Regression Analysis

Introduction to Regression

Introduction

- Let us consider two variables, years of experience and salary of software engineers working in a company
- Years of experience is the independent variable and Salary is the dependent variable
- In regression we try to find a relationship between the dependent variable and the independent variable i.e. between the salary and the years of experience
- To do a regression, the relationship between the variables has to be linear

Introduction (contd.)

- Consider the following table which gives the salary of software engineers with different years of experience

Years of Exp.	Salary
0	300,000
1	400,000
2	500,000
4	700,000

- From this data, can we tell what is the expected salary of a software engineer with 3 years of experience?

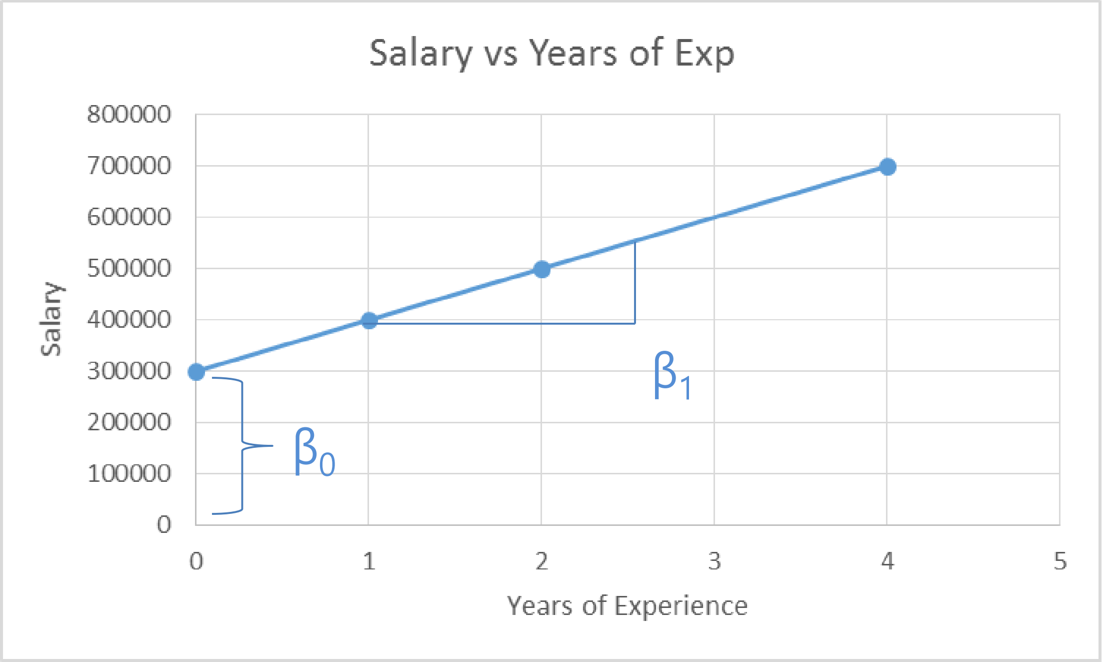
Introduction (contd.)

- Let us do a scatter plot of the data

- As you can see from the data, we can draw a straight line through the points

Regression Line

- The line is called the regression line



- The line is given by the equation:
- $\hat{y} = \beta_0 + \beta_1 (x)$
- β_0 is the y-intercept & β_1 is the slope of the line

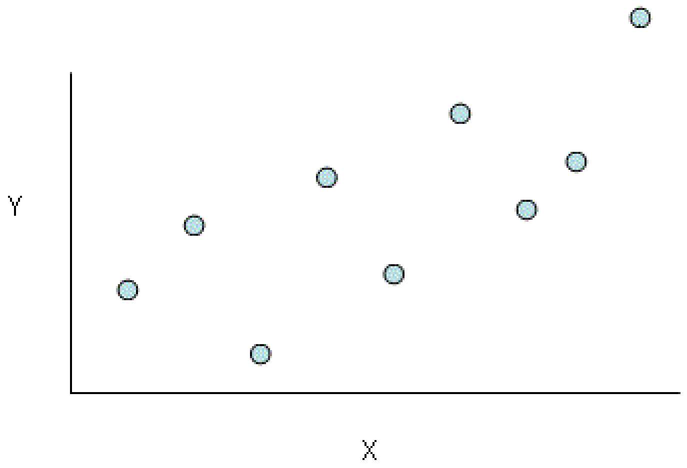
Regression Line (contd.)

