Statistical Analysis Plan:

Prevalence and correlates of depression and anxiety in a sample of Latinx college students in the later stages of the COVID-19 pandemic

Or

Prevalence and correlates of depression and anxiety in a sample of Puerto Rican college students in the later stages of the COVID-19 pandemic

1. **Data cleaning and Feature engineering**
   1. Exclude survey entries from non-consensuals
   2. Exclude those who weren’t enrolled in any classes
   3. Exclude those who 6.5 minutes or less to complete survey
   4. Separate multiple response questions into separate columns
   5. Recode answers for ease of use
   6. Change column names
   7. Set columns to correct variable types
   8. Change “DK”, “Refuse”, and “Prefer not to answer” to NA
   9. Correctly code NAs produced from questionnaire’s branching logic
2. **Data exploration**
   1. Categorical variable distributions
      1. Bar charts
      2. Frequency tables
   2. Verify categorical variables with unbalanced categories and consolidate categories when necessary/possible
   3. Numerical variable distributions
      1. Box plots
      2. Summary stats
   4. Search for outliers and correct when necessary/possible
   5. Etc.
3. **Missing Data Analysis**
   1. Missingness plots
   2. Visual inspection of missingness patterns
   3. Flux plot (for searching easily imputed variables and useful variables for imputing – *mice*)
   4. Flag variables (create dummies) with >50% missing data.
4. **Determine which variables will be outcome, predictors, interaction terms, and auxiliary variables.**
   1. Outcome = dependent variable, variable that will be modeled
   2. Predictors = variables that predict outcome, independent variables
   3. Interaction terms = variables that can interact with each other
   4. Auxiliary = variables used for imputation, but not as predictors

Before performing any sort of imputation or modeling, we should first predetermine which variables will be used for modeling. This can help facilitate downstream analyses (e.g. less variables to impute, less features to select from in model selection, etc.), given these are computationally intensive. The following table specifies preliminarily how features should be divided:

|  |  |  |
| --- | --- | --- |
| **Outcome** | **Predictors** | **Auxiliary** |
| 1. PHQ-9 2. GAD-7 | SEX  GENDER  SEX\_ORIENT  AGE  RACE  MARITAL\_STAT  INCOME\_YR  EMPLOY  ACA\_YR  CRD\_HRS  FACUL  ACA\_LVL\_BI  GPA\_SR  TECH\_ACSS  DIFF\_ACSS  PART\_DROP  EDUC\_MOD  PREEX\_COND  DEP\_HIST  ANX\_HIST  SOC\_MED  SM\_TIME\_DAY  SM\_TIME\_WK\_YAM  SM\_REAS  VID\_GMS  VG\_TYPE  VG\_VIOL  VG\_TIME\_WK\_YAM  VG\_TIME\_DAY\_YAM  VG\_REAS  VG\_IMP\_QOL  EXER\_LAST7  EXER\_DAYS\_IN\_LAST7  EXER\_DURAT\_MIN  SLEEP\_HRS  EAT\_OUT  EAT\_FRT  EAT\_VEG  ALC  ALC\_30  NICOT  NICOT\_30  CANNA  CANNA\_30  COCAINE  COC\_POWD\_30  CRACK\_30  MED  MED\_TYPE  MED\_30  HALUC  HALUC\_TYPE  HALUC\_30  REGION\_RES  PSS\_10  TILS  DEP\_TRT  ANX\_TRT | MUNIC  OTHR\_EMPLOY  WRK\_HRS\_WK  LOST\_JOB\_CV19  GOV\_ASSIST\_  UNIV\_ECON\_AST  CV19\_HIST  VAX\_STAT  VAX\_TYPE  VAX\_DOSE  SM\_TIME\_DAY\_BP  SM\_TIME\_W\_BP\_YAM  SM\_REAS\_BP  SM\_IMP\_QOL  VG\_TIME\_DAY  VG\_TIME\_WK\_DP\_YAM  VG\_TIME\_DAY\_DP\_YAM  VG\_TIME\_DAY\_BP  VG\_REAS\_BP  NICOT\_TYPE  COC\_TYPE |

1. **Multiple Imputation**

Missing values will be imputed using multivariate imputation by chained equations (MICE; Van Buuren, 2018). Useful predictors and auxiliary variables will be used for creating the imputation model. The number of imputed data sets will be dependent on computing power available and share of missingness in the data.

*Note:* the following steps will be applied to all imputed data sets and later pooled using Rubin’s rules to obtain the final results.

