Chapter 1

* Statistics: Science of collecting, classifying, analyzing describing and presenting data alongside scientific evidence being found from thing studied
* 3 main aspects
  + Research Design: Planning, designing appropriate ways of collecting data for the investigation of a particular scientific problem
  + Descriptive Stats: Description, summarization and presentation of data using both numerical methods and graphs
  + Inferential Stats: Drawing scientific conclusions and making predictions about a population based on data gotten from a sample of population. Usually has
    - Hypothesis tests
    - Confidence intervals
    - Estimate about population based of sample
* Population : Entire of all individuals/items under consideration in study
  + Population size is number of individuals or items in population under study, denoted by N.
* Census is collecting data about entire population
* Sample subset of population in which info is obtained
  + Small num of observations from population is investigated
  + Impossible/too expensive to measure variable of entire population
  + Sample size (denoted n) is the no. of observations in single sample
* Inferential Stats: Using info to make decisions, conclusions, and predictions about population
* Parameter: Descriptive measure of a population( µ🡪population mean, σ 🡪 population standard deviation)
* Statistic: A descriptive measure of a sample which is usually used to estimate a parameter. ( y with a bar on top is standard mean and S is sample standard deviation)
* Components of Research Design
  + Study Units
    - Individuals/subjects which info is required or on which measurements are recorded
    - Usually called cases/units of analysis
    - In experimental studies, they are called experimental units and in observational studies, they are called units of observations
    - They are the who.
  + Variables
    - Characteristic which varies from one study unit to another
    - Usually the what
    - Characteristic being measured on the study unit depends on research
    - Distribution of variable is all the values a variable takes on
    - Data is values of variable and datum is a single measurement
    - Types of variables
      * Recording
        + Categorical variables(Qualitative)

Recorded on nominal scale

Non-numerically valued

Classified by quantity or attribute

* + - * + Ordinal scale variables can be ranked but don’t have an in-between between the rankings
        + Quantitative variables

Numerically values

Constant interval size

Discrete: Only takes on specific values

Continuous: Infinite number of values between range

Ratio scale: true zero point

Interval Scale: no true zero point

* + - * + Indicator/dummy variables:

Categorical variables coded to get quantitative variables

* + - * Variables in Research:
        + Explanatory:

Variables of interest

Used to explain/affect other variables but aren’t affected by other variables

* + - * + Response:

Variable affected by explanatory

* + - * + Extraneous:

Explanatory variables that aren’t of interest or related to study

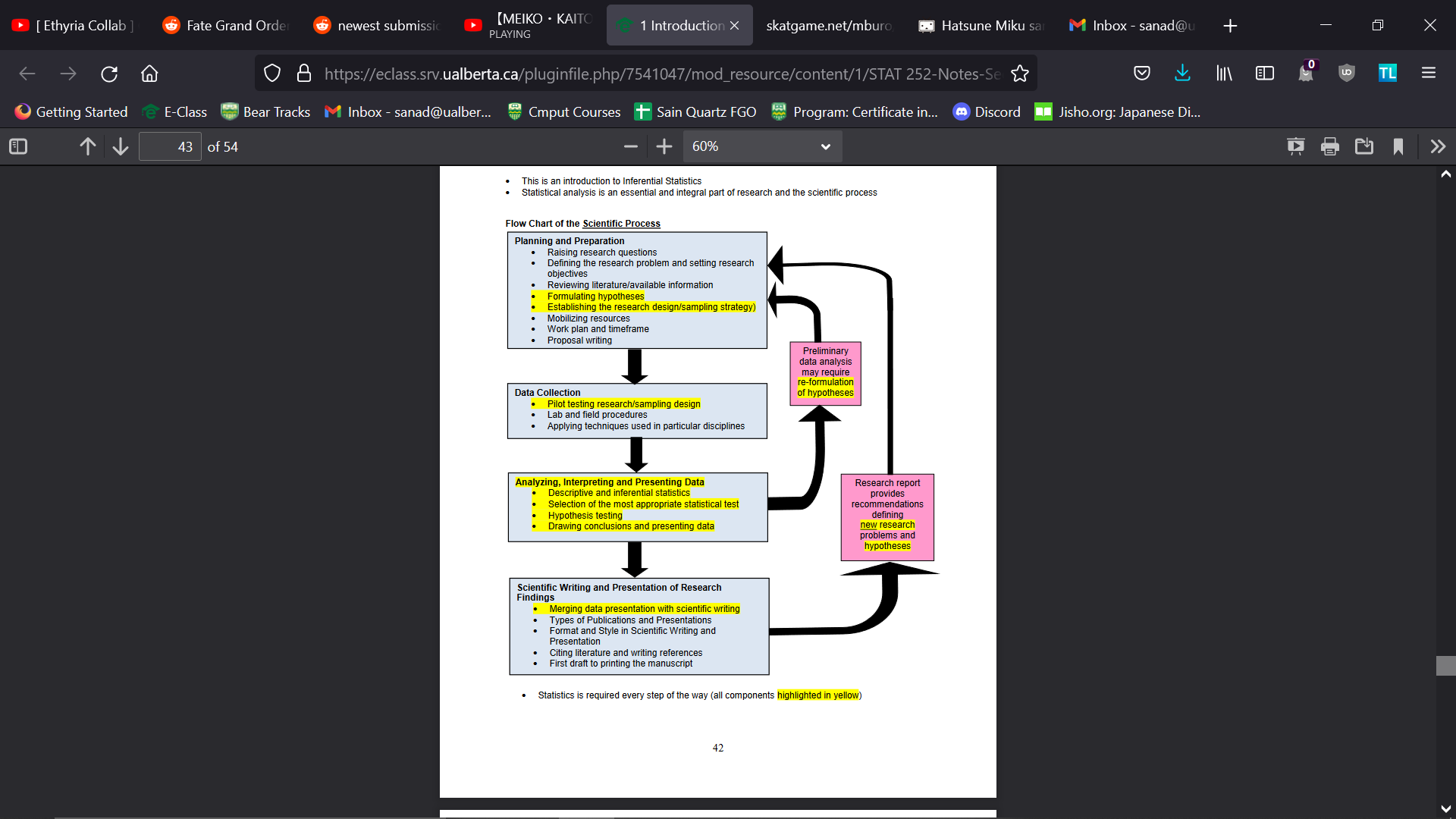
Could affect response variables

Unable to be measured/can’t be measured

* + - * + Factor:

Explanatory variables applied as treatments in experiment/levels in observational studies

