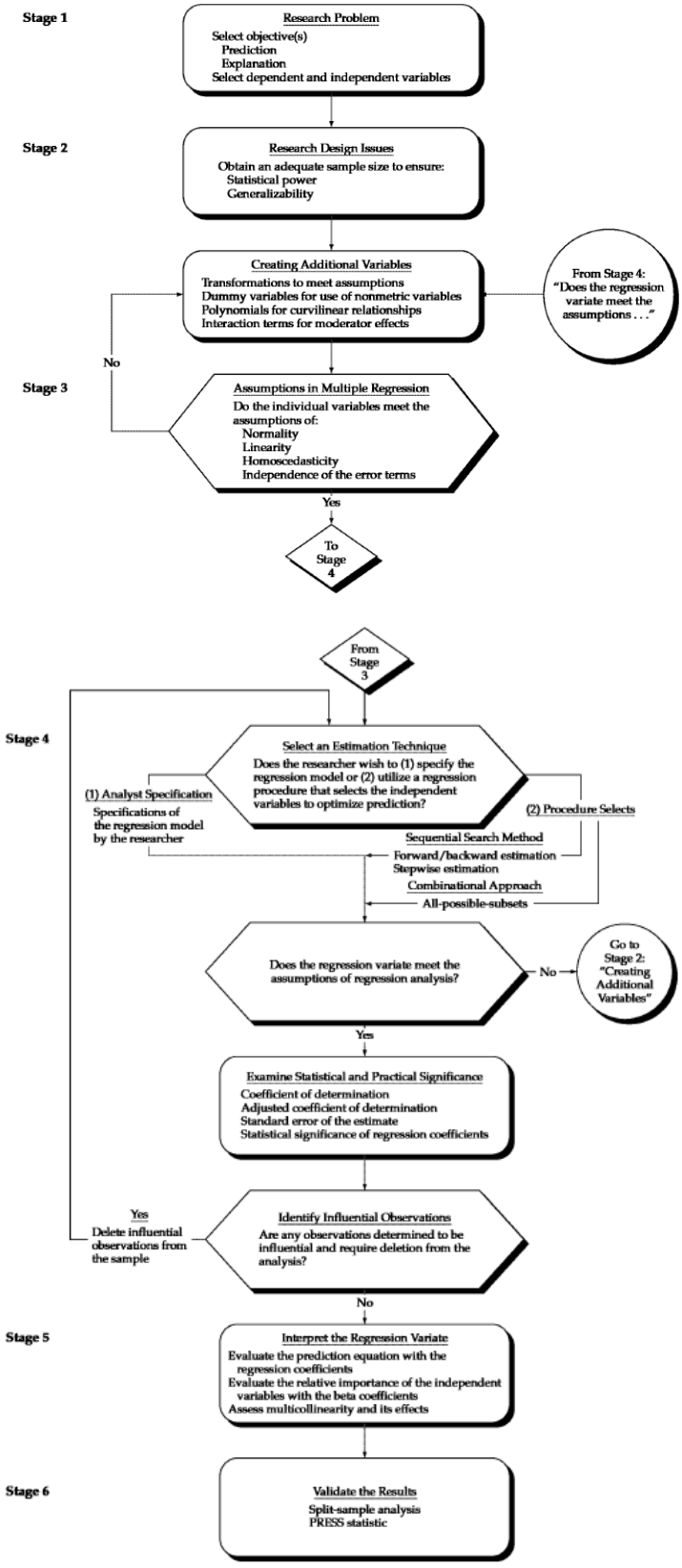
Multiple Linear Regression
(Concept Note)

1. Regression Modeling Algorithm



2. **Objectives of Model Fitting**
    a. Understanding the relationship between the variables [Statistical Approach]
    b. Predicting the outcome of new cases [Data Mining Approach]

3. **Applications**
    a. Predicting customer activity on credit cards from their demographics and historical activity patterns
    b. Predicting the time to failure of equipment based on utilization and environment conditions
    c. Predicting expenditures on vacation travel based on historical frequent flyer data
    d. Predicting staffing requirements at help desks based on historical data and production and sales information
    e. Predicting sales from cross selling of products from historical information
    f. Predicting the impact of discounts on sales in retail outlets

4. **Concept of Ordinary Least Squares**
    a. R^2 and Adj. R^2.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

$$Adj.R^2 = 1 - \frac{n-1}{n-p-1}(1 - R^2)$$

    b. F-Test
    c. t-tests, Standardized beta and Significance values
    d. Variate

$$\text{Variate value} = w_1X_1 + w_2X_2 + w_3X_3 + \ldots\ldots + w_nX_n$$

In Multiple Linear Regression the variate is so determined so as to best correlate with the variable being predicted.