- In our case, $\beta_0 = 300000$
- $\beta_1 = 100000$
- So the equation for our regression line is

salary = 300000 + 100000*(Years of Exp.)

Least Squares Fit

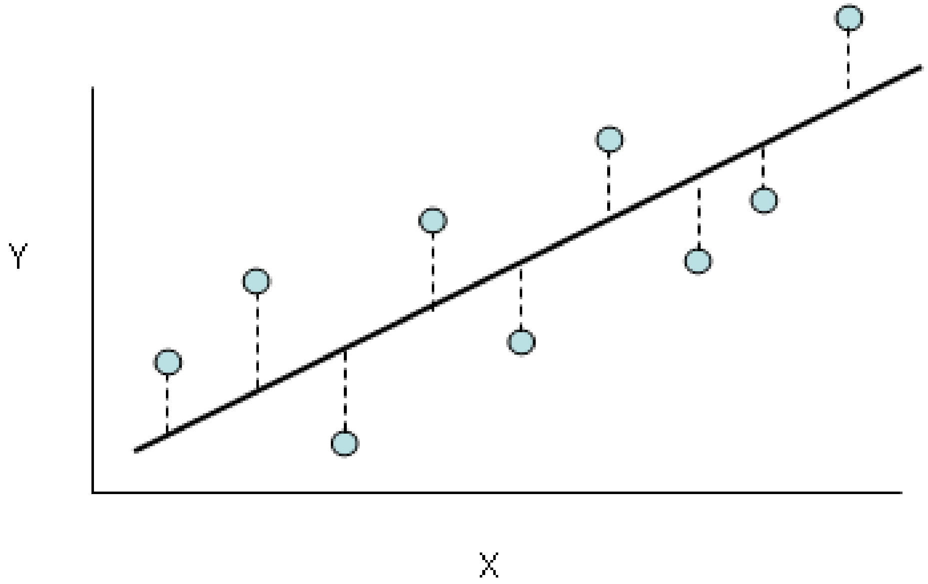
- In most practical cases, we won't able to fit all the points in a line
- For example consider the following data



- The relationship between dependent and independent variables is linear
- But we clearly can't fit a straight line through all the points

Least Squares Fit (contd.)

- Least squares fit allows us to draw a line that is close to all the points
- The difference between the line and the points should be minimum



Least Squares Fit (contd.)

- Every point in the graph can be expressed as (x_1,y_1), (x_2,y_2), (x_3,y_3) and so on..
- And for every point x_1, x_2, x_3.. we have the predicted values $\hat{y}_1$, $\hat{y}_2$, $\hat{y}_3$.. given by the regression equations $\hat{y}_1 = \beta_0 + \beta_1 (x_1)$, $\hat{y}_2 = \beta_0 + \beta_1 (x_2)$, $\hat{y}_3 = \beta_0 + \beta_1 (x_3)$…
- We need to reduce the square of the distan ce between y and $\hat{y}$
- So, for our regression line,

  $$\sum y_i - (\beta_0 + \beta_1 (x_i)) \text{ has to be minimum}$$
- We need to find β_0 and β_1 which minimizes the above function

Least Squares Fit (contd.)