1. **Sample Weighting (Raking)**

Known population totals from UPR’s institutional profile data will be used to construct population weights that will be applied when estimating the weighted prevalence of depression and anxiety in our sample, including standard errors (SE) and confidence intervals (CI). However, raking will not be applied to predictive models used later for feature selection. The *pewmethods* (Pew Research Center, 2020) package in R will be used for raking.

1. **Estimate weighted and unweighted prevalence of major depression**

After obtaining weights, the weighted and unweighted prevalence of depression (defined as having a PHQ-9 score of 10 or more), SE, and CI will be estimated for each imputed data set and later pooled using Rubin’s rules. Estimates of different severity levels, along with their SE and CI will also be calculated. A cutoff of 10 was found to maximize sensitivity (88%) and specificity (88%) for detecting likely cases of major depression (Kroenke, Spitzer & Williams, 2001).

1. **Estimate weighted and unweighted prevalence of generalized anxiety disorder**

After obtaining weights, the weighted and unweighted prevalence of depression (defined as having a GAD-7 score of 10 or more), SE, and CI will be estimated for each imputed data set and later pooled using Rubin’s rules. Estimates of different severity levels, along with their SE and CI will also be calculated. A cutoff of 10 was found to maximize sensitivity (89%) and specificity (82%) for detecting likely cases of major depression (Spitzer et al., 2006).

1. **Specify and fit measurement model for PHQ-9, GAD-7 and PSS-10**

A measurement model will be specified using different latent variable models: correlated factors model, hierarchical factor mode, and bifactor models. In these models, depression and anxiety will be conceptualized as separate but correlated constructs. Perceived stress (Perceived Stress Scale; PSS-10) will also be added to these measurement models as a correlated construct with anxiety and depression. Additionally, an over-arching, general factor (possibly defined as “Psychological Distress”) will be conceptualized and tested for fit. The final measurement model will be used as the outcome to be predicted in the structural model. Furthermore, all other scale variables (COVID-19 Impact on Quality of Life Scale (COV19-QoL) and Three-Item Loneliness Scale (TILS) scales) will be represented as latent variables in the structural model. They will be modeled according to how it’s been conceptualized in the literature (single-factor models).

1. Specify and fit structural model
   1. A structural equation model regression (SEM regression) will be fitted with the measurement model selected in the last step as the outcome and all other predictors.
   2. A feature selection algorithm called Recursive Feature Elimination (RFE; Guyon et al., 2002) that has been known to be applied to support vector machines for feature selection. However, this algorithm is not limited to any particular model, but is more so a feature selection strategy for reaching an optimal subset of predictive features. The process will be undertaken in the following set of steps:
      1. RFE reduces the number of features up to a predetermined amount. This amount can be decided by the researcher or by a process of cross-validation. For this analysis, the cross-validation method will be used. Furthermore, for feature selection, we will be fitting these models on a stacked version of the imputed data sets to ease computational load. This data set of stacked imputed sets will be weighted by .
      2. A series of SEM regressions will be fit to an increasing number of features, from 1 to p predictors. Each model will be cross-validated to a testing set using the *cvsem* package, and the Kullback-Leibler distance will be calculated between the model implied covariance matrix and the testing set covariance matrix. The split between training and testing sets will be 60/40. Given predictors, this process will be repeated for iterations.
      3. The KL distances will be plotted against the amount of predictors in each model, and the optimal number of features to leave in the final model will be selected.
      4. In the next step, we start with a fully saturated model that regresses all features to the outcome of interest. Each feature will be ranked by feature importance (calculated as permutation-based feature importance) from greatest to least importance. The two least importance features will be eliminated in each step until features have been reached. As we reach k features, the amount of features eliminated in each iteration will be reduced to just one feature.
      5. After reaching the reduced feature model, the results will be evaluated against theory and the literature. Depending on final results and what theory may suggest, certain variables will be left in due to there being evidence of them being useful predictors of depression and anxiety (such as stress).
2. **Sensitivity Analysis (missingness assumption)**
3. **Sensitivity Analysis (omitted variable bias)**

**References**

1. Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.
2. Pew Research Center (2020). "pewmethods". github.com/pewresearch/pewmethods.
3. Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: validity of a brief depression severity measure. *Journal of general internal medicine*, *16*(9), 606–613. <https://doi.org/10.1046/j.1525-1497.2001.016009606.x>
4. Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of internal medicine*, *166*(10), 1092–1097. <https://doi.org/10.1001/archinte.166.10.1092>
5. Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, *46*, 389-422.