* + Spatial Aspects:
    - Where
    - Way observations or replicates arranged in space
    - Linked to study unit(where they’re sampled and measured
  + Temporal Aspects:
    - When
    - Time period
    - Star, end, freq. of recordings
  + Techniques/Method:
    - How
    - Methods used to get methods & record data
  + Repeatability and Forward Planning for Data analysis
* Random Sampling: Selection of units/individuals from population without bias
  + Individuals from population equal chance of getting chosen and is independent
  + Ensures sample is as representative as possible of entire population
  + Most tests assume sampling is random
  + Sampling strategy and research design specify the way observations are recorded in space and time
  + Uses probability sampling
* Computer programs don’t generate truly random numbers (use of a seed helps though)
* Sampling without replacement means one can be selected only once whereas with replacement means they can be selected multiple times
* Types of random sampling
  + Simple: completely random and independently
  + Systematic: First is selected randomly then next samples are selected sequentially
  + Stratified: Population divided into strata based on pilot study or prior info
    - Each sub-population is relatively homogenous
    - Proportional allocation: sampling intensity in each stratum is proportional to estimated density of items
    - Most accurate if there are definite strata in the study area
  + Multistage random sampling
  + Cluster sampling
* Always make sure your sample size is large enough
* Sample size/N is sample fraction
* Different types of sample bias
  + Voluntary Response
  + Nonresponse bias
  + Incomplete sample frame
  + Undercoverage
* Observational Study:
  + Tries to estimate population parameters
  + Data about particular phenomenon as it occurs without meddling
  + Randomness of units from population
  + Can’t make causal inferences(cause and effect between response and explanatory variables)
  + Can make population inferences( can safely say that the sample data represents population) if random selection from population
  + They are prospective or retrospective. Prospective means subjects are identified beforehand and data are recorded as the study proceeds. Retrospective subjects identified and data recorded after events occur
* Experimental Studies:
  + Randomness in this case deals with random selection of sample AND random assignment to treatment groups
  + Usually have a control group with no assigned treatments
  + Can make both population and causal inferences as long as random selection AND random assignment
  + We could blind anything that could affect the results or who evaluate them or both.
* Frequency: Count of observations which fall into given observation
* Describing quantitative data has 3 aspects
  + Shape
    - Larger the size, better approximation of population distribution
    - Skewness vs Symmetrical
      * Left(-ve) = Longer left tail whereas positive skewness(right)=longer right tail
      * Symmetrical, both tails are equal
    - Uni, Bi, Multimodal
  + Center
    - Mean(not resistant to skewness), Median(resistant to skewness), Mode
  + Spread
    - Range
    - Variance and Standard Deviation(s=sqrt(v)) v=variance s = standard deviation
  + Degrees of freedom: number of independent observations. For s.d. it is n-1
* IQR=Q3-Q1
* Upper limit is Q3+1.5\*IQR whereas lower limit is Q1-1.5\*IQR anything ove or below the limits are outliers
* In a normal distribution curve, area under graph represents relative frequency/probability/% of observations
  + Bell shaped
  + Smooth curve
  + Centered at mean and its defined by mean and standard deviation
  + Any normally distributed variable is distributed according to these properties:   
    • 68.26% of all observations lie within one standard deviation on either side of the mean.   
    • 95.44% of all observations lie within two standard deviations on either side of the mean   
    • 99.74% of all observations lie within three standard deviations on either side of the mean
  + Z score is (y- mean)/standard deviation
* For samples of size n, mean of samples equals mean of population whereas standard deviation of sample is standard deviation of variable divided by the squared root of the sample size(also referred to as sample error).
* If x of a population of size n is normally distributed. Its shape is normally distributed, mean of sample distribution is equal to mean of x and s.d of distribution is the s.d of x divided by sqrt(n).
* Z=y-µ/(ơ/sqrt(n))
* Central Limit theorem.: Regardless of the distribution of the variable under study, for a relatively large sample size, the   
  variable y is approximately normally distributed. The approximation becomes better with increasing sample size(size>=30)
* If a variable isn’t normally distributed:
  + Shape: Regardless of the distribution of x, if the sample size is large (n ≥ 30), the   
    sampling distribution of all possible sample means (i.e., the distribution of the variable x )   
    is approximately normally distributed. Sampling distribution of all possible sample means   
    (known as variable x ) is also normally distributed.
  + Center: The mean of the sampling distribution is: μx= μ
  + Spread: The standard deviation of the sampling distribution is: σx =n  
    σ/sqrt(n).
* For inferential statistics, general flowchart below



* When performing research, we generally have a research question and objective
  + (Statistical Hypothesis)Null and Alternate Hypothesis are used. Null is usually no difference/relationship whereas alternate hypothesis means there is a difference or a relationship exists
  + Predicted outcome is our research hypothesis
* Test statistic is a statistic usually used as a basis in decided whether null is rejected or failed to be rejected
  + Calculated value (observed value) of the test statistic calculated from the data collected
  + Critical value of the test statistic
    - obtained from a table showing its theoretical distribution, which is compared with the calculated value in order to make a decision about the hypotheses
    - Forms the border separating the rejection and nonrejection regions (the critical value   
      itself is part of the rejection region)
  + Rejection region = the set of values for the test statistic that leads to rejection of the null hypothesis
  + Nonrejection region = the set of values of the test statistic that leads to nonrejection of the null hypothesis
* Type 1 error is when we reject the null hypothesis when we shouldn’t and type 2 error is when fail to reject the null hypothesis when we should
* Steps to follow when testing a hypothesis
  + Choose test
  + State H0 and Ha
  + Get test statistic
  + Reject(P<=α)/Fail to reject H0(P>=α)
  + Conclude
* P value is observed probability of type 1 error you find based on data obtained from test statistic and examination of appropriate statistics stable
* Confidence Interval: An interval or range of numbers derived from point estimate of parameter
  + For a certain percentage of all samples of size n, the population mean µ lies within the confidence interval of the sample mean x
  + Estimate+- Critical Value \* Standard Error of Estimate where critical value is the value obtained from a table showing the theoretical distribution of test statistic at given confidence level.