    e. Criteria for subset selection (in case of predictive modeling) – Mallow's C*p*

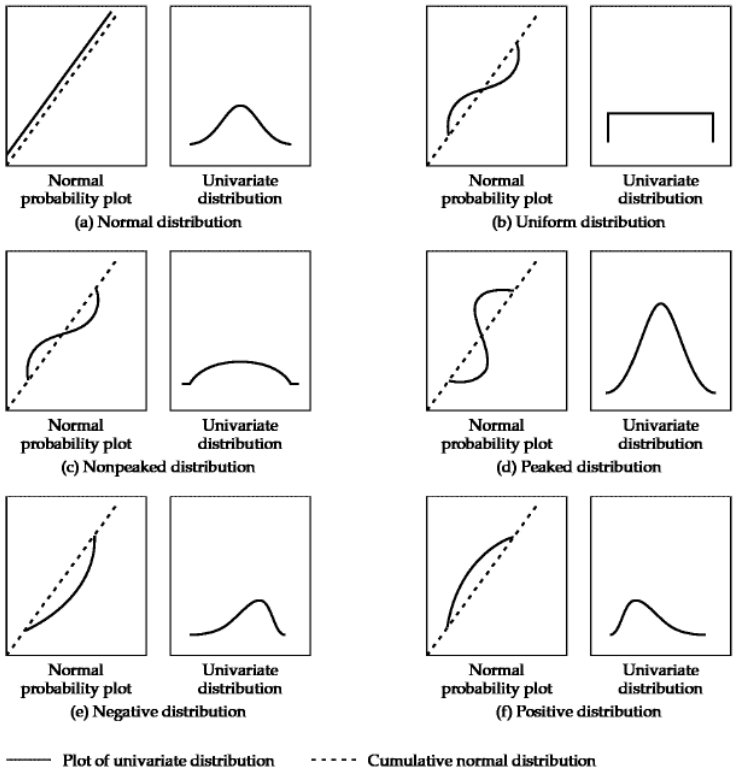
$$C_p = \frac{SSE}{\hat{\sigma}^2_{full}} + 2(p+1) - n$$

5. **Explanatory Vs Predictive Modeling**
    a. Important to have a model that predicts better on new values rather than a model that fits well on the data
    b. Division of dataset into training and test / validation dataset
    c. Split the dataset using seed
    d. Performance is measured by predictive accuracy (how well the model predicts new cases)

6. **Assumptions in Regression Modelling**
    a. Check for influential observations
        i. Check for Mahalanobis Distance (Check the option in Save tab in Regression). Compare with Critical Chi-square: $x^2_{0.05,\,k}$ where 'k' is the number of explanatory variables. Value greater than are possible outliers
        ii. Check for Cooks's distance. Value above 4/(n-(k+1)) are possible outliers, where k is the number of explanatory variables and n is the number of observations.

    b. Examine Heteroscedasticity (variance of residuals should be homogeneous across levels of predicted values)
        i. Examine residual Plot (Plot between standardized residuals with predicted value)
        ii. Remedy is to Transform predictors

    c. Multivariate Normality (The noise or the dependent variable follows a normal distribution)
        i. If the variation from normality is sufficiently large, all statistical tests are invalid
        ii. Check using Histogram of unstandardized residual
        iii. Q-Q Plot of unstandardized residual
        iv. Skewness and Kurtosis should be near zero
        v. Kolmogorov-Smirnov Test (Should be insignificant for normality)
        vi. Shapiro-Wilk's Test (Should be insignificant for normality)
        vii. Normality assumption may be relaxed when split sample validation is done (Predictive Modeling)
        viii. REMEDY: Transformation (Also check other assumptions first)



