The Definitive Guide to ECON 257D2

Understanding basic economic statistics (or atleast the second semester)

Chapter 1: Distributing Distributions

The first part of the course mostly dealt with reviewing distributions we had seen last semester. Some of the distributions included here were presented later in the semester

Distribution: A link between the potential outcomes of a random phenomenon and the probabilities of them arising.

- **Discrete distribution:** A distribution with a finite amount of potential outcomes.
- **Continuous distribution:** A distribution with an infinite amount of potential outcomes.

Description of distributions:

- **Probability density function (or probability function for discrete case) (PDF):** Function that maps potential outcomes of the phenomena to the probability that they arise.
- **Cumulative distribution function:** Function that maps an outcome to the probability of getting an outcome or an outcome smaller that outcome.
- **Expected value:** Commonly known as the mean, gives the probability-weighted average value returned by the distribution (if you were to sum up each potential outcome, multiplied by its probabililty, what would you get?)
    - Measure of location (meaning that it tells you roughly where the distribution is situated)
- **Variance:** The expected value of the squared deviation of a random variable from its mean (on average, how far away is a random variable from its mean)
    - Measure of dispersion (describes how the data is spread out

Generating probability functions: Anyone can create distribution functions as long as they match the following criteria. Existing functions that match this criteria are often used as distribution functions.

- The range of the function must be positive (can include 0)
    - Since there cannot be any negative probabilities, the probability that each outcome is mapped to must be positive (the domain however can be the real numbers)
- The sum of all probabilities of all outcomes must be exactly 1
- The outcomes must be mutually exclusive

Distribution parameters: Luckily, distributions aren't one size fits all; you can use a parameter to modify the probability density function of that distribution. In doing so, the value of the parameter will usually affect the expected value and the variance

Important distributions

- **Discrete distributions**
  - **Poisson distribution:** Distribution that takes one parameter, lambda. Was created using the power series for e^x
    - **PDF:** $P(x = k) = \frac{\lambda^k e^{-k}}{\boldsymbol{k}!}$
    - **CDF:** $P(x \leq k) = \sum_0^k \frac{\lambda^k e^{-k}}{\boldsymbol{k}!}$
      - To find the probability that a r.v. following the poisson distribution has a value above k, you can use 1 – probability that it has a value less than k
    - $E(x) = \lambda$
    - $V(x) = \lambda$
    - Converges to normal by CLT since the poisson distribution can be thought of as the sum of lambda r.vs. distributed as Poisson(1)
  - **Binomial distribution:** Distribution that models a random even with outcome 1 (probability p) and outcome 0 (with probability 1-p), the random event being observed n times. The two parameters are **n (number of trials)** and **p (the probability of success)**
    - **PDF:** $P(x = k) = \binom{n}{x} p^x (1-p)^{n-x}$
      - $\binom{n}{x} = \frac{n!}{x!(n-x)!}$
    -
    - **CDF:** $P(x \leq k) = \sum_0^k \binom{n}{k} p^k (1-p)^{n-k}$
    - $E(x) = np$
    - $V(x) = np(1-p)$
    - Like the poisson, the binomial converges to the normal (fairly quickly, can use normal approximation if np>5 and n(1-p) > 5). Due to it being the sum of n bernouilli random variables.
  - **Multinomial distribution:** Similar to the binomial but instead of there being just 2 possible outcomes, there is a finite set of possible outcomes. To each possible outcome, there is a probability of it arising and we observe N events.
- **Continuous distributions**
  - **Normal distribution:** Arguably the most important statistical distribution, it comes from the integral $\int_{-\infty}^{\infty} e^{\frac{-x^2}{2}} dx = \sqrt{2\pi}$. It takes two parameters, $\boldsymbol{\mu}$ being the mean (simply a value that is added to the standard normal distribution centered at 0) and $\boldsymbol{\sigma}$ being a value that multiplies the standard normal distribution (typically multiplied by 1)
    - **PDF:** $P(x = k) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$
    - **CDF:** $P(x \leq k) = \int_{-\infty}^{k} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx$
    - $E(x) = \mu$
    - $V(x) = \sigma^2$