Chapter 2

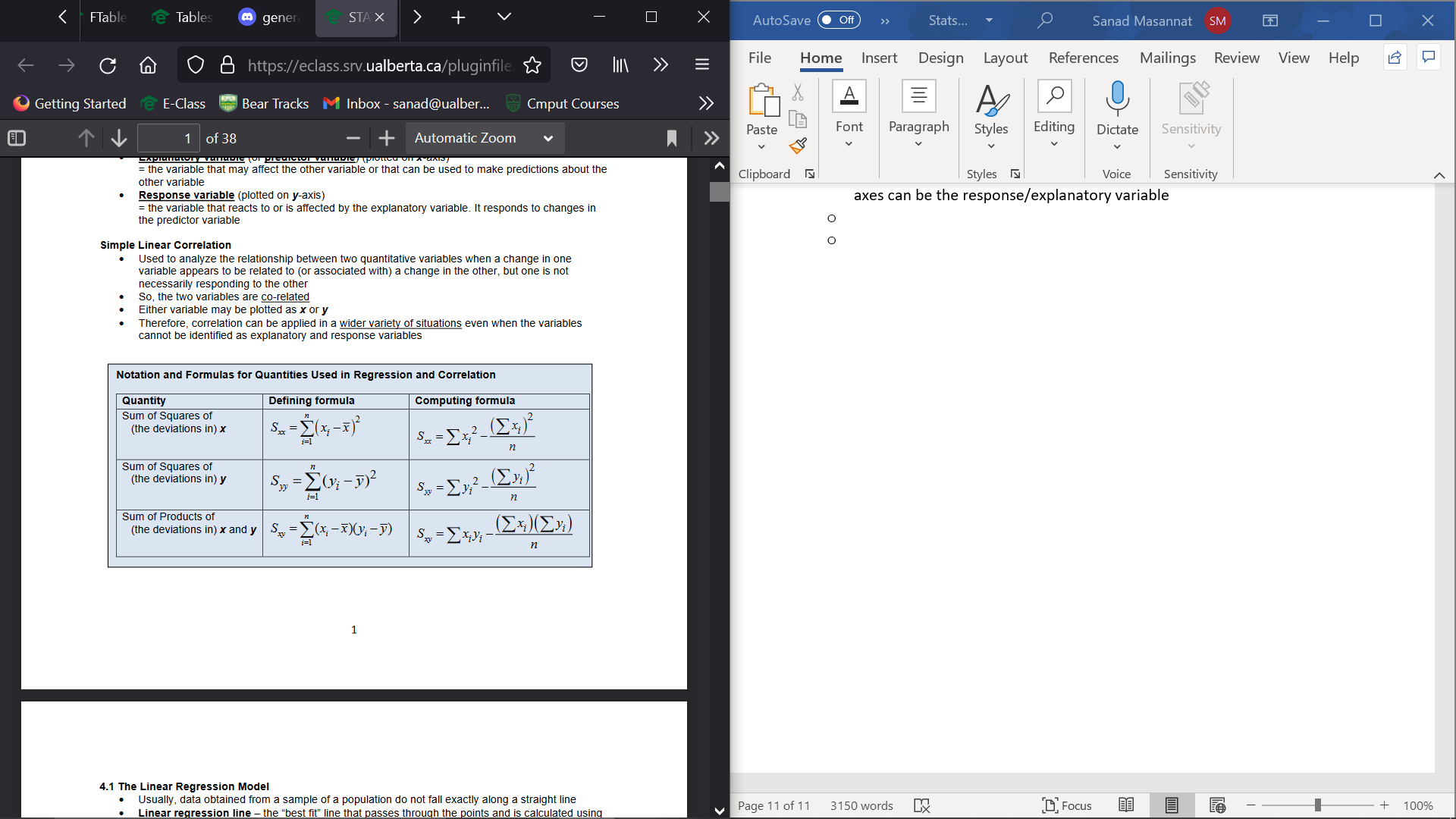
* Suppose a variable y of a population is normally distributed with mean y, then for samples of size n: t=y- µ/(s/sqrt(n)) where the variable t has n-1 df. (One sample t-test)
* T-distribution is more spread out than normal distribution
* Properties of t-curve
  + Total area under curve=1
  + Symmetrical about 0
  + Extends infinitely in both directions but nut never touch the axes
  + different t-curve identified by df
  + As df increases, it becomes closer to normal curve until df=infinity in which the two are equal
* T-table is 1-tailed/2-tailed or both
* One Mean/One Sample t-test
  + Assume σ is unknown, simple random sample and sample size is large
  + Null hypothesis is that µ=µ0 (value given to use as µ is usually unknown) whereas the alternate hypothesis is that µ>µ0, µ<µ0, µ≠µ0( Right, left, two tailed tests respectively)
  + To get the test statistic, t= y- µ0/(s/sqrt(n))
  + We then reject or fail to reject null hypothesis if p-value of test statistic at df=n-1 <=a or P>a(fail,fail to reject)
  + s/sqrt(n) is the same thing as Standard Error of estimate of sample mean.
  + Note if you are doing a two-tailed test multiple P-Value bounds by 2
  + Ones alpha is usually the significance level in decimal form
  + 1-a gives us our confidence interval
  + Estimate+- Critical Value \* Standard Error of Estimate where critical value is the value obtained from a table showing the theoretical distribution of test statistic at given confidence level using a/2. (Gives confidence range)
  + If theoretical estimate of mean lies in the confidence range, fail to reject H0.
  + A wider confidence interval is less precise hence we aim for a larger confidence interval and sample size
  + Note must be consistent in the type of tailed test you perform
  + Sometimes P-Vale is called Sig. (SPSS).
* Two Sample mean Case
  + One variable but we compare two population mean(independent/Paired)
  + Independent
    - µ = µ1 - µ2
    - σ=σ1/n1+σ2/n2
* Pooled t test(Two-sample test):
  + We assume standard deviation is equal
  + H0: µ2= µ1 whereas Ha: µ1>µ2, µ1<µ2, µ1≠µ2( Right, left, two tailed tests respectively)
  + T= y1-y2-∆/SE(y1-y2)
    - SE(y1-y2)=Sp\*sqrt((1/n1)+(1/n2)) where sp = sqrt(s12(n1-1)+ s22(n2-1) /(n1+n2-2) )
  + Df =n1+n2-2
  + To get confidence interval it is (y1-y2)+-ta/2\*SE(y1-y2) at df n1+n2-2
  + If 0, is in confidence interval, fail to reject H0
* Paired t-test
  + Applies when two populations/measurements are paired in space/time by some relationship.
  + Same Null and alternate hypothesis as above
  + T= mean difference-∆/(Sd/sqrt(n))
  + Sd=sqrt(∑d2-((∑d) 2/n)/(n-1))
  + Df=n-1
  + D+-ta/2\*SE(d) at df=n-1
    - SE-Sd/sqrt(n)
* Assumptions of Statistical Inference
  + Simple Random Sampling
  + Independence of sampling
    - Only for One-Way ANOVA and two-sample t-test(pooled and non-pooled)
  + Normally Distributed Data
    - One/Two sample t-test
    - Paired sample t-test
    - One-way ANOVA
  + Equal Standard Deviations:
    - Pooled t test and ANOVA
  + When comparing S.Ds divide the largest by smallest and check if ratio is less than 2, they are similar else they differ
  + TO check if normality is violated, check if points lie close to line(Q-Q plot) and if they don’t they violate normality
* Robustness: Ability to withstand departures of violations of assumptions without being too affected .
  + Mean is not robust whereas median is
* You can apply transformations to data(log, sin, sqrt etc.)
  + Usually, log transformations done on following
    - Fits all assumptions
    - Normal but unequal s.d.s
    - Non normal with high sd
    - Two extreme outliers in data
  + You would transform data before performing tests and back-transform after tests are done
* If data does not fit assumptions of parametric tests, use non-parametric tests
  + They don’t use estimates in calculations
  + Less powerful