- The β_0 and β_1 which minimizes the squared error is given as follows:

$$\hat{\beta}_1 = Cor(Y, X)\,\frac{Sd(Y)}{Sd(X)}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

- Where,
  - Cor(Y,X) is the correlation between Y and X
  - Sd is the standard deviation
  - $\overline{x}$ and $\overline{y}$ are the means of X and Y respectively

Covariance

- Covariance measures how two variables vary together
  - If one variable increases with increase in another variable, then covariance is positive
  - If one variable decreases with increase in another variable, then covariance is negative
- Covariance doesn't tell us anything about the strength of the relationship

$$\mathrm{Cov}(X, Y) = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{n - 1}$$

Correlation

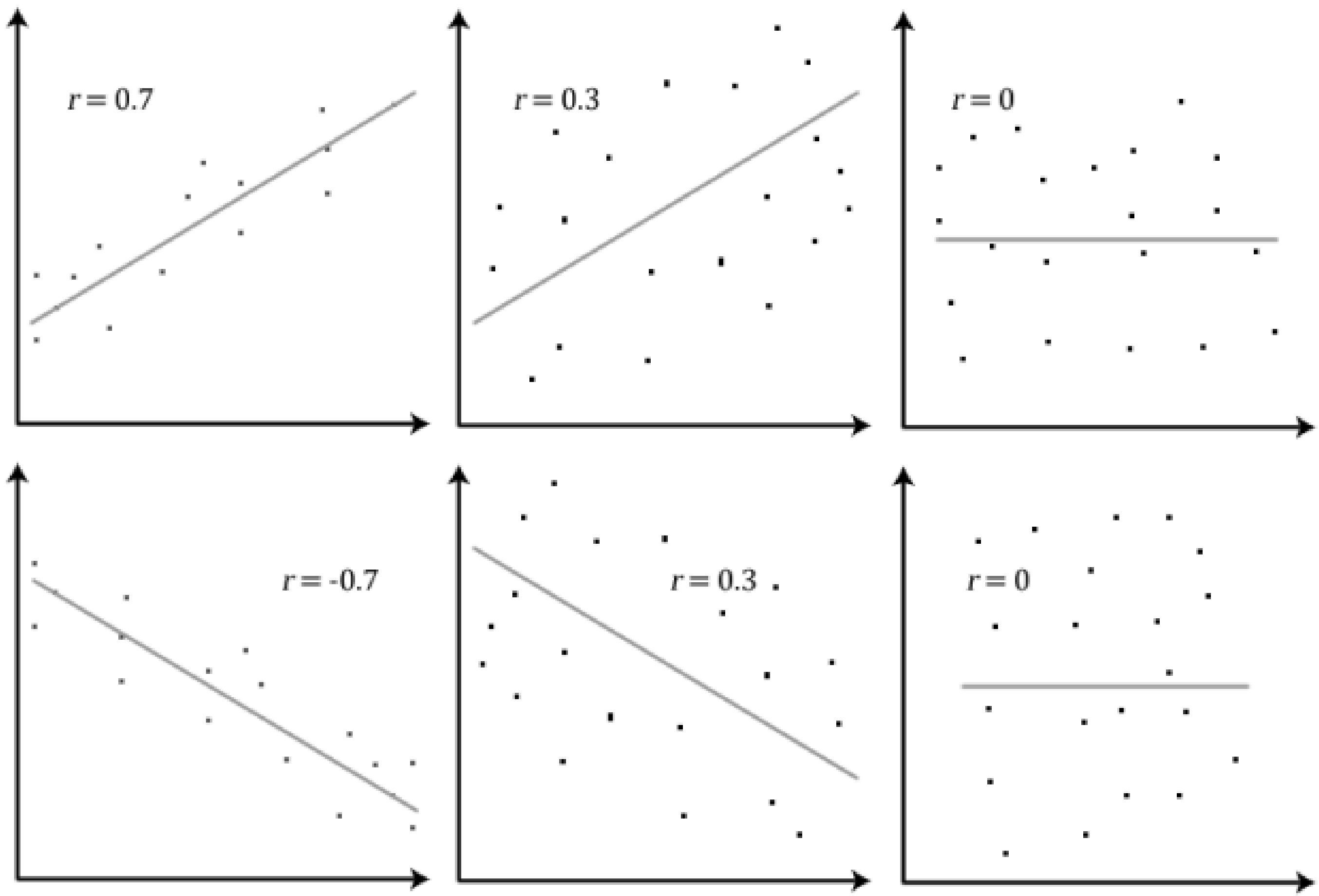
- Correlation is another measure that explains the relationship between two variables
- Correlation explains both the direction and the strength of the relationship
- Correlation ranges between -1 to +1

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

Correlation (contd.)

