MODULE 2

Simple linear regression is a technique that estimates the linear relationship between 1 independent variable x, an 1 continuous variable y.

Best fit line

- The line that fits the data best by minimizing a loss function or error.
- We need to measure error to find the best fit line

Residual

- The difference between observed value and predicted value (estimated y value)
- $Residual = observed\ value - predicted\ value$

  $\varepsilon_i = y_i - \widehat{y_i}$
- The sum of the residuals is always equal to 0 for OLS estimators

Sum of Squared Residuals (SSR)

- The sum of squared differences between each observed value and its predicted value
- $\sum_{i=1}^{n}(Observed - Predicted)^2$

Ordinary Least Squared (OLS)

A method that minimizes the SSR to estimate parameters in a linear regression model

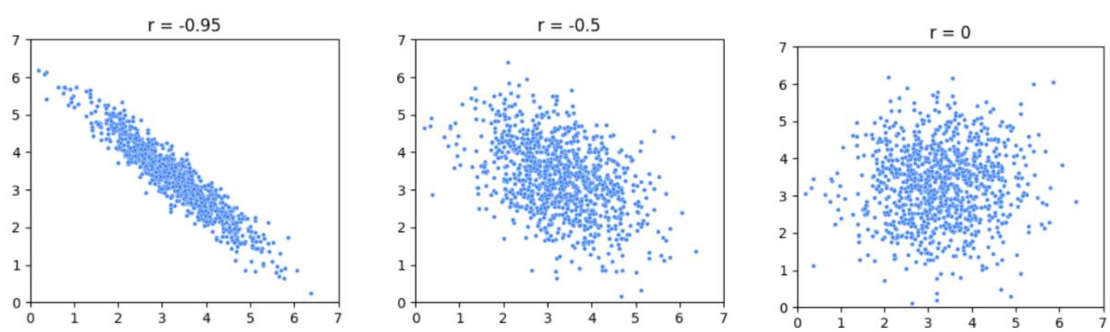
Estimating beta coefficients

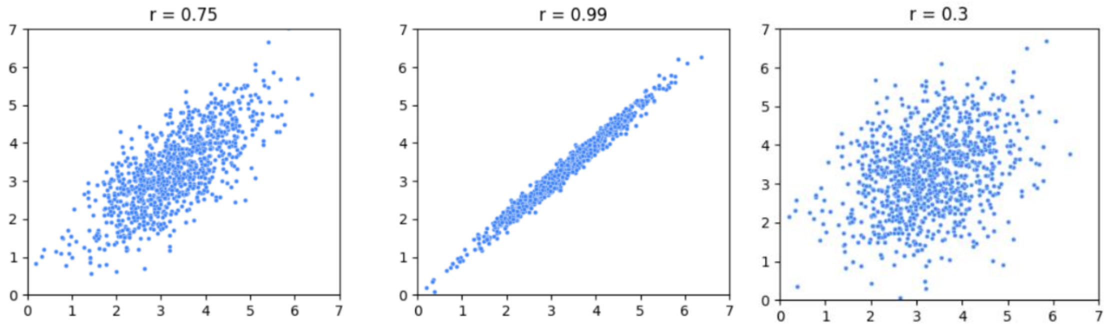
- $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$
- $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X}$

r (Pearson's correlation coefficient), quantifies the strength of the linear relationship between two variables using value of -1 to 1.

- r is negative = negative correlation
- r is 0 = no correlation
- r is positive = positive correlation

r only tells about the strength of the correlation, it doesn't include other information such as the gradient of the slope.



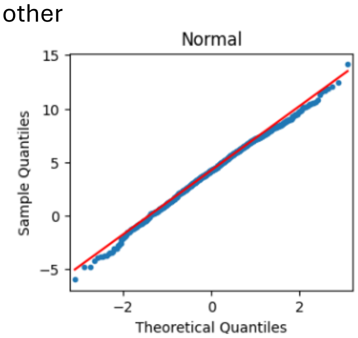


$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Slope for regression line is $m = \frac{r(SD\ y)}{SD\ x}$, the calculate the y-intercept.

Liner regression assumptions:

1. **Linearity:** Each predictor variable x, is linearly related to the outcome variable y
   Make sure that the points on the plot appear to fall along a straight line
2. **Normality:** Residual values are normally distributed
   Use quantile-quantile plot (QQ plot) to check the assumptions
   QQ plot is used to compare two probability distributions by plotting their quantiles against each other



3. **Independent observations:** Each observation in the dataset is independent
4. **Homoscedasticity (having the same scatter):** The variance of the errors is constant or similar across the model
   Plot a scatter graph of fitted value vs residuals, the assumption is true if the shape is random cloud

What to do if the assumptions is violated

1. **Linearity**
   Transform one or both variables, for example taking the logarithm
2. **Normality**
   Transform one or both variables, for example taking the logarithm
3. **Independent observations**
   Take a subset of the available data

4. **Homoscedasticity**
   Define different outcome variable or transform the y variable

Measures of uncertainty:

- **Confidence** intervals around beta coefficients
- **p-values** for the beta coefficients
- **confidence band** around the regression line

Hypothesis test on regression results (to know if x is correlated with y or not) :

- **H_0** (null hypothesis - Difference in x is not correlated with difference in y): $\beta_1 = 0$
- **H_1** (alternative hypothesis): $\beta_1 \neq 0$

Common evaluation metrics:

- **R^2 (coefficient of determination)**
  Measures the proportion of variation in the dependent variable y, explained by the independent variable(s) x
  The value are in range of 0 - 1
- **Mean Squared Error (MSE)**
  Average of the squared difference between the predicted and actual values
  Very sensitive to large errors
- **Mean Absolute Error (MAE)**
  Average of the absolute difference between the predicted and actual values
  Use when the is outliers to ignore and it is not sensitive to large errors

Hold-out sample is a random sample of observed data that is not used to fit the model