Here are some topics that you should understand.

**1) The core focus of this course consists of linear regression (ordinary least squares regression) and topics from multivariate analysis (principal components analysis and cluster analysis). You should understand the underlying assumptions of each model or method, how to specify or apply each model or method, and how to interpret the results from each model or method.**

1. note initial assumption

2. calculate estimate intercept and slopes

3. examine how well the model fits

Look at correlation and slope (closely related to two variables). Correlation depends on the units of X and Y, outliers

Regressions predictions based on x will always be at least as good as using just the mean bc line has to go through the mean

a and b estimates that minimize squared residuals

Equation to predict intercept and slope:

1. assumption of linear model
   1. minimize the residuals within a dataset
      1. Model to estimate linear relationship
      2. e = y – y knot
      3. b =

Assume relationship is linear

Assume both A and B are true intercepts

Assume true error term, portion not explained by A or b

SSE sum of squared errors:

Sum(yi-yhati)^2

Y hat = model

“better predictor” of the data = less residual or “error” between predicted y and actual y

Total sum of squares = sum(y hat – mean(y) ) ^ 2

* if we are using mean of y to predict actual values of y. lower the better, less error
* equals regression sum of squares + error sum of squares

Regression sum of squares = sum( yhat – mean(y))^2

* distance between predicted values and actual mean. Total error is due to difference of line and mean

Error sum of squares = sum(y – yhat)^2

* actual to predicted value

Residuals can show if there is a linear relationship or not

We also assume the error term has mean = 0

Assume that e and x are uncorrelated

* If the relationship is linear

Check the residual plot for any patterns, want to be a scatter plot around e = 0

<https://statisticsbyjim.com/regression/ols-linear-regression-assumptions/>

<https://statistics.laerd.com/spss-tutorials/principal-components-analysis-pca-using-spss-statistics.php>

Linear regressions assume Both homoscedastic and normally distributed residuals.

**2) You should understand the concept of multicollinearity. How do we detect multicollinearity, how does it affect our regression results, and how do we mitigate multicollinearity?**

What X values don’t have an effect on Y, suspect and T-tests fail to reject null hypothesis.

<https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/>

**3) When developing predictive models, we typically build and compare alternative models. What statistical metrics can be used to compare the alternative models? What statistical tests can be used to compare models? Which metrics are only valid when comparing nested models and which metrics can be used to compare any set of models? You should understand the underlying assumptions of these metrics and tests and be able to interpret these metrics and tests. Metrics of particular interest include: mean error, mean absolute error, mean square error, R-squared, adjusted R-squared, Mallow’s Cp, AIC, and BIC.**

<http://www.sthda.com/english/articles/38-regression-model-validation/158-regression-model-accuracy-metrics-r-square-aic-bic-cp-and-more/>

**4) You should understand how to perform principal components analysis, and how could you apply PCA in conjunction with cluster analysis. You should understand how to select the number of components to retain under different selection rules, and why we would use PCA.**

**5) You should understand in complete detail how forward and backward variable selection work in the context of OLS regression. How does each method work at each step and what statistical test is being performed at each step?**

**6) You should understand how the F-test is used to test nested models.**

F test tells us that we can safely reject the null hypothesis that all the coefficients are zero. Using the mean is better than the regression model

Nested model tests: f test allows us to compare the influence of subsets of regressors relative to a larger encompassing model

**7) You should understand how to define, use, and interpret dummy variables as a means of coding categorical variables for use in a regression model.**

**8) You need to be able to perform all of the computations in Assignment #4.**

How many observations are in the sample data?

**To determine this, we need to calculate the observations as n = df + p + 1. The variables indicate that df = degrees of freedom, p = predictor variables, and n = number of observations.**

Write out the null and alternate hypotheses for the t-test for Beta1.

**Full explicit model:**

**Reduced Model: while defining the null and alternative hypotheses as:**

**H0: Beta 1 = 0, which states that the null hypothesis for Beta 1 is 0 and X1 has no meaningful contribution to the prediction of the response variable**

**Ha: Beta 1 ≠ 0, which states that the alternative hypothesis for Beta 1 is not 0 and that it has statistically significant effect on the prediction of the response variable**

Compute the t- statistic for Beta1.

**T-statistic calculation: Estimate / Standard Error**

**Beta 1 T-Statistic: 2.18604/.41043 = 5.3262**

Compute the R-Squared value for Model 1.

