***Himesh Buch***

***CS 486***

***Notes***

# Extra:­­

* simple linear regression is one quantitative variable predicting another
* normal distribution means most of the data is closer and symmetrical around the mean
* multiple linear regression is basically the same as simple linear but with more independent vars
* OLS = ordinary Least Square Regression = Simple Linear Regression
* you always want to minimize MSE in any regression model
* Regression analysis seeks to find the relationship between one or more independent variables and a dependent variable

SSR = Sum of squares for regression = sigma(yi(hat) – y(bar))2 => df: 1

SST = Sum of squares of total = sigma(yi – y(bar))2 => df: n-1

SSRes = Sum of squares of residuals = sigma(yi – yi(hat))2 => df: n-2

MSR = SSR and MSRes = SSRes / n-2; F0 = MSR / MSRes

MSE = sigma(yi -yi(hat))^2 = SSE

## Logistic regression

Deals with the quantitative data and binary data will work as well. It models the probability of an event occurring depending on the values of independent variables. Estimates the probability of an event occurring for a randomly selected observation vs the probability of it not occurring. *Logistic regression is a method for fitting a regression curve, y = f(x), when y is a categorical variable. The typical use of this model is predicting y given a set of predictors x. The predictors can be continuous, categorical or a mix of both. The categorical variable y, in general, can assume different values. In the simplest case scenario y is binary meaning that it can assume either the value 1 or 0. A classic example used in machine learning is email classification: given a set of attributes for each email such as number of words, links and pictures, the algorithm should decide whether the email is spam (1) or not (0). In this post we call the model “binomial logistic regression”, since the variable to predict is binary, however, logistic regression can also be used to predict a dependent variable which can assume more than 2 values*

# Maximum likelihood test:

The goal is to find the optimal way to find fitted distribution of the data. You try to find the optimal value for mean and/or sd for a distribution assuming it is normal. And in normal distribution the observations are close to the mean and if it is not you shift it until they are. We basically shift the mean of the data to where most of the datapoints are and that means that that is the place where the likelihood of most of the data to be is high. Meaning, we find the location of mean which maximizes the likelihood of the observations. The same process applies for standard deviation (<https://www.youtube.com/watch?v=XepXtl9YKwc>)

# Maximum-Likelihood Estimation:

Maximum-Likelihood Estimation (MLE) is a statistical technique for estimating model parameters. It basically sets out to answer the question: what model parameters are most likely to characterize a given set of data? First you need to select a model for the data. And the model must have one or more (unknown) parameters. As the name implies, MLE proceeds to maximize a likelihood function, which in turn maximizes the agreement between the model and the data.

# Newton-Raphson method:

used to find 0’s of a function. Only drawback is that it sometimes doesn’t converge to root. Basic formula => xn+1 = xn – f(xn)/f’(xn). you keep on doing this until xn+1 - xn is the smallest or you might keep on getting the same answers for the xn that you compute, stop then. It won’t be useful when the derivative of the function is close to 0. Guess or x values on x axes, function value on y axis

# Permutation Tests:

it gives an alternative that needs no assumption of normality. It’s not random.

# Monte-Carlo Simulations:

Monte Carlo Simulation Methods are used in situations where an exact solution or a closed form to a problem is not available. Here are the steps:

1. Describe the experiment and all of its possible outcomes
2. Characterize the probabilities associated with each outcome
3. Match these probabilities up with what is produced by some random number generator
4. Generate a large number of random experiments according to this rule

# Negative Binomial Distribution:

read the pdf in lec2 folder. Binomial distribution describes the number of successes k achieved in n trials, where probability of success is p. Negative binomial distribution describes the number of successes k until observing r failures (so any number of trials greater than r is possible), where probability of success is p

# Trimmed Mean:

The trimmed mean is the mean with some extreme values taken out.; for example, in the 10% trimmed mean the largest and smallest 10% of the values are removed and then the mean is taken on the remaining 80%