        - The normal distribution is symmetric around its mean, with half the probability mass being on each side. Thus, $P(x \leq z) = 1 - P(x \leq$-z)
    - **T distribution:** Arises from a normal divided by the root of a Chi-square (if Z ~ N(0, 1) and $W \sim X_k^2$, then, $\boldsymbol{X = \frac{Z}{\sqrt{\frac{W}{k}}} \sim t(k)}$**.** The only parameter it takes is the number of degrees of freedom (otherwise it resembles the standard normal.)
        - **PDF and CDF and variance:** Just look at the tables
        - $E(x) = 0$
        - Converges asymptotically (as the degrees of freedom increase) to a standard normal
        - Usually used when testing the mean of a normally distributed variable for which the variance is unknown (since the sample variance has a chi-squared distribution)
        - Gives a more conservative estimate than just a normal test
    - **Chi-Squared distribution:** Arises from the sum of k squared standard normal variables with the only parameter it takes being k = number of summed squared standard normal variables. K is also equal to the number of degrees of freedom
        - **PDF and CDF:** Look at the table
        - $E(x) = k$
        - $V(x) = 2k$
        - The chi-squared distribution is **not symmetric**
    - **F distribution:** Arises from the division of one chi-squared random variable by another (which means that squaring a student's t distributed r.v. gives a F distribution as well). Takes 2 parameters (first being the degrees of freedom of the chi-squared in the numerator and the second being the degrees of freedom of the chi-squared in the denominator)
        - **PDF and CDF:** Look up in table
        - The F distribution is **not symmetric**

Chapter 2: Experts Estimate a…

In this relatively short chapter, we looked mostly at what estimators are, their properties and some commonly used ones.

Estimator: An estimator is a random variable that, as the name implies, estimates the value of a population parameter (usually through the use of sample data). This random variable has its own distribution with its own expected value/variance.

Estimator properties: The following properties are useful for evaluating/comparing estimators:

- **Bias:** The bias of an estimator relative to the population parameter it estimates is equal to the difference between expected value and the population parameter ($E(\bar{x}) - \mu$)

- If the estimator is **unbiased**, we have that $E(\bar{x}) = \mu$
- **Consistency:** An estimator is considered consistent if, as N approaches infinity, the estimator converges towards the true population parameter.
    - For this to be true, we need the estimator to be unbiased and for its variance to converge to 0 as N approaches infinity
- **Efficiency:** The variance of the estimator (smaller is obviously more desirable)
    - The BLUE (Best Linbear Unbiased Estimator) is the one with the highest efficiency (lowest variance)
    - **Relative efficiency:** The efficiency of an estimator relative to another is the ratio of their variances

Common estimators and their distributions: The following estimators are commonly used in the tests in the upcoming section. They are provided with their distribution

- **Estimators when $X \sim N(\mu, \sigma^2)$**
    - **Sample mean:** $\bar{x} = \frac{1}{N}\sum x_i \sim N\left(\mu, \frac{\sigma^2}{N}\right)$
    - **Sample variance:** $s^2 = \frac{1}{N-1}\sum(x_i - \bar{x})^2 \sim X^2_{N-1}$
- **Estimators when X ~ Bin(n, p)**
    - **Estimated parameter:** $\hat{p} = \frac{successes}{trials} \sim N(p, \frac{p(1-p)}{N})$

Chapter 3: Putting Things to the Test

And here lies the biggest part of the course, learning about various ways of testing hypotheses for different distributions. To do so we, need the essential elements of a statistical test.

Elements of a statistical test:

- **Hypotheses:** 2 mutually exclusive possibilities
    - **Null hypothesis:** The hypothesis we are trying to reject (formulated as population parameter = a certain value)
    - **Alternative hypothesesis:** The hypothesis we are comparing it to (formulated as the population parameter is bigger or smaller than a certain value, or simply different)
- **Test statistic:** It is a random variable with its own distribution that satisfies 2 criteria
    - Needs to be a function of the sample measurement and the population parameter we are looking for (where the population parameter is the only unknown).
    - Its probability distribution should not be a function of the population parameter.
- **Rejection Rule:** A rule basically saying if the test statistic is above or below a certain value, we reject the null.

When we perform these tests, there is always a certain chance that our ultimate verdict is wrong (probability that we make a certain error)

Types of errors

- **Type I error (α):** Probability of rejecting the null when it is true
  - Usually only care about a type I error since it is considered more harmul to make one
  - Alpha is also called the **level of the test**
- **Type II error (β):** Probability of accepting the null when it is false
  - $\mathbf{1-\beta}$ is the **power of the test**
- **Error tradeoffs:** For a fixed N (and as such, a fixed amount of information), reducing the probability of making one type of error increases the probability of making the other type of error.
  - Alternatively, reducing the power of the test also reduces the level of the test

Additionally, when performing our tests, we need to consider the sidedness of the test which affects the way we formulate the alternative hypothesis and ultimately the decision rule we use.

