**ST201 Project Coversheet**

**Project Title: Studying Generalized Anxiety Levels during COVID-19 Using Linear Regression Analysis**

**Word count:**

**Candidate IDs: 45448**

**Date: 4 May 2023**

Permission to use as an example

We give permission for my assignment to be used as an example in ST201 (in an electronic form).

**Introduction**

**Data and Exploratory Analysis**

**Dataset**

The dataset used contains 1115 observations of participants in a European survey on the COVID-19 pandemic. The study records demographics information such as sex, age, and education of the participants, and their survey responses to the extent at which the COVID-19 pandemic has posed difficulties to different aspects of their lives. Gad\_score, a standardized score for generalized anxiety levels during the pandemic, is the variable of interest.

|  |  |
| --- | --- |
| Data type | Variable |
| Numeric | Age, Pandemic\_Difficulties\_\*(All), Covid\_risk, Social\_support |
| Factor | Sex, Education, IncomeContinuity, HealthStatus, Unemployed, Student |
| Outcome variable | Gad\_score (numeric) |

**Descriptive statistics**

Table

Description automatically generated

**Table 1 | Descriptive Statistic for numeric variables.** The youngest participant is 18 yearls old and the oldest is 85 years old. Pandemic\_Difficulties\_\* have been renamed to PD\*. There are no missing data. The Gad\_score variable is standardized, with mean 0 and standard deviation around 1.

Table

Description automatically generated

**Table 2 | Descriptive Statistic for categorical/dummy variables.** Sex0= Male, Sex1=Female. around 1. Dummy variables are generated for each category of Education and HealthStatus, both originally stored as integers. Education2 = Vocational education, Education3 = Secondary education, Education4 = Post-secondary education, Education5 = University education and above. Ommited group for education is primary education. HealthStatus3 = No pre-existing health condition, HealthStatus 2 = don’t know, HealthStatus omitted group = Has pre-existing health condition. IncomeContinuity have 391 missing values and NAs are generated and used as a category.

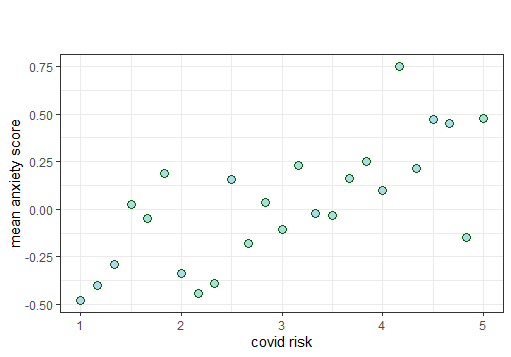
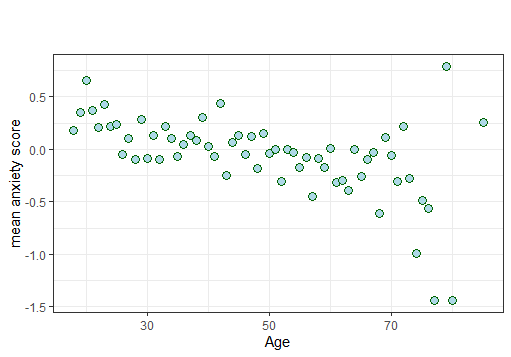
**Exploratory Data Analysis**

Exploratory data analysis was carried out to ensure the correct functional form is used when building the model. First, a correlation matrix was constructed between variables with numeric values (Appendix A) to address concerns with perfect or near-perfect collinearity in the data. When correlation between two variables is relatively high (e.g., ), their estimated coefficients in a linear model can become unstable as it becomes difficult to identify the effect of an individual regressor. Multicollinearity also poses difficulty in inference as standard errors are likely to increase. Namely, high levels of correlations are expected between values for different types of pandemic anxiety, as an individual who face more difficulties in one aspect of the pandemic are expected to be more prone to other pandemic difficulties. The correlation matrix yielded positive pair-wise correlations between all pandemic difficulty levels as expected, but no pair-wise correlations exceeded 0.50 between the numeric variables. As a result, multicollinearity was not of particular concern in the regression analysis.

One other concern was that pandemic difficulties record discrete responses, and the true functional form might not be linear in the pandemic difficulty value, then perhaps generating dummies for each level of difficulty (e.g., PD1==1, PD1==2, etc) better captures the true functional form. Plotting Gad\_score against different pandemic difficulties (Appendix B) showed that the increment at which Gad\_score increased with each discrete jump (from 1 to 2, from 2 to 3, etc) in pandemic difficulties are mostly constant throughout.