**R-Squared value calculation: Sum Squared of Residuals (SSR) / Total Sum of Squares (SSTO)**

Compute the Adjusted R-Squared value for

**Adjusted R-Squared calculation:**

Write out the null and alternate hypotheses for the Overall F-test.

**Test: hypothesis that all predictor variables have no explanatory influence which would leave each coefficient equal to zero.**

**Full Model (Ha):**

**Reduced Model (H0):**

**There needs to be confirmation that at least one coefficient does not equal 0.**

Compute the F-statistic for the Overall F-test.

**F-Stat calculation: Mean Square due to Regression (MSR) / Mean Square due to Error (MSE)**

**F-Stat for Model 1: 531.50226/9.40835 = 56.4926**

Now let’s consider Model 1 and Model 2 as a pair of models. Does Model 1 nest Model 2 or does Model 2 nest Model 1? Explain.

**If we look at the reduced models of both model 1 (M1) and model 2 (m2), it appears model 1 nests model 2. Reduced model 1 has less predictors than model 2. Model 1 also excludes predictors, which would make this a unique case. To see if model 1 would fit better than model 2, the F-Test would be performed.**

**Model 1 and model 2 would be compared based on their full and reduced models to see if their independent variables are statistically significant. The p-values for each model suggests which variables can be determined to fit positively to the regression.**

Now let’s consider Model 1 and Model 2 as a pair of models. Does Model 1 nest Model 2 or does Model 2 nest Model 1? Explain.

**If we look at the reduced models of both model 1 (M1) and model 2 (m2), it appears model 1 nests model 2. Reduced model 1 has less predictors than model 2. Model 1 also excludes predictors, which would make this a unique case. To see if model 1 would fit better than model 2, the F-Test would be performed.**

**Model 1 and model 2 would be compared based on their full and reduced models to see if their independent variables are statistically significant. The p-values for each model suggests which variables can be determined to fit positively to the regression.**

The term ‘simple linear regression model’ refers to a linear regression model with one predictor variable.

The term ‘linear model’ refers to any model that is linear in the parameters. It does not require that the relationship between X and Y is a linear relationship.

Regression models are intended to be used for interpolation, not extrapolation, and hence regression models should not be used to predict response values for values of the predictor variables outside of the range used in the estimation. This is easy to understand in the case of simple linear regression, but more difficult in the case of multiple linear regression.

The regression residual is computed as e(i) = Y(i) – Y-hat(i), i.e. the residual is always computed as the actual value minus the predicted value.

The R-Squared value for the no intercept model is computed with a similar formula as in the case of the intercept model. The formula is the same if you set Y-bar equal to zero, even if Y-bar is not equal to zero in the data. This is where the difference occurs. The R-Squared value for the no intercept model assumes a sample mean of zero, or ties the 'center' to the origin and not the overall mean.

R-Squared is monotonic in the number of predictor variables. Adding variables to an existing model will never cause the R-Squared to decrease.

Regression coefficients are estimated with respect to the other predictor variables (and their coefficients) in the model, hence they are sometimes called ‘partial regression coefficients’ since they represent the partial effect of the predictor variable.

Since R-Squared is monotonic in the number of predictor variables, we would always choose the model with the larger number of predictor variables. When comparing models of different sizes, we should use a metric like adjusted R-Squared, which provides a trade-off between model fit and model complexity.

Under the assumption of normality the parameter estimates of a linear regression model will be the same whether estimated by maximum likelihood or least squares.

The term X1\*X2 is the interaction term. In this case these three predictor variables fit a surface to the response variable Y.

Often times, residual plots as well as other plots of the data will suggest some difficulties or abnormalities in the data. Which of the following statements are not considered difficulties?

* The variance of the error term (and of Y) is constant

The Analysis of Variance (ANOVA) table in linear regression can be used to compute the R-Squared, Adjusted R-Squared, and the Overall F statistic

The hat matrix is given by X\*inv(X’X)\*X’

Beta calculation: Beta1 = COV[X,Y] / VAR[X]

Beta0 intercept: Y-Bar - Beta1 \* X-bar

Diagnostics for assessing the Goodness-of-Fit for a linear regression model include Plots of Y-hat versus Y, a Quantile-Quantile plot of the residuals, and Y against each continuous predictor variable

Heteroscedasticity can be detected graphically by plotting the residuals against the in-sample predicted value Y-hat by visualizing these shapes: a funnel, a double bow, or any nonlinear pattern