- Direction of relationship: If the value of correlation is
  - > 0, then the relationship is positive
  - < 0, then the relationship is negative
  - Close to 0, then no relationship
- Strength of relationship: If the value of correlation is
  - > 0.8 or < -0.8, then the relationship is strong
  - Between 0.4 to 0.8 or between -0.8 to -0.4, then the relationship is of medium strength
  - Between -0.4 to 0.4 then the relationship is weak

Correlation (contd.)



https://statsmethods.wordpress.com/2013/05/10/pearson-correlation-coefficient-r/

Simple Linear Regression

Simple Linear Regression

- Involves two variables one independent variable X and one dependent variable Y
- The dependent variable Y has to be a quantitative variable
- We need to find a line of best fit

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon$$

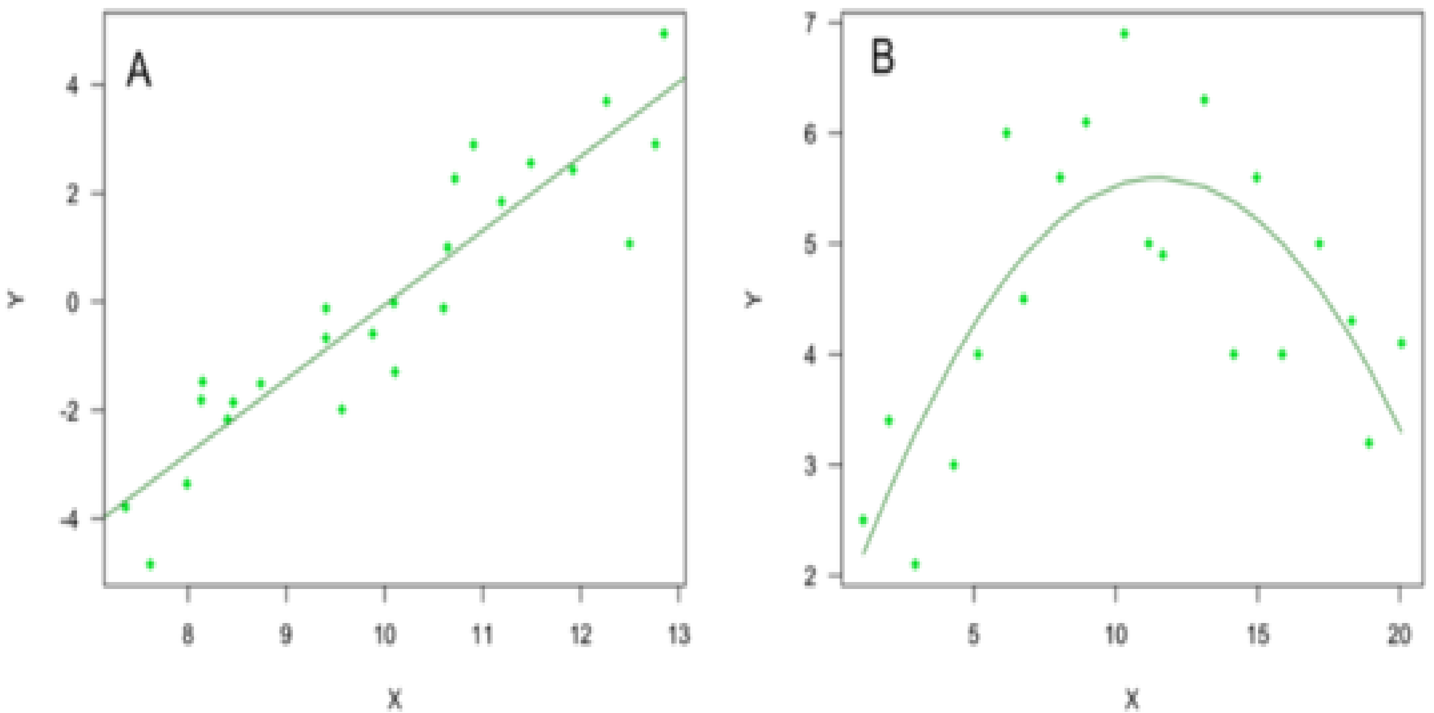
- We need to minimize the error and find and using least squares fit
- Using this equation we can find the predicted Y values using the equation