Sides of test: Since we usually test the hypothesis at a certain probability, the side of the test will affect which boundary we use for the decision rule.

- **One-sided test:** The alternative hypothesis is that the population parameter is below or above a certain value.
  - Since we are testing against the parameter being above a certain value, we only reject the null when the test statistic is above a value that makes it such that the probability of the null being true is below a certain percentage.
- **Two-sided test:** The alternative hypothesis is that the population parameter is different than a certain value.
  - Since we are testing against the parameter being different than a certain value, we only reject the null when the test statistic is above or below a value that makes it such that the probability of it being above or below that value is a certain percentage (thus we split the probability we are checking for between the areas above and below)

Instead of testing at arbitrary levels, we can also get the probability of getting a value as extreme as the one taken on by the test statistic which brings up the notion of p-value.

P-value: The smallest α for which we would reject the null hypothesis

- A small p-value indicates more compelling evidence against the null
- The p-value can be used to test the hypothesis (if the p-value is below the alpha we want to test for, we can reject)
- For a two-sided test, the p-value is the probability of getting a value of the test statistic that's that extreme, divided by two

Finally, we can use confidence intervals to determine a likely interval in which the population interval lies

Confidence interval: Another random variable that consists of an upper and lower bound for the parameter. A $100(1-\alpha)\%$ confidence interval is an interval such that the probability that the parameter lies in it is $1-\alpha$

- If the null hypothesis isn't in the $1-\alpha$ confidence interval, then we would reject the null for a two-sided test at level α
  - For a one-sided test, we just get an interval with one bound

Great! We've now got all the tools to perform hypothesis tests. Now for a long list of tests that vary based on what we are testing.

Test of a mean of a normally distributed variable (unknown variance, $X \sim N(\mu, \sigma^2)$)

- **Hypothesis:** $\mu = c$
- **Test statistic:** $T = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{N}}} \sim t(N-1)$
- If the variance was known, we would simply use it in the calculation of the test statistic and use the normal distribution

Test of population proportion (X ~ Bin(N, P))

- **Hypothesis:** $p = p_0$
- **Test statistic:** The test statistic we use depends on the value of NP and N(1-P). If both are above 5 we can use the normal approximation, or else we use the binomial.
  - **Binomial:** $N\hat{p} \sim Bin(N, p)$
  - **Normal:** $T = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{N}}} \sim N(0, 1)$

Test of difference in means, matched pairs, dependent samples ($X \sim N(\mu_x, \sigma_x^2), Y \sim N(\mu_y, \sigma_y^2)$)

- In this test, we have a sample of matched pairs with both elements coming from separate, normal distributions. We want to test if their mean is the same or different. We can construct a new random variable which consists of the difference between one element of the matched pair and the other.
  - Let $D = X - Y$
    - To compute the sample data for D, calculate the difference for each matched pair. You can then consider the resulting sample of data as D and calculate its sample average and variance as you normally would
- **Hypothesis:** $\mu_D = D_0$
- **Test statistic:** $T = \frac{\bar{D}-D_0}{\sqrt{\frac{s_D}{N}}} \sim t(N-1)$

Test of difference in means, independent samples ($X \sim N(\mu_x, \sigma_x^2), Y \sim N(\mu_y, \sigma_y^2)$)

- Unlike last time, we have two samples of data but they are not linked in matched pairs. The number of sample points for each sample can differ.
- **Hypothesis:** $\mu_D = D_0$
- **Test statistic:** The test statistic we use depends on what information we know about the distributions. 3 cases are possible

- **Both variances are known**
  - $T = \frac{\bar{D}-D_0}{\sqrt{\frac{\sigma_x^2}{N_x}+\frac{\sigma_y^2}{N_y}}} \sim N(0,1)$
- **Variances are the same but unknown**
  - $T = \frac{\bar{D}-D_0}{\sqrt{\frac{\sigma^2(N_x+N_y)}{N_xN_y}}} \sim t(N_x + N_y - 2)$
  - Where $\sigma^2 = \frac{1}{N_x+N_y-2}\left(\sum(x_i - \bar{x})^2 + \sum(y_i - \bar{y})^2\right)$
- **Variances are different and unknown**
  - $T = \frac{\bar{D} - d}{\sqrt{\frac{s_x^2}{N_x}+\frac{s_y^2}{N_y}}} \sim t\ (incredibly\ long\ degrees\ of\ freedom\ formula)$