# Winsorized Mean:

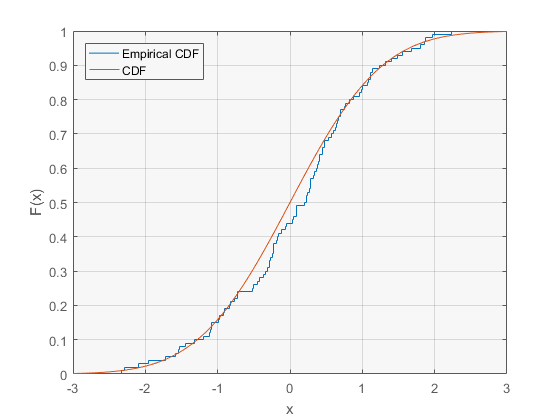
The Winsorized mean is similar to the trimmed mean, except that rather than deleting the extreme values, they are set equal to the next largest (or smallest) value.

# CDF (Cumulative Distribution Function):

These functions are used to calculate the area under the curve, or to calculate the area left to some point of interest. When the probability distribution is continuous, probability = (area under the curve). Total area is always 1. You get the CDF function by integrating the given function in the limit of whatever, could be (0) to (point of interest) (<https://www.youtube.com/watch?v=3xAIWiTJCvE>)

# ECDF vs CDF:

Empirical is something you build from data and observations. CDF basically shows the shape which is theoretical as it is not made by looking at the data. Look at the given graph:



# Bootstrapping:

A sampling method. When you don’t know the population parameter and you want to do any sampling methods you use bootstrapping. Conducted with replacement. Here are the steps:

1. Randomly pick something from the data
2. Take note of it
3. Put it back

Repeat the process for large amount of time

# GLM (Generalized linear model):

In a generalized linear model (GLM), each outcome Y of the dependent variables is assumed to be generated from a particular distribution in an exponential family

# Likelihood ratio test:

The LR test is performed by estimating two models and comparing the fit of one model to the fit of the other. Removing predictor variables from a model will almost always make the model fit less well (i.e., a model will have a lower log likelihood), but it is necessary to test whether the observed difference in model fit is statistically significant. The LR test does this by comparing the log likelihoods of the two models, if this difference is statistically significant, then the less restrictive model (the one with more variables) is said to fit the data significantly better than the more restrictive model. If one has the log likelihoods from the models, the LR test is fairly easy to calculate. Here is the formula:

LR=−2ln(L(m1) / L(m2))=2(loglik(m2)−loglik(m1))

# LAD (Least Absolute Deviation):

The Least Absolute Deviation model minimizes the absolute value of the residuals (<http://www.real-statistics.com/multiple-regression/lad-regression/>)

MAE regression

# Robust Regression:

Robust regression is an alternative to least squares regression when data are contaminated with outliers or influential observations, and it can also be used for the purpose of detecting influential observations. Ordinary least squares have favorable properties if their underlying assumptions are true but can give misleading results if those assumptions are not true; thus, ordinary least squares are said to be not robust to violations of its assumptions. Robust regression methods are designed to be not overly affected by violations of assumptions by the underlying data-generating process.

# Least Square Regression:

If your data shows a linear relationship between the X and Y variables, you will want to find the line that best fits this linear relationship. That line is called a Regression Line and has the equation ŷ= a + b x. The Least Squares Regression Line is the line that makes the vertical distance from the data points to the regression line as small as possible. It’s called a “least squares” because the best line of fit is one that minimizes the variance (the sum of squares of the errors). This can be a bit hard to visualize but the main point is you are aiming to find the equation that fits the points as closely as possible.

1. to see how the variables, enter the model, we check adj R squared and see which one has the highest value and then move on to the only models that contain only those who had the highest value. See lec 16 for more info on that
2. The variable that should be removed is the one with highest p value
3. You might get different results when u do different regression analysis (forward, backward, stepwise, all, AIC etc.) and in order to select the best among them, you use cross validation
4. By adding a predictor if R squared increases it is due to overfitting but there is no guarantee on that
5. Maximize R squared, R squared adj, minimize RMSE and MAE in k fold cross validation
6. LOESS, SMA 🡪 need to study