#### STA302 - Part 2 Flowchart Diagram Key Madhav Kanna Thenappan , Happy Nasit, Khushil Nagda, Kevin | November 13, 2023 Module 1 & 2 Module 3 Start: Choose Research Question + necessary Module 4 variables Module 5 Module 6 **Treat Categorical Variables Correctly** Miscellaneous 1. Turning the Categorical variable Fit the prelimnary model and state the into a Dummy variable. relationship between: a. Encode the gender variable, Male -Gender 0 and Female 1 b. The coefficient will represent the -Regular Sleeping Hours - Direct HDL Cholestrol difference in number of bad -Amount of physical activity per day mental health days between -Age Males and Females -Amount of alcohol used per year 2. For further Analysis, we can add -BMI interaction terms. -Systolic BP a. Gender: Age - allows us to see -Diastolic BP whether bad mental health affects With categorical variables being Gender older/younger males and females with categories MALE & FEMALE. differently. The model is created such that: DaysMentHlthBad ~ Gender + Analyse Assumptions SleepHrsNight + DirectChol + for the Multivariate PhysActiveDays + Age + AlcoholYear + BMI regression model + BPSysAve + BPDiaAve Plot the graphs to analyze assumptions. - residuals vs fitted values Analyze the graphs - response vs predictor: and check for Dayswmentalhealthbad vs violation of -Gender assumptions. -Regular Sleeping Hours -Direct HDL Cholestrol -Amount of physical activity per day We use a Box-Cox -Amount of alcohol used per year transformation on the predictor -Systolic BP • Our aim is to find the best value of **Violation in Linearity?** -Diastolic BP λ that makes the data as close to - Do we not see approximately null - residual vs each predictor variables. normal as possible. This is ideally plots for residual vs fited-values? done using maximum likelihood - standardised residual histograms Do we not see linearity between estimation but due to the complex QQ Plots between standardised residuals response vs predictor and vs fitted nature of it we aim for simple and the variables in the multi linear model powers such as a log values? transformation. Applying Box-Cox Transformation on response variable From the part 1 QQ plot, Normality was violated. To correct it, we use a Box-Cox iolation in Normality transformation on the response variable. • Our aim is to find the best value of λ that - Does the QQ plot not makes the data as close to normal as approximately look like a line? possible. This is ideally done using Standardized residuals should be maximum likelihood estimation but due approximately normal to the complex nature of it we aim for simple powers such as a square-root transformation. Νo -Apply a Variance Stabilizing Transformation. - The goal of the transformation is to make the variance of the residuals approximately constant across different Violation in Constant Variance? levels of the predictors. Does the plot of residuals not vai -This transformation will be applied on without obvious patterns around 0? the DaysMentHlthBad (reponse) variable. - We start by applying a square root transformation because it is often used with counts, in this case number of days.

# STA302 - Part 2 Flowchart **Diagram Key** Madhav Kanna Thenappan , Happy Nasit, Khushil Nagda, Kevin | November 13, 2023 Module 1 & 2 Module 3 Module 4 Module 5 Module 6 Miscellaneous

Confirm that both Linearity Assumptions and Conditionas have been satisfied.

- If we calculate a confidence interval and/or a prediction interval using data that violates the linearity assumption, predictions made for new observations will be systematically understated or overstated because the model does not represent the true relationship between the variables.

We want to estimate the effect of the predictors on the population

Calculate the 95% confidence interval for each predictor in the multlinear

The confidence interval tells us the range within which we can be confident that the true population parameter lies given a certain level of confidence.

- The confidence interval tells us that If we were to repeatedly take random samples of the same size from the population and construct a 95% confidence interval for a particular regression coefficient each time using the same model, then we would expect 95% of those confidence intervals to contain the true population coefficient for that predictor
- The value of the confidince interval that we obtain can tell us two things.
- 1. A CI that doesn't have zero means the effect of the predictor on the response is statistically significant.
- 2. The width of the interval gives an indication of the precision of the estimate i.e. narrow CI suggests more precise estimate.

#### Perform partial F test with the reduced model:

- 1) Construct Null and Alternative hypotheses:
  - Null Hypothesis: the reduced model is sufficient
  - Alternative Hypothesis: the reduced model is not sufficient
- 2) Define and compute the test statistic:

The random test statistic is the ratio of mean sums of squares with sampling distributions given by:

$$F = rac{MS_{reg}}{MSR} \sim F(p-r,n-p-1)$$

F\* is the computed value of the test statistic with the data

- 3) Calculate the p value, the probability that F\* is at least as large or larger than the random test statistic. We asume alpha = 0.05.
- 4) Conclusion based on p value:
  - If p value < 0.05, then we reject the null hypothesis, and thus the reduced model is not sufficient.
  - If p value >= 0.05, then we do not reject the null hypothesis, and thus the reduced model is sufficient.

End

We are now close to answering our research question: To what extent can biological and lifestyle markers predict depressive symptoms among

survey respondents 18 - 80 years old?

We'd like to know the range within which we expect future individual observations of depressive symptoms to fall, with a 95% level of confidence.

### So, we calculate a prediction interval

Calculate the 95% prediction interval for the response variable.

- Prediction interval is a range of values within which we expect a single new observation to fall, with a 95% level of confidence
- If we were to repeatedly sample from the population according to the regression model and also a new Y\*, x\* and compute a 95% prediction interval each time, then 95% of the intervals would include the population response value.

Start with i = 1 th coefficeint when considering hypothesis tests for individual coefficients

## Perform hypothesis tests for individual coefficients of predictor variables.

- 1) For the response predictor variable pair, construct:
  - Null Hypothesis: coefficient i = 0
  - Alternate Hypothesis: coefficient i not = 0
- 2) The test statistic follows the T distribution:

$$T = \frac{\hat{\beta_i}}{\hat{se}(\hat{\beta_i})} \sim t_{n-(p-1)}$$

- 3) Find the p value, the probability that T\* (computed test statistic from data) is at least or more extreme than the random test statistic. We assume a value of alpha = 0.05.
- 4) Conclusion based on p-value:
  - If p value < 0.05, then we reject the null hypothesis
- If p value >= 0.05, then we do not reject the null hypothesis. If we do not reject the null hypothesis, we do not include the predictor variable in our reduced model for the F test, and we include the predictor variable in the reduced model if we reject the null hypothesis.