Test of difference in proportion for large samples

- Here we are testing whether two binomially distributed sets of data share the same probability of success
- **Hypothesis:** $p_1 - p_2 = c$
- **Test statistic:** The test statistic we will use depends on our null hypothesis (if we are testing if they are the same, then they will have the same variance, otherwise they will not)
  - $\boldsymbol{H_0: p_1 - p_2 = 0}$
    - $T = \frac{\widehat{p_1}-\widehat{p_2}}{\sqrt{\frac{\tilde{p}(1-\tilde{p})(N_1+N_2)}{N_1N_2}}} \sim N(0,1)$
    - $\tilde{p} = \frac{N_1\widehat{p_1}+N_2\widehat{p_2}}{N_1+N_2}$
  - $\boldsymbol{H_0: p_1 - p_2 = c}$
    - $T = \frac{\widehat{p_1} - \widehat{p_2} - c}{\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{N_1}+\frac{\widehat{p_2}(1-\widehat{p_2})}{N_2}}} \sim N(0,\ 1)$

Test of variance of normally distributed variable ($\boldsymbol{X \sim N(\mu, \sigma^2)}$)

- Here the goal is to test whether the variance of a parameter has a certain value, even though we don't know the population mean.
- **Hypothesis:** $\sigma^2 = a$
- **Test statistic:** $\frac{s^2\ (N-1)}{\sigma^2} \sim \mathrm{X}^2_{(N-1)}$

Test for equality of variance (works for independent samples) ($\boldsymbol{X \sim N(\mu_x, \sigma_x^2), Y \sim N(\mu_y, \sigma_y^2)}$)

- In this test, we have 2 independent samples and we are checking if their variances are equal
- **Hypothesis:** $\frac{\sigma_x^2}{\sigma_y^2} = 1$
- **Test statistic:** $\frac{s_x^2}{s_y^2} \sim F(N_x - 1,\ N_y - 1)$
  - Make sure to pick the largest sample variance as numerator

Chapter 4: My Goodness of Fit!

In this chapter, we looked at tests to see if a given sample fit distributions that we know the PDF of. In general these tests consist of knowing certain moments of an existing distribution and checking if the corresponding moment of the sample data matches that moment of the existing distribution.

Test for normality

- We know that the skewness of a standard normal is 0 and the kurtosis is 3
- **Test of skewness**
  - **Hypothesis:** $skewness = 0$ (against alternative it is equal to 0)
  - **Test statistic:** $\frac{\widehat{sk}}{\sqrt{\frac{6}{N}}} \sim N(0,\ 1)$
    - Where $\widehat{sk} = \frac{\frac{1}{N-1}\sum(x_i - \bar{x})^3}{s^3}$
- **Test of kurtosis**
  - **Hypothesis:** $kurtosis = 3$ (against alternative it is not equal to 3)
  - **Test statistic:** $\frac{\hat{k} - 3}{\sqrt{\frac{24}{N}}} \sim N(0,\ 1)$
- **Jarque-Bera test:** Test whether skewness is equal to 0 and if kurtosis is equal to 3
  - **Hypothesis:** Data is normally distributed (against alternative it is not)
  - **Test statistic:** $JB = \frac{N(\widehat{sk})^2}{6} + \frac{N(\hat{k})^2}{24} \sim X_2^2$
    - We reject the test statistic exceeds the value of X_2^2 at level alpha (one-sided test)

Test for multinomial: Instead of testing whether the data is normally distributed, we test whether the data fits a multinomial where the probability of each outcome being picked is equal

- **Hypothesis:** $p_1 = p_2 = \cdots = p_c = \frac{1}{c}$ against the alternative that this is not true for any pair
- Using the observed counts for each outcome, we compare them to the expected counts for each outcome, giving us a normally distributed test statistic for each outcome.
  - $\frac{N_i - e_i}{\sqrt{e_i}} \sim N(0,1)$
- We then sum the square of these test statistics to get a chi-squared test
  - $X^2 = \sum \frac{(N_i - e_i)^2}{e_i} \sim X_{c-1}^2$ $(where\ N_i\ is\ the\ observed\ count\ and\ e_i\ is \frac{N}{c})$
- Since the test statistic is distributed as a chi-square,we reject if it exceeds the value at level alpha