Chapter 3

* ANOVA is used to find difference among 3+ groups(more than 2)
* Only 1 variable recorded(response) but come from different populations, treatments, or groups.
* All ANOVA uses F-distribution gotten by 2 types of variations and dividing one by the other
* 2 df’s
  + Numerator df – type of variation placed in numerator when getting test statistic
  + Denominator df- variation placed in denominator of f statistic.
* F curve
  + Area under curve =1
  + Starts at 0 in x axis the extends infinitely to right but never touches x-axis
  + Right skewed
  + At df(inf, inf) F=1
* Pooled t-test compares one variable measured in 2 populations; ANOVA compares one variable in 3+ populations
* One-Way ANOVA(Single Factor)
  + Compare one value between groups/populations affected by 1 factor(explanatory variable)
  + Different values of factor are called levels
* Two-Way ANOVA(Two-Factor)
  + Compares values of 1 variable among populations that are classified or grouped to 2 factors
* Factors are either categorical or quantitative
* Multiway Factorial ANOVA deals with comparisons where more than two factors affect population
* Randomized block ANOVA is an extension of paired-sample t test in which you have more than 2 blocked in time or space by some relationship
* ANOVA analyzes variances among population with variances within population
* F statistic is either
  + Between Group Variability/ Within Group Variability
  + F=Treatment Mean Square/Error Mean Square
* One-Way ANOVA
  + Total Sum Of Squares(SSTotal)=Total variation between and within samples/groups
  + Treatment Sum of Squares(SSTreatment)=Variation Between Treatment
    - ∑nj(yj-y) In this case yj is mean of group and y is overall mean
  + Error Sum of Squares(SSError)=Variation with Treatments
    - ∑∑n(yij-yj) In this case yj is mean of group and yij is the data value
  + SSTotal=SSTreatment+ SSError
  + Treatment Mean Square (MSTreatment)= SSTreatment/k-1 where k is number of populations compared
  + Error Mean Square (MSError)= SSError/n-k where k is number of populations compared and n is total number of observations
  + F = MSTreatment/ MSError
  + H0: All means are equal whereas alternate hypothesis is that not all means are equal.
  + Df is (k-1,n-k) k-1 is numerator and n-k is denominator
* If we reject H0 for one-way ANOVA, we perform comparisons to see which mean is different(pairwise comparison)
  + Number of comparisons we need to perform is k(k-1)/2 where k=number of groups being compared
  + Different types of comparisons:
    - Tukey Multiple Comparisons
      * Sample sizes for all groups
      * When equal Cis are shorter than other methods and therefore more likely to show differences
    - Bonferroni
      * Used for a general case where sizes are different
      * Can control overall error rate
    - Fisher
    - Scheffe-wider CI than Tukey
    - Least Significant difference
    - Student-Newman Keuls test
* Tukeys
  + Df=(k,n-k)
  + Get confidence interval
  + (yi-yj)+-Critical Value/sqrt(2) \* SE
    - SE = sqrt(MSE) \*sqrt((1/ni)=1/nj)
  + Put results in matrix and 0 isn’t in interval, we can declare it is different
  + Conclude via means comparison diagram( rank groups in size of mean)
* Bonferonni
  + Get individual comparison-wise(ai) error rate based on family error rate(1-a usually. af) where ai= af/number of comparisons
  + Find t at ai/2 at df=n-k
  + Margin of error is Crit value \* SE
    - SE = sqrt(MSE) \*sqrt((1/ni)=1/nj)
  + Means are different if absolute difference between means >=0 and put results in matric and conclude using mean comparison diagram
* Linear Combinations(planned comparisons)
  + Decide which mean or group of means you want to compare
  + γA-B = (µ1,1+…+µ1,n)/n - (µ2,1+…+µ2,m)/m where A and B are the combinations being compared and n/m are means within those comparisons
  + Then γ=C1µ1+….+ Ckµk where C1+….Ck=0
  + H0: γ=0 whereas Ha: γ≠0
  + Estimate(γe)= C1y1+….+ Ckyk where y is mean of that group
  + SE(γ)=sp\*sqrt(C12/n1+… Ck2/nk)
    - Sp=sqrt(MSE)=sqrt((n1-1)\*s12+…+ (nk-1)\*sk2/n-k)
  + T=γ-0/SE(γ) at df=n-k
  + Confidence interval id γe+-Crit Value\* SE(γ)
* Extra Sum of Squares test
  + Classifies two models: Reduced and full
    - Null hypothesis is reduced model whereas Alt. hypothesis is the full model
  + Extra Sum of Square(extra SS)= SSE(Reduced)- SSE(Full)
  + Extra df= dfE(Reduced)- dfE(Full)
  + F=(Extra SS/Extra df) / (SSE(Full)/Extra df)
    - Same thing as ( (SSE(Reduced)- SSE(Full))/ dfE(Reduced)- dfE(Full) )/ (SSE(Full)/Extra df)
  + When looking at F table df is [Extra df, dfE(Full)=n-k]
* Kurskal-Wallis test(Non-parametric equivalent of ANOVA)
  + Used where there is 3+ independent samples
  + 95% as powerful as ANOVA
    - When data doesn’t fit anova assumption, KW test more powerful
  + Data ranked from highest to lowest and calculations performed on ranks
  + When there are tied observations, assign average rank to tied observations
  + If data is not normally distributed not log normal and size is less than 30 you must used KW test
  + Not affected by outliers
  + Assumptions
    - Independent, simple random samples
    - Same Shape populations
    - 5+ sample size
  + Null hypothesis is that the k populations are equal whereas the alternate hypothesis is that not all populations are identical.
  + H(test stat.)=(12/n(n+1)) \* ∑Rj2/nj -2(n+1 ) sum is from j=1 to k
  + Df=k-1
  + Look at Chi squared table