$$\hat{y}_i = \beta_0 + \beta_1 X_i$$

Assumptions of Linear Regression

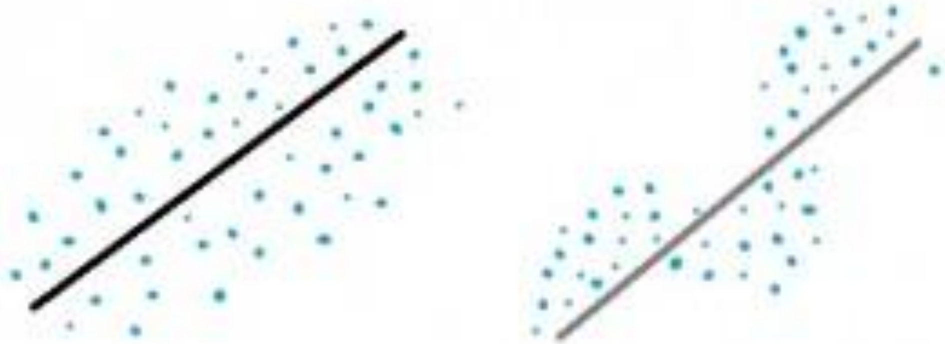
- The relationship between dependent and independent variables is linear
- The error term should be:
  - Normally distributed
  - Independent
  - Homoscedastic (the error term should have the same variance across different X values)