Test for Poisson: In this, test we get count data (ex: we observed 1 person passing by twice, 2 people passing by 5 times , ...). We want to test if this count data matches a Poisson distribution

- **Hypothesis:** The data is distributed as poisson and thus the probability of getting a certain count is determined by the probability function
- **Test statistic:** Before calculating the test statistic, we must first estimate λ using the sample mean of the counts
    - Using that estimated λ, we can calculate the expected counts for each value (N times the probability function at that value)
    - $X^2 = \frac{(N_0 - e_0)^\wedge 2}{e_0} + \cdots + \frac{(N_c - e_c)^2}{e_c} \sim X^2_{c-2}$(c is the number of categories)
- **Outliers:** If we observe outliers in the count data, we can group them into one category of say observations bigger than 15)
    - When using the outliers to estimate lambda, multiply the number of observations bigger than 15 by 15 instead of their original value
    - The expected count for that category is the sum of the expected counts of the grouped outliers.

- Helpful to construct a contingency table of the type

k	observations	p	expected
0	10	0.66	10
1	4	0.2	3
2	1	0.066	1

Test for independence: Unlike the other tests, we are not comparing the sample data to a distribution. Instead, we have 2 different characteristics (ex: university major and their GPA category) or not as well as counts for intersections of those characteristics (ex: there are 50 Arts with a GPA above 3, 70 Science students, ...).

We then want to compare if the the two characteristics are independent, in which case the probability of being an arts student and having a GPA above 3 would be equal to the probability of being an arts student * the probability of having a GPA above 3.

- **Hypothesis:** $p_{xy} = p_x * p_y \ for \ all \ x, y$
- **Test Statistic:** $X^2 = \sum \frac{(N_{ij} - \widehat{e_{ij}})^2}{\widehat{e_{ij}}} \sim X^2_{(r-1)(c-1)}$
    - $N_{ij} = observed\ count,\ \widehat{e_{ij}} = N\widehat{p_i}\ \widehat{p_j}$

Chapter 5: Dropping Off Mr. Parametric

Most of the tests we've done so far have depended on knowing something about the distribution of the underlying data. Nonetheless, it is possible to perform tests without assuming/knowing the distribution of the data. These tests are considered **non-parametric**. In performing these tests, we are usually evaluting the median of the data rather than the mean.

Sign test for matched pairs: We have matched pairs of sample data and we want to see if one the values is generally higher or lower than the others. Instead of using the raw data, we can compute a new variable S which is equal to 1 if the difference between the values of the matched pair is positive and equal to 0 otherwise (if they are the same, discard)

- **Hypothesis:** $p_+ = 0.5$
- Since under the null $S \sim Bin(N, 0.5)$ we can perform a standard test of proportion

Wilcoxon signed-rank test for matched pairs: Similar to the previous test but we want to take into account the magnitude of the differences.

- **Hypothesis:** $\mu_+ = \mu_-$ (the average rank of the positive pairs is equal to the average rank of the negative pairs
- **Test statistic:** Either test statistic can be used
  - $T = \frac{w_+ - n\mu}{\sigma} \sim N(0,\ 1)\ or\ T = \frac{w_- - n\mu}{\sigma} \sim N(0,\ 1)$
  - $n\mu = \frac{n(n+1)}{4}$
  - $\sigma^2 = \frac{n(n+1)(2n+1)}{24}$

Rank test for independent samples: The goal of this test is to evaluate whether two independent samples of data have the same distribution (in which case they should have the same median).

- **Hypothesis:** X and Y have the same distribution
- **Test statistic:** We construct a test statistic based on the sum of the ranks which should be normally distributed.
  - $T = \frac{W_x - N_x\mu}{\sigma} \sim N(0,\ 1)$
  - $E(W_x) = N_x\mu = \frac{N_x(N_x+N_y+1)}{2}$
  - $V(W_x) = \frac{N_xN_y(N_x+N_y+1)}{12}$

Chapter 6: Going Back to Regression

The final chapter of the course was on the basics of linear regression (how to construct regression models, the math behind least squares regression, important theorems and how to perform tests on the regression coefficients).

