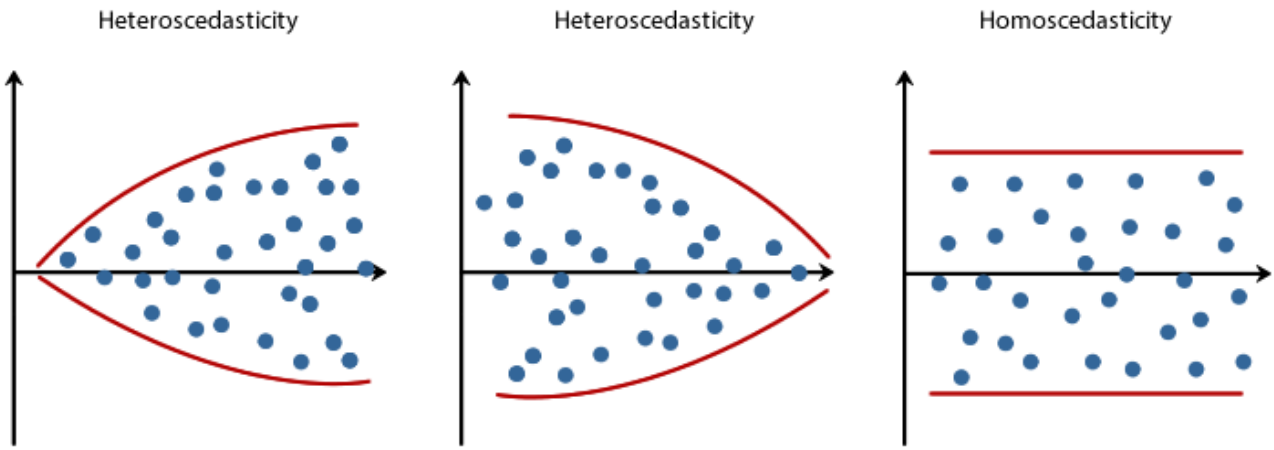
- multicollinearity is when independent variables are not independent from each other
- it indicates that changes in one predictor are associated with changes in another predictor

- we use heatmaps and calculate VIF (Variance Inflation Factors) scores which compares each independent variable's collinearity with other independent variables



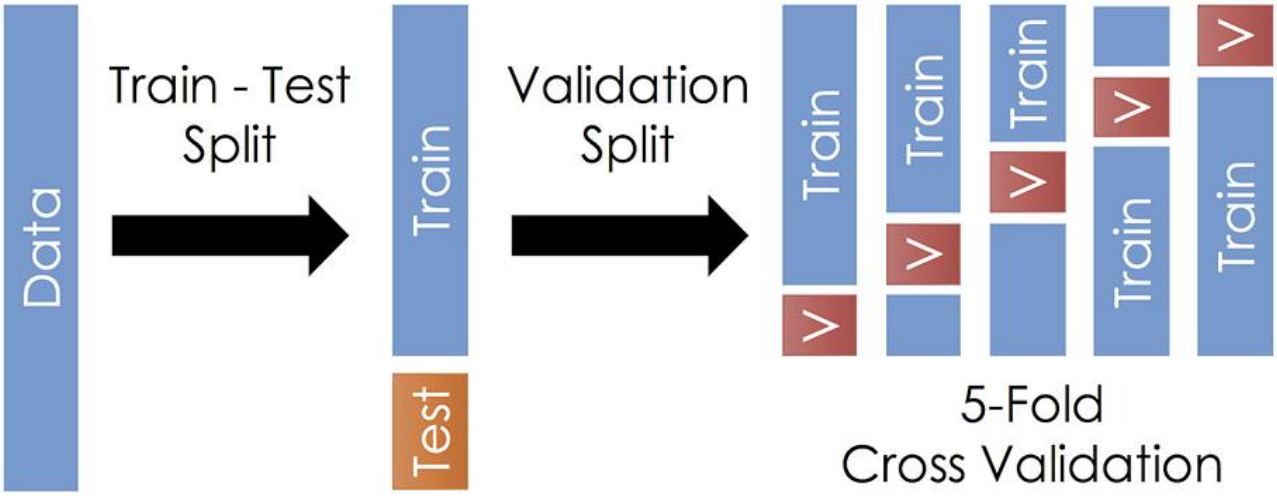
Homoscedasticity: the variance of residual is the same for any value of x, fancy word for "equal variances"

- the model does not fit all parts of the model equally which lead to biased predictions
- it can be tackled by reviewing the predictors and providing additional independent variables (and maybe even check that the linearity assumption is respected as well)



How can we validate our regression model results?

We have finished building the model, but how can we validate if the model can predict the correct outcome. We use **Train-Test Split** to *randomly* split the data into two separate samples, sample for "training" the data and the other sample for "testing" the data. The general rule is to split the data into 70% training data and 30% testing data, but similar percentage splits can work as well.



To calculate the results for both train and test data, a popular metric is the Root Mean Squared Error (RMSE).

Mean Squared Error (MSE): average squared difference between the estimated values and the actual value

- the smaller the MSE, the closer the fit is to the data

Root Mean Squared Error (RMSE): square root of MSE

- easier to interpret since it is the same units as the quantity plotted on the x axis
- the RMSE is the distance on average of a data point from the fitted line, measured along a vertical line

5-Fold Cross Validation

Running a model with different Train-Test Split will lead to different results. This is where **5-Fold Cross Validation** comes in where we split the data into "*k*" equal sections of data, with each linear model using a different section of data as the test data and all other sections combined as the training set.

Sources

- https://flatironschool.com/career-courses/data-science-bootcamp/online
- https://www.statisticssolutions.com/what-is-linear-regression/

- https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R5_Correlation-Regression/R5_Correlation-Regression4.html
- https://towardsdatascience.com/verifying-and-tackling-the-assumptions-of-linear-regression
- https://www.vernier.com/til/1014