Linear and Non-Linear Relationships



http://www.edanzediting.com/content/study_planning_statistics

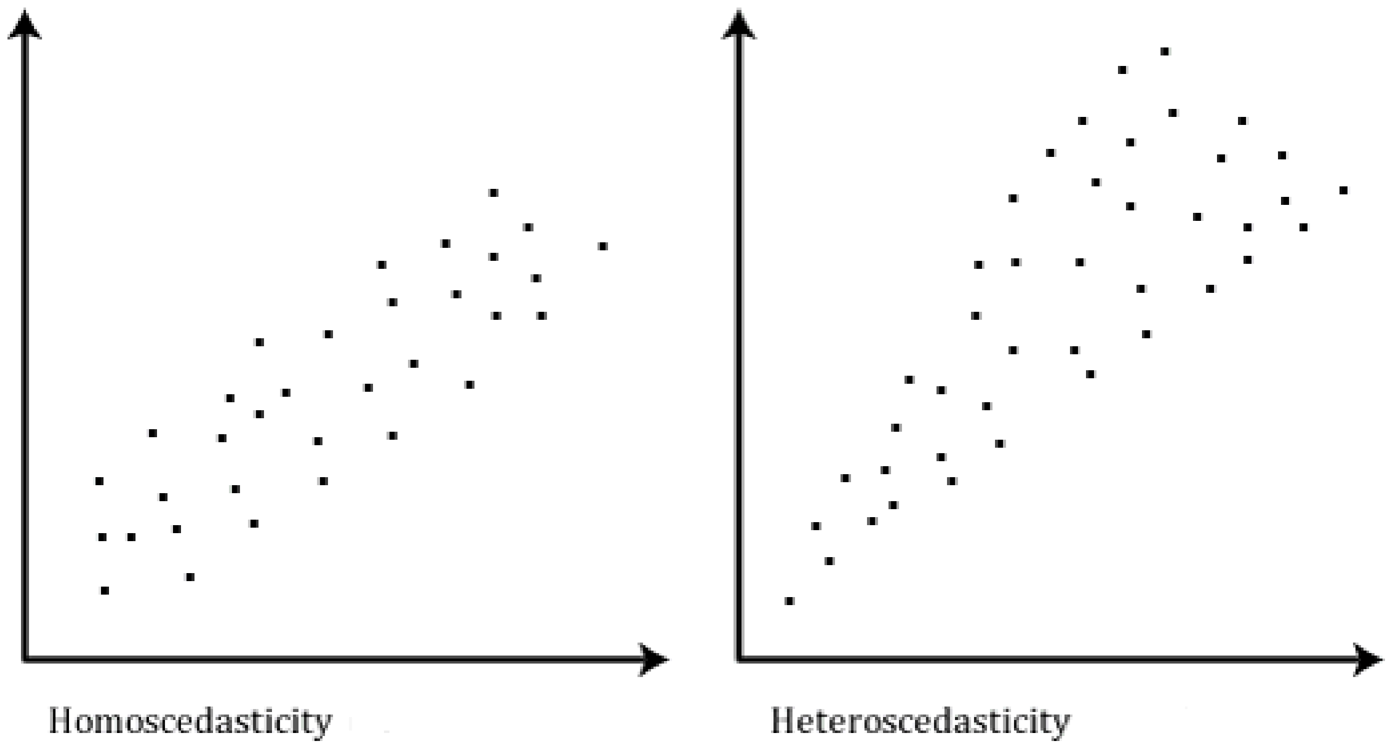
Independence of Errors



- The above figures show residual errors
- The errors in the first one are independent
- But in the second is not independent as we can see a trend

http://www.statisticshowto.com/assumptions-conditions-for-regression/

Homoscedasticity



https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide-2.php

Multiple Linear Regression

Multiple Linear Regression

- Involves one dependent variable (Y) and more than one independent variables ($X_1, X_2 \ldots X_n$)
- The goal is to find the relationship between independent variables and the dependent variable
- We need to find a line of best fit

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in} + \varepsilon$$

- We need to minimize the error and find and using least squares fit
- Using this equation we can find the predicted Y values using the equation

$$\hat{y}_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in}$$

Evaluation Criteria

Measures

- Mean Squared Error (Mean Squared Residual)
- Coefficient of Determination (R^2)
- Adjusted R2
- Mallow's C_p

Mean Squared Error

- Mean squared error is the average of the square of errors
- The model is good if the mean squared error is low

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$$

Coefficient of Determination (R^2)

- R^2 gives a measure of how much total variance in the data is explained by the model
- R^2 takes values in the range 0 to 1
- If R^2 is 1, then the regression line perfectly fits the data, if R^2 is 0, then the regression line doesn't fit the data at all

$$\textit{Coefficient of Determination} \rightarrow \quad R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

$$\textit{Sum of Squares Total} \rightarrow \quad SST = \sum (y - \bar{y})^2$$

$$\textit{Sum of Squares Regression} \rightarrow \quad SSR = \sum (y' - \bar{y'})^2$$

$$\textit{Sum of Squares Error} \rightarrow \quad SSE = \sum (y - y')^2$$

http://www.saedsayad.com/model_evaluation_r.htm

Adjusted R^2

- When we add more independent variables to a model, R^2 increases irrespective of whether the additional variables improve the model or not
- Adjusted R^2 penalizes the model if the new variable doesn't fit the model
- Adjusted R^2 can take any value unlike R^2 which can be within the range of 0 to 1

$$Adjusted\ R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

- Here p is the number of independent variables

Mallow's C_p

- Mallow's C_p is used in selecting the best regression model
- Mallow's C_p value should to be small and close to the number of independent variables used in the model

$$C_p = \frac{SSE_p}{S^2} - N + 2P,$$

- Here N is the sample size and P is the number of independent variables used in the model