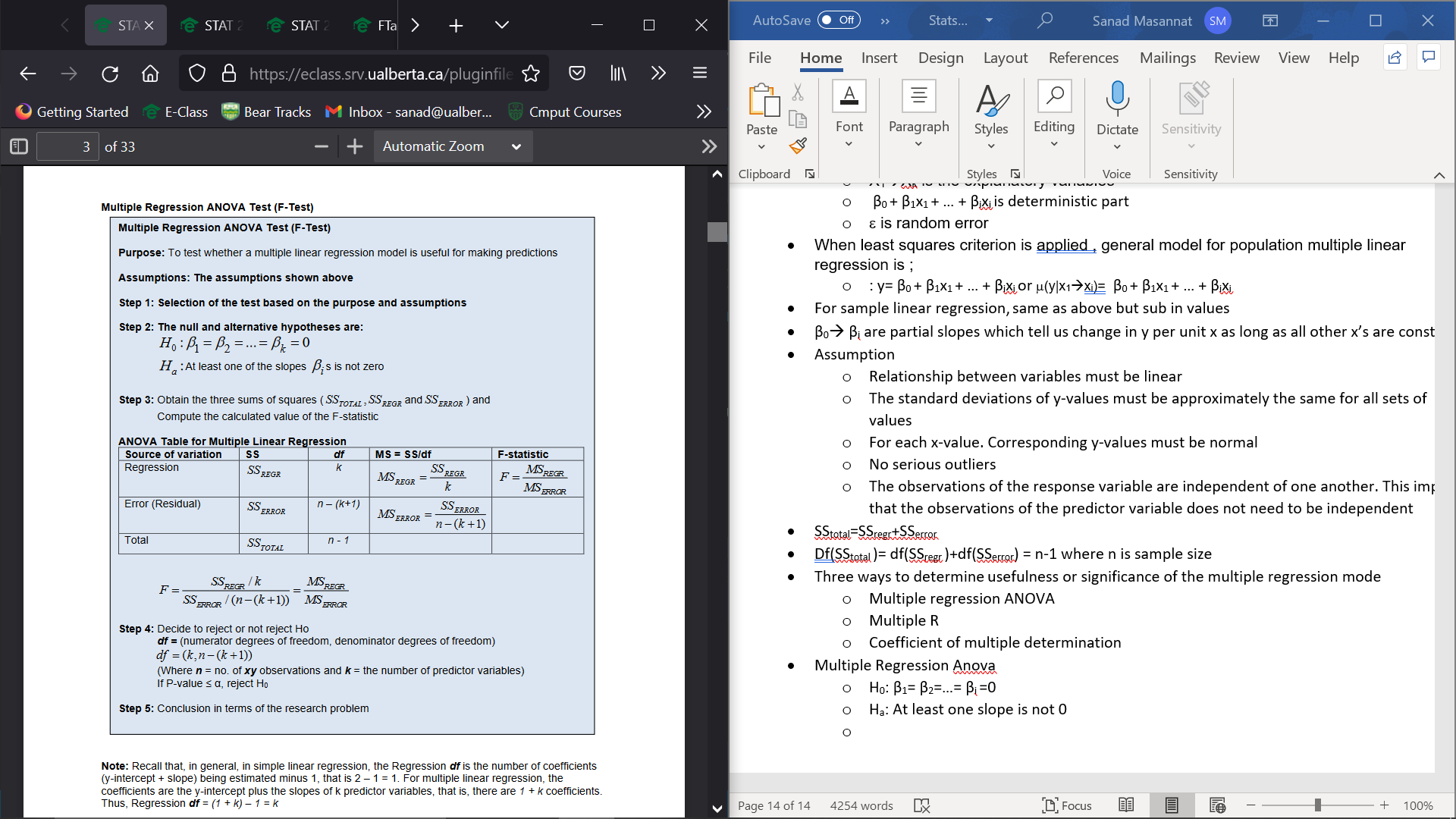
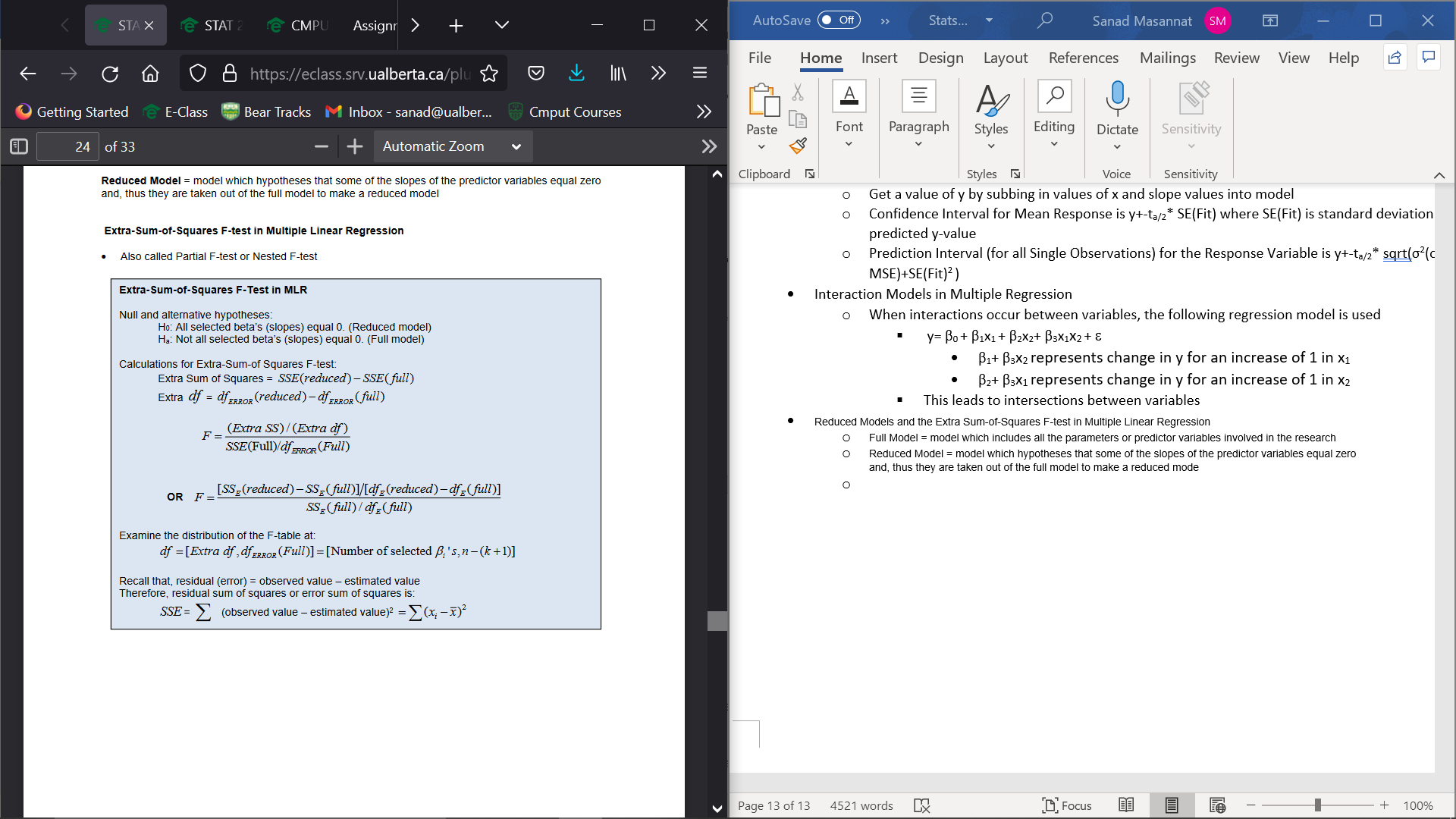
Chapter 4

* Linear Regression and Correlation
  + Analyze relationship between quantitative variables
  + Simple = two variables
  + Scatterplot/ diagrams show relationship between the two
    - If points appear linear, linear relationship exists
  + Linear regression is used to analyze the relationship between 2 quantitative when one variable responds to the other with the explanatory on the x and the response on y axis
  + Simple linear is used to analyze the relationship between two quantitative variables when a change in one variable appears to be related to a change in the other but not responding to another. i.e. two variables are co-related which leads to EITHER x and y axes can be the response/explanatory variable

