Correlation and Regression

Covariance is statistical measure of the degree to which two random variables moving together. Positive Covariance is means two variable move in same direction, Negative is opposite to each other, Zero is no relation between two variables.

$$COV(xy) = \frac{\sum (x - \bar{x})(y - \bar{y})}{n-1}$$

Correlation coefficient measures strength of linear relationship between two variables. Have no units and range from -1 to +1.

$$r(xy) = \frac{\text{cov}(x, y)}{s_x s_y}$$

Correlation 1 represents variables are perfectly correlated, -1 is perfectly negatively correlated.

Outliers: Extreme observations in the sample have very big or very small values.

Spurious Correlation: Appearance of linear relationship by coincidence.

Non linear relationships can not be measured with correlation.

A t-test to determine correlation is statistically significant is

$t = \frac{r(\sqrt{n-2})}{\sqrt{(1-r^2)}}$; Significance is accepted test stat is outside the critical range with n-2 degrees of freedom.

Assumptions underlying linear regression:

- A linear relationship exists among dependent and independent variables.
- Independent variable is not correlated with residuals.
- Expected value of the residue is zero.
- Variance of residue term is constant.
- Residual tern is independently distributed.
- Residual term is normally distributed.

Linear regression model: $Y = b_0 + b_1 x + \varepsilon$

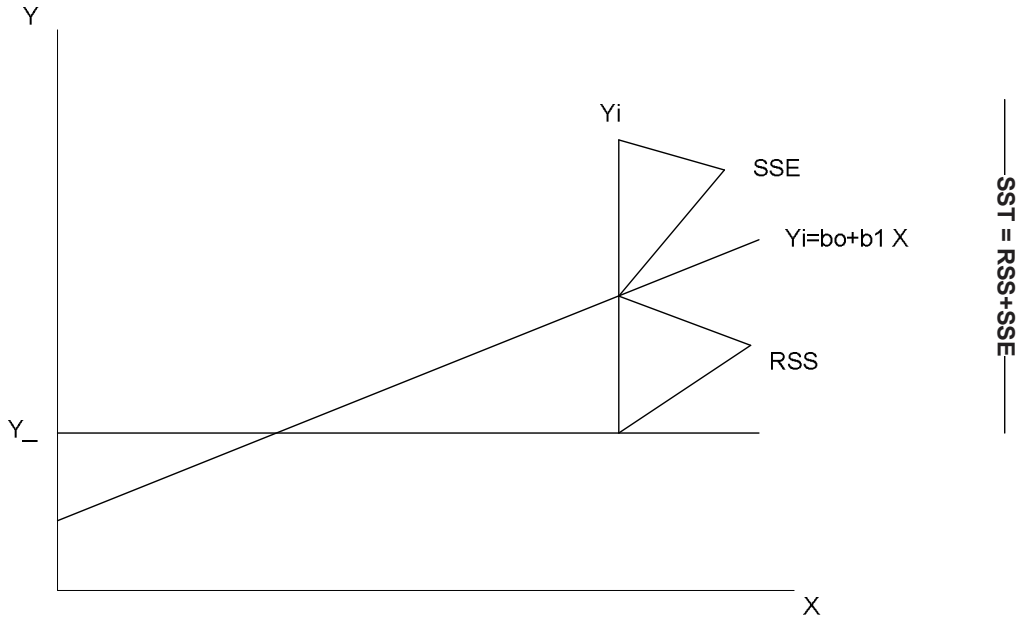
Coefficient of determination (R^2) is percentage of dependents variation explained by independent variable. It is square of correlation coefficient.

Confidence interval for regression coefficient b_1 is $b_1 \pm std.Err$

Hypothesis test of estimated regression parameter t-test with n-2 degrees of freedom is

$$t = \frac{b_1^{\wedge} - b1}{s_{b1}}$$

Confidence interval for Y-value is = $Y^{\wedge} \pm t * s_f$



ANOVA table:

ANOVA			
Source of Variation	**DOF**	**Sum Squares**	**Mean Sum Squares**
Regression (Explained)	**1**	**RSS**	**MSR**
Error (Unexplained)	**n-2**	**SSE**	**MSE**
Total	**n-1**	**SST**	

SST is total variation of independent variable from mean Y value.

RSS Regression Sum Square: measures variation of dependent variable explained by independent variable. It is the sum square distance between predicted Y and mean Y value.

SSE Sum Squared Error or sum squared residual: measures unexplained variation in independent variable. It is sum square distance between actual dependent variable and predicted value.

SST is sum of RSS and SSE

Coefficient of determination (R^2) and standard error estimate (SEE) calculated as

$$R^2 = \frac{TotalVariation(SST) - Un\exp ainedVariation(SSE)}{TotalVariation(SST)}$$

$$R^2 = \frac{ExpainedVariation(RSS)}{TotalVariation(SST)}$$

SEE (standard deviation of regression error) = $\sqrt{MSE}$ = $\sqrt{MSE}/n-2$

F Statistic is used to test how well a set of independent variables explain variation in dependent variable. It is a one tailed test.

$$F = \frac{MSR}{MSE}$$

Limitations of regression Analysis:

Parameter instability
Limited usefulness of regression, from public data.
Heteroskedasticity and auto correlation.