Linear Regression t-test

- We can use the t-test to test whether each coefficient used in the linear model is 0
- Example:
  - The t-test with the intercept tests the null hypothesis H_0: $\beta_0 = 0$
  - The t-test with the first variable tests the null hypothesis H_0: $\beta_1 = 0$ to be true

Linear Regression Partial F-test

- Partial F-test is used to compare two linear models
- Let us say there are two models on the same data
- One with 3 variables var1, var2 and var3 and the other with only 2 variables var1 & var2
- Partial F-test between these two models tests the null hypothesis that the coefficient of var3 is 0
- i.e. $\beta_3 = 0$
- This is similar to the t-test
- However using partial F-test we can compare more than one variable

Partial F-test (contd.)

- For example, let us say one model has 5 variables var1, var2, var3, var4 and var5 and the second one has only 3 variables var1, var2 and var3
- Partial F-test between these two models tests the null hypothesis that both the coefficients of var4 and var5 = 0
- i.e. $\beta_4 = \beta_5 = 0$

Linear Model Selection

Linear Model Selection

- Involves the process of selecting a subset of relevant features or independent variables to be used in building a linear model
- The goals is to identify a subset of independent variables or predictors that influence the dependent variable and to fit a least squares model using the subset of independent variables
- If there are p independent variables, the number of possible models is 2^p
- We are going to study three methods:
  - Backward Elimination
  - Forward Selection
  - Stepwise Regression

Backward Elimination

- In this method, we start with all the independent variables or predictors in the model
- Remove the variable with highest p-value greater than the alpha value
- Refit the model after removing the variable
- Remove the next variable with highest p-value greater than the alpha value
- Refit the model and repeat the process until all the p-values are less than alpha

Forward Selection

- This is the reverse process of backward elimination
- For all the predictors which are not in the model, check their p-value if they are added to the model
- Choose the one with the lowest p-value which is lower than the alpha value
- From the rest of the predictors, check the p-value if they are added to the model
- Choose the one with the lowest p-value which is greater than the alpha value
- Repeat the process until there are no predictors with p-value lower than alpha

Stepwise Regression

- Combination of backward elimination and forward selection
- Helps us when we have added or removed a variable early in the process and we want to remove them at a later stage
- This is because the significance of one predictor might be influenced by the presence or absence of another
- In stepwise regression, after a new predictor is added, all other predictors which are already in the model are checked to see if the p-value falls below alpha level
- If the p-value of any of any of the predictors falls below alpha level then that predictor is removed before moving on to the next step

Multicollinearity

Multicollinearity

- A phenomenon in which two or more predictor variables in a multiple regression model are highly correlated
- Since one predictor variable is correlated with the other, it is possible to predict one with another using linear fit
- Multicollinearity can be detected using:
    - Correlation Matrix
    - Variation Inflation Factor (VIF)

Effects of multicollinearity

- Multicollinearity increases the uncertainty about the estimated coefficients and as a result the confidence intervals on the coefficients will be large
- Individual p-values can be misleading because of multicollinearity. A variable can have a high p-value even though it is important

Correlation Matrix

- Correlation Matrix offers a simple measure of multicollinearity
- High correlation between two or more predictor variables indicate presence of multicollinearity
- One disadvantage of using correlation to determine mutlicollinearity is that correlation is bivariate i.e. it measures correlation between two variables and it isn't particularly useful when one variable is a linear combination of many variables put together

Variation Inflation Factor

- Assesses how much the variance of an estimated regression coefficient increases if your predictors are correlated
- If none of the predictor variables are correlated, then VIF will be 1
- We say multicollinearity is present if the VIF of any variable is ≥ 5

$$\mathrm{VIF} = \frac{1}{1 - R_i^2}$$

- Where R_i^2 is the coefficient of determination of the linear regression where dependent variable is the ith predictor variable and the independent variables are the rest of the predictor variables