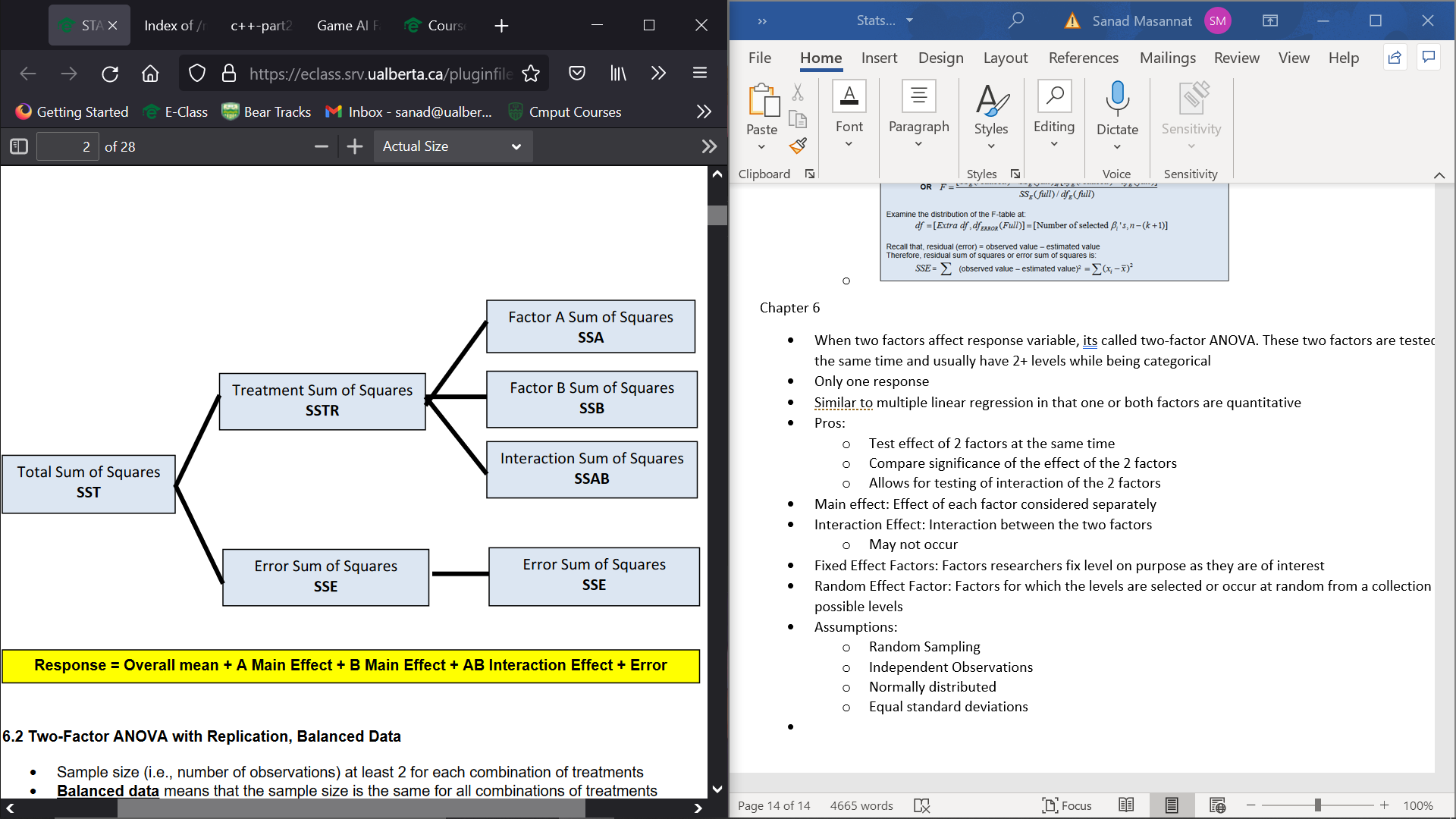
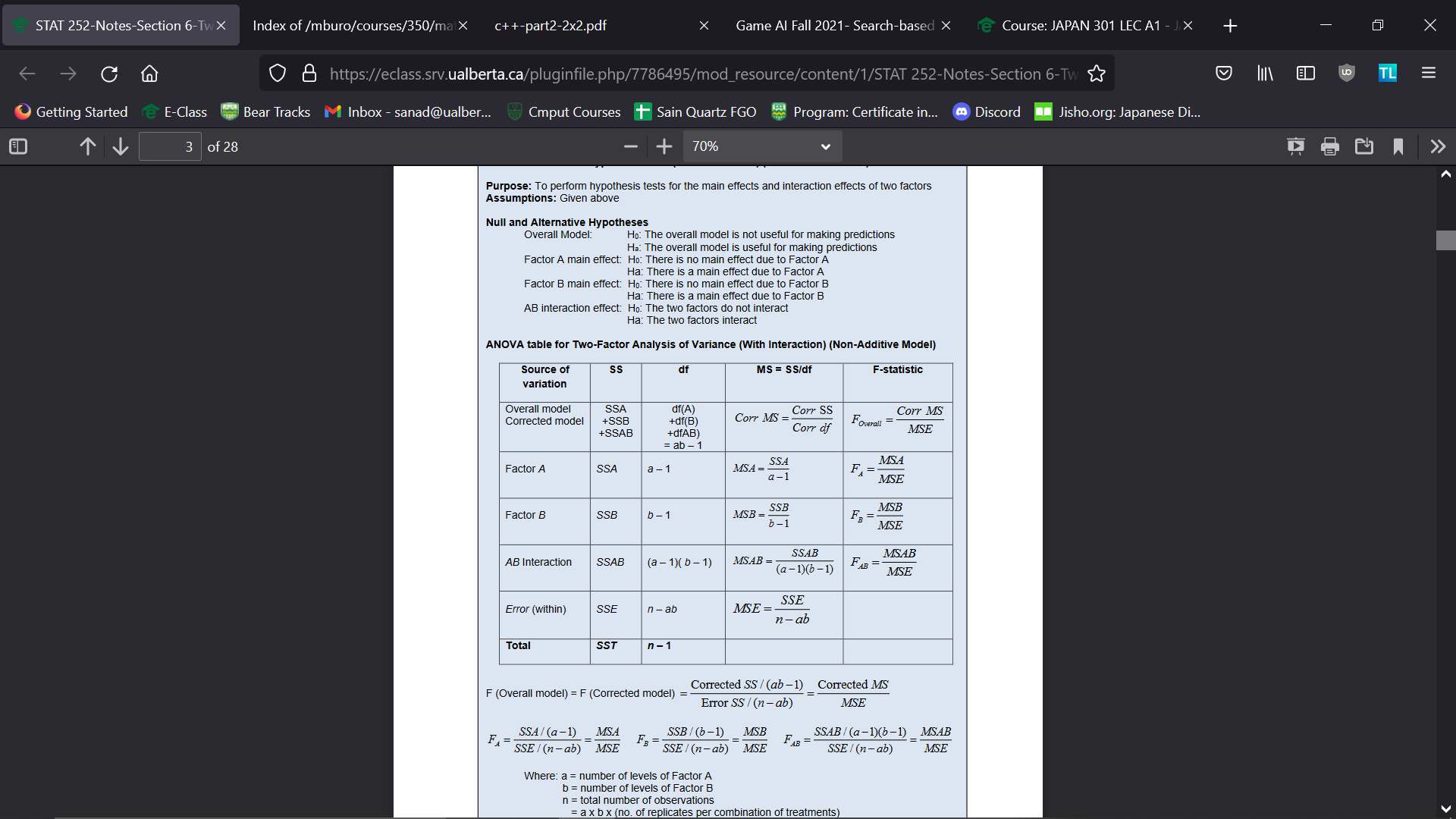
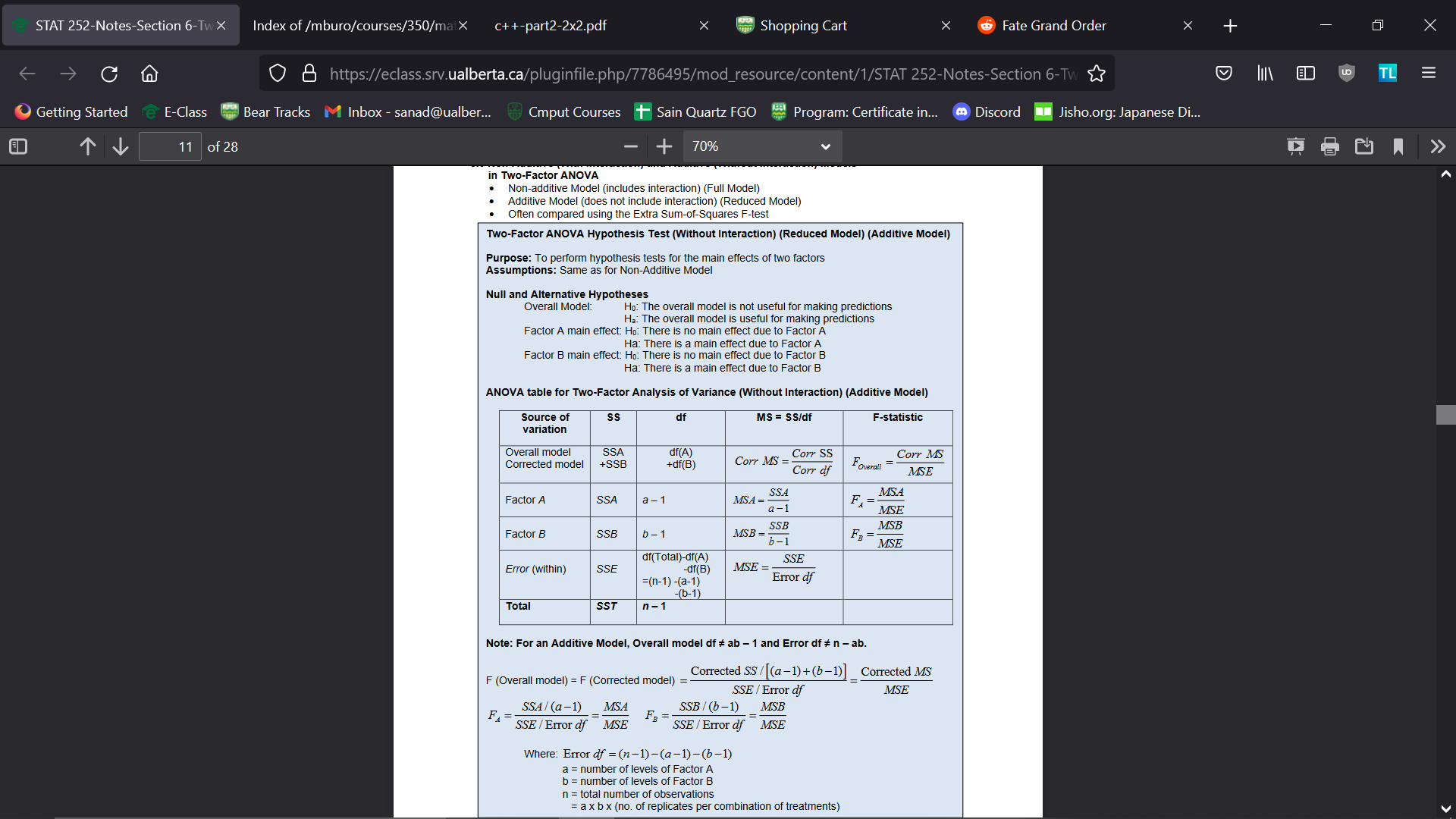


