Data science club meeting note
Nov. 28th

Linear regression – by Ascari

- Beta value - model coefficient

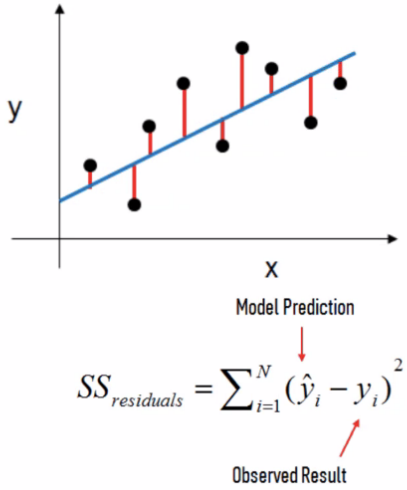
Linear Regression Introduction: Part 2

The β values in the previous equation are called the **model coefficients**:

- These values are estimated (or "learned") during the model fitting process using the **least squares criterion**.
- Specifically, we are trying to find the line (mathematically) that minimizes the **sum of squared residuals** (or "sum of squared errors").
- Once we've learned these coefficients, we can use the model to predict the response.

In the diagram to the right:

- The black dots are the **observed values** of x and y.
- The blue line is our **least squares line**.
- The red lines are the **residuals**, which are the vertical distances between the observed values and the least squares line.



-
- Goal: find the best fitting line that minimize the residual error

Linear Regression Assumptions

Let's start with some basics. Linear Regression assumes:

- Data is **normally distributed** (but doesn't have to be - good topic to research)
  - residuals should be normally distributed, however
  - test with **histogram/ Q-Q plot/** or **Kolmogorov-Smirnov Test**
- X's significantly explain y (**low p-values**)
- X's are independent of each other (**little to no multicollinearity**)
  - test with tried and true **correlation matrix** or a **variance inflation factor (VIF)** from statsmodel
- Resulting values pass a **linear assumption**
  - use **scatter plots/pairplot** to check for **linear relationships** between **target/features**
- There should be **little to no autocorrelation** in the data (i.e. **residuals** should be independent from each other)
  - Use **Durbin-Watson test** or **scatter plots** to check
- residuals must be **equal** across the regression line (i.e. **homoscedasticity** assumption)
  - check with **lmplot/Levene's test/NCV test**, etc.

-
- By Visual:
  - Divide the data into two dataset: one predict dataset, one test dataset

    - ??? need someone to fill out the method Vishual was talking about

Linear regression on Jupyter Notebook

- Simple linear regression example:
    - Link:
-