*Himesh Buch*

*CS [463]: Regression Methods*

* If the value of the parameter specified by the null hypothesis is contained in the 95% interval, then the null hypothesis cannot be rejected at the 0.05 level. If the value specified by the null hypothesis is not in the interval, then the null hypothesis can be rejected at the 0.05 level.
* Beta = (X’X)-1 \* (X’y)
* Residuals = y-y(hat)
* Read: Variable Selection Methods\_using\_olsrr
* In statistics, overfitting is "the production of an analysis that corresponds too closely or exactly to a particular set of data and may therefore fail to fit additional data or predict future observations reliably". An overfitted model is a statistical model that contains more parameters than can be justified by the data
* If we are building a model with p terms, then adding more terms will increase R2
* If we are building a model with p terms, then adding more terms will not necessarily increase R2 adj
* We want both R2 and R2adj to be as big as possible and minimize MSE (mean squared error)
* Residual mean square decreases as p increases
* Cp is a measure of variance in the fitted values and (bias)2. Small Cp values are desirable. If Cp >> p then there is a significant bias
* If there are k regressors or terms, then we will have to investigate 2k possible regression equations
* Forward selection: no variables in the data initially, but they are added one at a time
* Backward elimination: variables already exists, examines them one at a time and removes it if not necessary
* Stepwise regression: kind of like forward selection but removes the variable if it is not significant
* Leverage point: if the x value is way higher than other x values
* Influential point: if removing some point changes the fitted line drastically than it’s an influential point. You basically compare the regression lines before and after removing the point, and see if the deflection is noticeable than that is an influential point
* For a pth order polynomial regression model, E(y) = beta0 + beta1x + beta2x2 + ... + betapxp, you need n to be p+1
* Multicollinearity exists when two or more independent variables are highly or moderately corelated