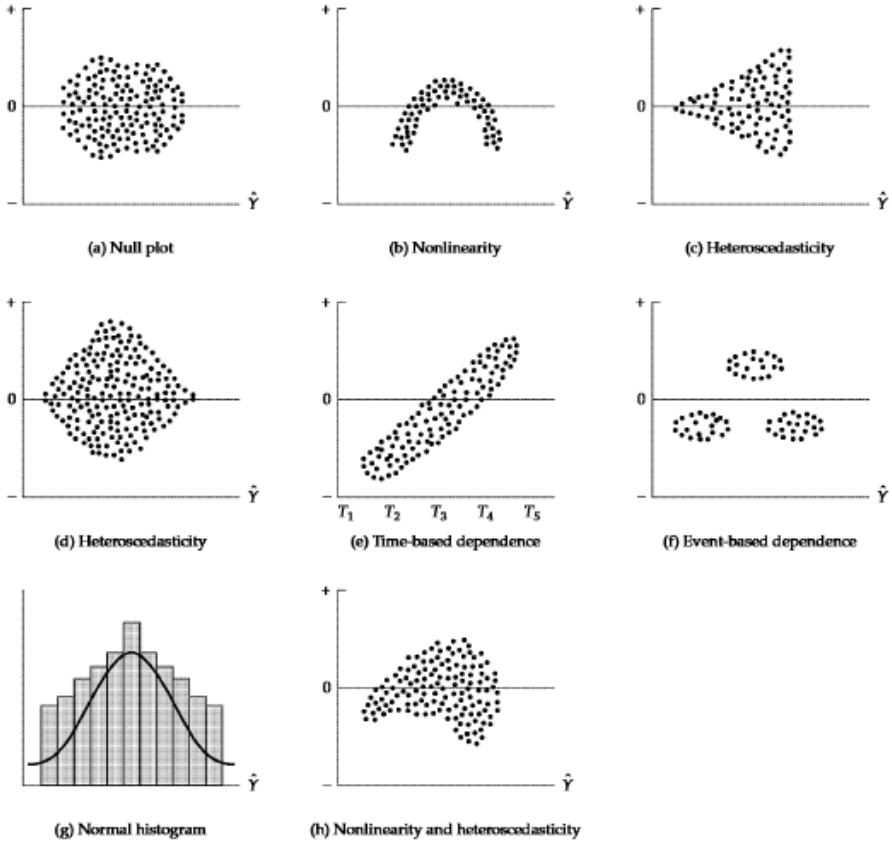
d. Multicollinearity
    i. Bivariate Correlations: If correlations are greater than 0.8, multicollinearity is very like to exist. Ok if less than 0.6.
    ii. Tolerance and VIF values: VIF values greater than 4 indicate possible multicollinearity
    iii. Collinearity Diagnostics (Check for variance proportions): Condition index greater than 30 indicate serious multicollinearity
    iv. An excluded variable may be tested for its possible inclusion by checking its actual t-value = sqrt(VIF)*t-value

e. Non-Linearity (Variables are linearly related to the dependent variable)
    i. Scatterplots
    ii. REMEDY: Transformations (Usually Log Transformation)

f. Model Mis-specification
    i. Take predicted value and squared predicted value as predictors of the actual value and run the regression. If the squared of predicted value is significant then the model is mis-specified and more variables need to be added.

g. Absence of correlated errors (The cases are independent of each other)
    i. Durbin-Watson Statistic: The Durbin-Watson statistic has a range from 0 to 4 with a midpoint of 2. 2 implies no autocorrelation. Value below 2 is positive autocorrelation and value above 2 is negative autocorrelation
    ii. REMEDY: Include the omitted causal factor into the multivariate analysis

h. Test for Linearity, Homoscedasticity and Correlated Errors
    i. Plot of Studentized Residual Vs Predicted Dependent values



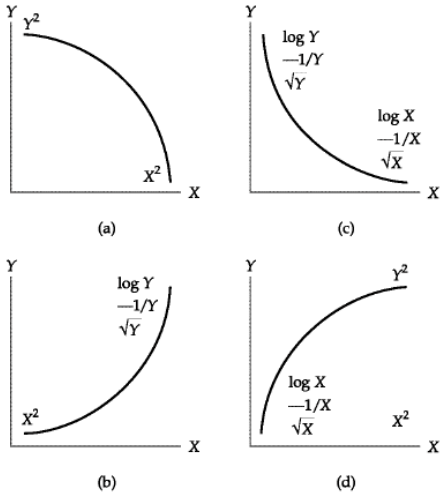
7. Transformations

a. For non-normal distributions, the two most common patterns are 'flat' distributions and 'skewed' distributions. For the flat distribution, the most common transformation is the inverse transformation ($1/y$, or $1/x$).

b. Skewed distributions can be transformed by taking the square root, logarithms or even the inverse of the variable. Negatively skewed distributions are best transformed by using a square root transformation and positively skewed distributions are best transformed by using logarithmic transformation.

c. Heteroscedasticity: If the cone in residuals opens to the right, take the inverse transformation. If the cone opens to the left, take the square root transformation

d. Some transformations to achieve linearity are shown below:



8. Model Validation (On Test / Validation Data Set)

a. Root Mean Square Error: $RMSE = \sqrt{\sum_{i=1}^{n}\left(\frac{Predicted_i - Actual_i}{n}\right)^2}$

b. Mean Absolute Percentage Error: $MAPE = \sum_{i=1}^{n}\left(\frac{Predicted_i - Actual_i}{Actual_i}\right)$