Regression: The general set of techniques for determining relationships between variables.

- **Regression model:** Model that seeks to determine the expected value of a random variable based on a set of conditioning characteristics
- The simplest way to do this is to assume there is some linear relationship between the dependent and independent variables and try to find the coefficients for that relationship

Data: The data we use for a regression is usually structured in the following form.

- $y = \begin{bmatrix} y_1 \\ \cdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & \cdots & x_{k1} \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_{kn} \end{bmatrix}$
    - First column of X is filed with 1s with the rest of the columns being the value of independent variables for each sample point
    - Y is a column vector with the value of the independent variable for each sample point

Data generation process: The true process that generates the data that we are trying to model

Basic linear regression model: $Y = \beta X + \epsilon$ (the dependent variable is equal to a linear coefficient multiplied by each independent variable + an error term for each sample point

- In matrix form $\begin{bmatrix} y_1 \\ \cdots \\ y_n \end{bmatrix} = \begin{bmatrix} \beta_1 & \cdots & \beta_k \end{bmatrix} \begin{bmatrix} 1 & \cdots & x_{k1} \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_{kn} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \cdots \\ \epsilon_n \end{bmatrix}$

Computing $\boldsymbol{\beta}$: To compute the regression coefficients, we want to find the coefficients that minimize the sum of squared errors which can be done by taking the derivative set it equal to 0

- In doing so we get $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$
    - Betahat is an unbiased consistent estimator and we can estimate its variance using $\widehat{\sigma^2} = \frac{\hat{\epsilon}'\epsilon}{n-k}$

Variances: It is important to know the covariances matrices for $\hat{\beta}$ and ϵ to perform tests on the regression coefficients. These are derived by simplifying the expressions for the variance.

- **Covariance of $\widehat{\boldsymbol{\beta}}$:** $cov(\hat{\beta}) = \sigma^2(X'X)^{-1}$
- **Covariance of ϵ**: $cov(\epsilon) = \sigma^2 I$

Residuals: The discrepancies from the model for each sample point ($\hat{\epsilon} = y - X\hat{\beta}$)

- **SSR (Sum of squared residuals):** $\sum\widehat{\epsilon_i^2}$
- **SST:** $\sum(\boldsymbol{y}_i - \overline{\boldsymbol{y}})^2$
- $\boldsymbol{R}^2 = \boldsymbol{1} - \frac{\boldsymbol{SSR}}{\boldsymbol{SST}}$
    - Also known as the goodness of fit of the model, it measures how much of the independent variable can be predicted from the dependent variables

Performing tests: There is a general procedure for testing the value of the regression coefficients

- **Hypothesis:** $a'\hat{\beta} = 0$
  - $a = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$ $(some\ vector\ of\ weights\ applied\ to\ each\ coefficient)$
- **Test statistic:** $\frac{a'\hat{\beta}}{SE(a'\hat{\beta})} \sim N(0,1), or\ t(n-k)$
  - $var(a'\hat{\beta}) = a'var(\hat{\beta})a$

Other considerations:

- **Dealing with qualitative variables:** Create a dummy variable that takes on the value of 1 if the sample has that characteristic or 0 otherwise
- **Non-invertibility:** If X is not of full rank (there are linearly dependent columns or rows), then X'X is not invertible either
- **Taylor's theorem:** We can approximate a non-linear functions using a linear function (however, the quality of the approximation worses as we get farther away from the point of approximation)
- **Correlation and omission issues:** If you omit a variable that is correlated to another that you include and they both have an effect on the independent variable, then the regression will overestimate the impact of the variable that was included
  - Not an issue for a predictive model but can cause problems if you're examining causation

Gauss Markhov Theorem: States that with certain assumptions, the OLS estimator is BLUE. The assumptions required for this theorem are:

- $y = X\beta + \epsilon$
  - There is a linear relationship between y and X
- X is an n x k matrix of full rank
  - The columns of X are linearly independent
- $E(\epsilon|X) = 0$ or $E(X'\epsilon) = 0$
  - On average, the residuals have value 0 for any value of X
- $E(\epsilon\epsilon'|X) = \sigma^2 I$
  - Essentially states 2 things:
    - **Homoskedasticity**: The variance of all epsilons is the same
    - The errors terms are independent (covariance is 0)
- X is generated by a mechanism unrelated to ϵ (can be fixed or random)