Dealing with multicollinearity

- Collect additional data
  - Additional data might break multicollinearity
- Remove predictor variables from the model
  - see if VIF of any of the predictor variable is > 5
  - If yes, then remove the predictor variable with the highest VIF from the model
  - Re-compute VIF values
  - Repeat the steps until VIF of all the predictor variables is below 5

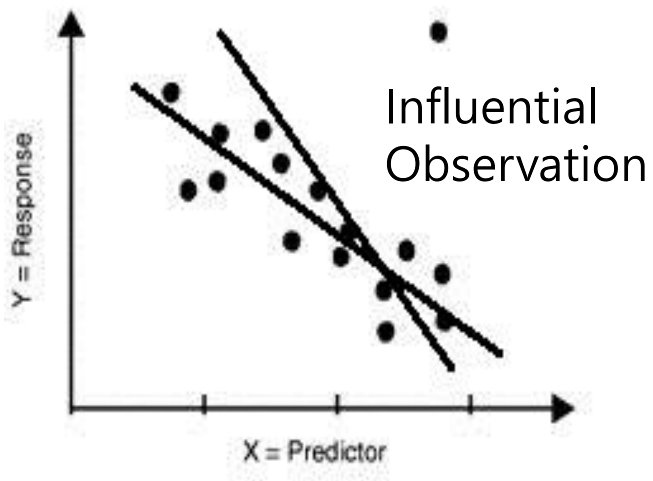
Handling Outliers

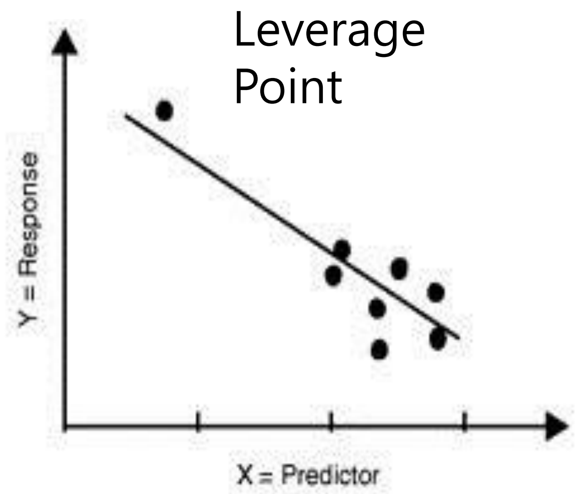
Leverage points & Influential observations

- Leverage points are those observations, which have extreme or outlying values of independent variable but graphically they lie close to the pattern described by other points
- Influential points also have extreme or outlying values of independent variable but they are far from the pattern described by other points
- Leverage points don't have much effect on the regression coefficients but Influential points have great influence on them
- Influential points in a regression can be detected using Cook's distance or Cook's D statistic

Leverage points & Influential observations (contd.)

- Examples of influential observation and leverage point





- There are two regression lines on the left. One is a fit obtained by including the influential observation in the regression and another without including the influential observation

http://www.bexcellence.org/Regression.html

Cook’s distance

- Cook’s distance is used estimate the influence of a data point in linear regression
- Cook’s distance for the ith observation is:

$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})^2}{p \, \mathrm{MSE}},$$

- Where,
    - $\hat{Y}_j$ is the prediction from the full regression model for observation j
    - $\hat{Y}_{j(i)}$ is the prediction for observation j from a refitted regression model in which observation i has been omitted
    - p is the number of fitted parameters in the model
    - MSE is the mean square error of the model

Cook's distance (contd.)

- The generally accepted cut-off values for spotting highly influential points is $D_i > 4/(n-k-1)$ where n is the number of observations, k is the number of independent variables
- Can we remove observations with $D_i > 4/(n-k-1)$ from our regression model?
  - We can remove them if we have enough evidence that the influential observation was due to measurement error or other errors
  - If they are genuine observations, we need to include them in our model even though they are influential