* + Linear Regression line: A best fit line that passes through the points and Is calculated using the least squares criterion
  + Form is μ(Y|X)= β0+ β1X
    - μ(Y|X) means the predicted mean of Y at a given x
    - β0 is the y-intercept β1 is the slope, Y is response and X is explanatory
    - Estimated Model or sample regression line is y=μ(Y|X)= β0+ β1X
    - y is denoting y-valye predicted by regression and β0 & β1 are least square estimates
    - β1=Sxy/Sxx
    - β0 = y -β1\*x
  + Three Sum of Squares in Regression
    - SSTotal=Syy
    - SSRegr=(Sxy)2/Sxx
    - SSError=SSTotal- SSRegr=SSRES
  + Residual = error (denoted as e) yi-yi
  + Least squares criterion tries to minimize SSError in order to get the best fit line which means they try to minimize the errors due to deviations not explained by regression equation
  + Residual plots can be used to check whether data fits regression analysis assumptions
  + If the assumptions for regression inferences are met, the plot of the residuals against the values of the predictor variable should fall roughly in a horizontal band centered and symmetric about the x-axis
  + Standard Error of model Se= σ= sqrt(SSError/n-2)
  + Interpolation is using the regression equation to make predication about the response variable, within the range of observed values of x
  + Extrapolation is using the regression equation to make predication about the response variable, outside the range of observed values of x. This can lead to bad predictions and errors
* Linear Correlation(r) and Co-efficient of Determination(R2)
  + 3 main aspects:
    - Direction: if r is +/-, slope is also +/-
    - Form: Linear or curved
    - Strength:
      * If r is close to -1/1 that means there is a strong negative/positive relationship, the closer they are to extreme value. 0 indicates no relationship
  + Outliers: Data which doesn’t follow pattern of the rest of the data
  + Leverage: Data points whose x-values are far from mean of x.
  + Influential points: A data point which if left out will affect the model
  + Serious outliers and inferential observations:
    - Makes weak correlations look stronger, change slope considerably and could sometimes make a +ve correlation appear -ve.
  + r= Sxx/(Sx\*Sy)=sqrt(SSRegr/SSTotal) between -1 and 1
  + R2=r2= SSRegr/SSTotal
  + R2Adj=1-(MSE/MST)
* Assumptions for Regression Inferences and Analysis of Residuals
  + Relationship between two variables are linear
  + S.d of y values must be the same for all x values
  + For each x, corresponding y values must be normal
  + No serious outliers
  + Observations of response variable are independent of one another.
* Using Anova regression
  + MSRegr=SSregr/dfregr df is usually 1 in this case
  + MS Error=SSError/n-2; n-2= df; This is the standard error of the model squared=variance of model
  + F statistic = MSReg/MSError
  + Null Hypothesis is β1=0 whereas the alternate hypothesis is β1≠0;
    - This means we are checking if there is no relationship between the two variables
  + Perform similar test to standard Anova
  + T2=F
* Inferences for the slop of Population Regression Line
  + Null Hypothesis is β1=0 whereas the alternate hypothesis is β1≠/>/<0;(Two tailed i.e there is a relationship, right-tailed i.e. there is a negative relationship and finally left tailed in which there is a positive relationship)
  + SStotal=Syy
  + SSregr=(Sxy)2/ Sxx
  + SSerror =SStotal-SSError
  + Standard Error of model is sqrt(SSerror/n-2)
  + SE(β1)= Standard Error of model/sqrt(Sxx)
  + T-statistic= (β1)/ SE(β1)
  + Df=n-2
  + Confidence interval= ( β1 – ta/2\* SE(β1) , β1 + ta/2\* SE(β1)
  + To get the confidence interval for mean response:
    - Y=(β0)+ (β1)x This is the point estimate
    - Interval=(pointEstimate- ta/2\*σ\*sqrt(1/n + (x-x)2/Sxx), pointEstimate+ ta/2\*σ\*sqrt(1/n + (x-x)2/Sxx)) x is the mean
    - Prediction Interval is( pointEstimate- ta/2\*σ\*sqrt(1+1/n + (x-x)2/Sxx), pointEstimate+ ta/2\*σ\*sqrt(1+ 1/n + (x-x)2/Sxx))
* Hypothesis testing for linear correlation
  + H0: No correlation between two variables Ha: There is a connection (!= >(+ve) < (-ve)leads two a two tailed right tailed left tailed test respectively)
  + R= sqrt(SSRegr/SSTotal) df=n-2
* Assumptions:
  + Linear: Do not use regression if data isn’t linear and consider transformation
  + Equal S.D: transform y if not equal
  + Normal
  + No Serious Outliers
  + Independent Observations
* Transformations
  + For Y vs X
    - Additive change of k units in X → Additive change of kβ1 in the mean of Y
    - Suppose (in separate circumstances) that the following natural log transformations were required.   
      Interpretation of the model effect on the original scale will follow.   
      i) a natural log transformation was used on the response variable only. (ln Y vs. X)   
      ii) a natural log transformation was used on the predictor variable only. (Y vs. ln X)   
      iii) a natural log transformation was used on both variables. (ln Y vs. ln X)
  + for ln Y vs. X:
    - Additive change of k units in X → Multiplicative change of 1βke in the   
      median of Y. (Take antilog of the slope
  + IN GENERAL, for Y vs. ln X:
    - Multiplicative change by a factor → Additive change of β1ln(k) in the   
      of k in X (X × k) mean of Y.
  + IN GENERAL, for ln Y vs. ln X:
    - Multiplicative change by a factor → Multiplicative change of 1βk in the   
      of k in X (X × k) median of Y.

Chapter 5

* General Form : y= β0 + β1x1 + … + βixi + ε
  + Y is response variable
  + X1🡪Xk is the explanatory variables
  + ­ β0 + β1x1 + … + βixi is deterministic part
  + ε is random error
* When least squares criterion is applied , general model for population multiple linear regression is ;
  + : y= β0 + β1x1 + … + βixi or μ(y|x1🡪xi)= β0 + β1x1 + … + βixi
* For sample linear regression,same as above but sub in values
* β0🡪 βi are partial slopes which tell us change in y per unit x as long as all other x’s are constant
* Assumption
  + Relationship between variables must be linear
  + The standard deviations of y-values must be approximately the same for all sets of values
  + For each x-value. Corresponding y-values must be normal
  + No serious outliers
  + The observations of the response variable are independent of one another. This implies that the observations of the predictor variable does not need to be independent
* SStotal=SSregr+SSerror
* Df(SStotal )= df(SSregr )+df(SSerror) = n-1 where n is sample size
* Three ways to determine usefulness or significance of the multiple regression mode
  + Multiple regression ANOVA
  + Multiple R
  + Coefficient of multiple determination
* Multiple Regression Anova
  + H0: β1= β2=…= βi =0
  + Ha: At least one slope is not 0
  + 
  + Df=(k,n-k+1) n is num of observations and k is number of predictor values
* Multiple R
  + Multiple R = sqrt(R2)
  + R2 = SSregr/SSTotal
  + R2Adj= 1- MSregr/MSTotal
* Multiple regression t-test allows us to determine if a particular predictor variable is useful for making predictions
  + H0: βi=0 whereas Ha: βi is either !=,<,> 0 (Twotailed,left,right)
  + T= value of slope/SE(value of slope) for a df of n-k+1
  + Confidence interval is : value of slope+-ta/2\* SE(value of slope)
* Confidence Interval and Prediction Interval for the Response Variable
  + Df = n-k+1
  + Get a value of y by subbing in values of x and slope values into model
  + Confidence Interval for Mean Response is y+-ta/2\* SE(Fit) where SE(Fit) is standard deviation of predicted y-value
  + Prediction Interval (for all Single Observations) for the Response Variable is y+-ta/2\* sqrt(σ2(or MSE)+SE(Fit)2 )
* Interaction Models in Multiple Regression
  + When interactions occur between variables, the following regression model is used
    - y= β0 + β1x1 + β2x2+ β3x1x2 + ε
      * β1+ β3x2 represents change in y for an increase of 1 in x1
      * β2+ β3x1 represents change in y for an increase of 1 in x2
    - This leads to intersections between variables
* Reduced Models and the Extra Sum-of-Squares F-test in Multiple Linear Regression
  + Full Model = model which includes all the parameters or predictor variables involved in the research
  + Reduced Model = model which hypotheses that some of the slopes of the predictor variables equal zero   
    and, thus they are taken out of the full model to make a reduced mode
  + 

Chapter 6

* When two factors affect response variable, its called two-factor ANOVA. These two factors are tested at the same time and usually have 2+ levels while being categorical
* Only one response
* Similar to multiple linear regression in that one or both factors are quantitative
* Pros:
  + Test effect of 2 factors at the same time
  + Compare significance of the effect of the 2 factors
  + Allows for testing of interaction of the 2 factors
* Main effect: Effect of each factor considered separately
* Interaction Effect: Interaction between the two factors
  + May not occur
* Fixed Effect Factors: Factors researchers fix level on purpose as they are of interest
* Random Effect Factor: Factors for which the levels are selected or occur at random from a collection of possible levels
* Assumptions:
  + Random Sampling
  + Independent Observations
  + Normally distributed
  + Equal standard deviations
* 
* Two-Factor ANOVA with Replication,Balanced Data
  + The sample size must at least be 2 for each different combination
  + Balanced Data: Sample Size Is same for all combinations of treatments
* Total Sum of Squares = Factor A Sum of Squares + Factor B Sum of Squares   
  + AB Interaction Sum of Squares + Error Sum of Squares
  + SST=SSA+SSB+SSAB+SSE
* Df(SSR)= df(SSA)+df(SSB)+df(SSAB)+df(SSE) = n-1 = a-1+b-1+(a-1)(b-1)+(n-ab)
* Two-Factor ANOVA Hypothesis Test (With Interaction) (Non-Additive Model)
* 
* Non-Additive (With Interaction) and Additive (Without Interaction) Models
  + Non-additive is full model whereas additive is the reduced model
  + 
* When comparing additive vs non-additive, use the Extra Sum of Squares Test
* Randomized Block Design