Continuous variables were plotted against Gad\_score (Figure 1) to check functional form assumptions. For covid risk, age, and social support (Appendix C), there seem to be linear relationships with Gad\_score. The signs (positive for covid risk, negative for a) were expected. There might exist potential outliers for participants with very high age (Figure 1, b). Nonetheless, the high variance can be induced by the lack of observations for individuals beyond 70 years of age paired with high variance of anxiety score for any age group.

**a**  **b**

**Figure 1 | Relationship between Covid\_risk/Age and anxiety score.**

**a,** Mean anxiety score (Gad\_score) plotted against covid\_risk. For easier visualization the mean value for all Gad\_scores in each Covid\_risk group is plotted using group\_by.mean() in R.

**b**, Mean anxiety score plotted against Age. Similarly, the mean Gad\_score is generated for each age and plotted.

From the preliminary analysis there is no reason to believe functional forms with higher degree polynomials exist in the population. Therefore, only linear terms of those variables will be included in the first stage of model building. Note that the potential for the true function to contain higher polynomials is not dismissed as pair-wise scatterplots in Figure 1 do not control for other features. For instance, if residual plots during model building yield different results, then adding polynomial regression will be considered.

**Regression Analysis and Results**

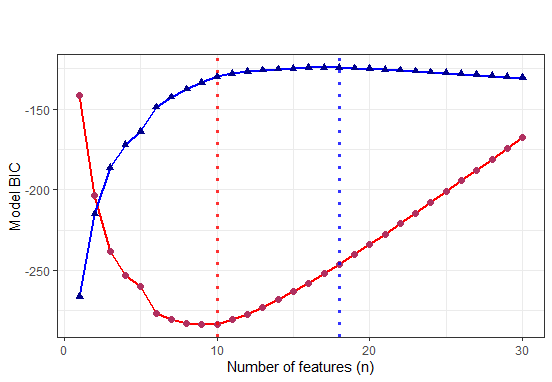
**Initializing the model**

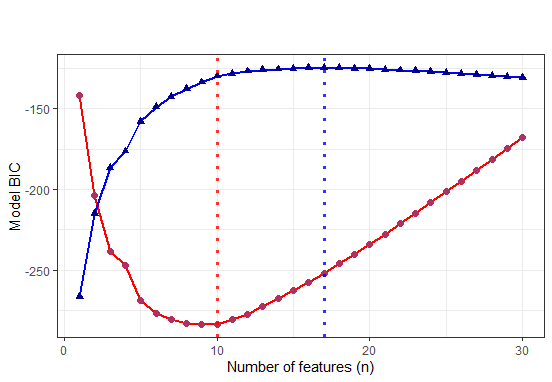
The model building process began with an initial linear regression model (Appendix E) containing all available features in the dataset. All variables were included because for each independent variable in the dataset there exist arguments that they should affect the generalized anxiety level. The initial linear regression model yielded a value of , so of the variations in generalized anxiety level were explained by the model. However, the value is likely inflated as regressors were included in this initial model. With the OLS approach, adding more regressors in a linear regression always yields higher . The adjusted value, , is more indicative of the model’s performance.

Observing coefficients in the initial model, most coefficients in the initial regression were not statistically significant at the 5% level. It is however unwise to remove all variables that are not statistically significant, as 1) a set of variables can appear as not statistically significant due to multicollinearity, and 2) some variables may appear statistically significant by chance. Therefore, stepwise feature selection was used to reduce the number of features and add interpretability to the model. Feature selection can also prevent overfitting and allows better predictions for new data.

**Feature Selection**

Forward and backward selection was used to reduce the number of features. (Appendix E) Both forward and backward selection are stepwise selection methods that aim to produce the model that best fits the data with a reduced number of features. Forward selection begins with a model with only intercept and add single features at a time, with the feature improves model fit having higher priority, while backward selection begins with a model with all features and eliminates one feature that undermines model fit the least iteratively.

**a**   **b**



**Figure 2 | Adjusted and BIC of models of different numbers of features used in stepwise selection.**

**a,** Forward stepwise selection. The red line represents BIC score, and blue line represents adjusted at different number of features. As shown with respective dotted lines, Here n=10 minimizes BIC and n=18 maximizes adjusted

**b,** Backward stepwise selection. Same with forward selection, n=10 also minimizes the BIC score. n=17 now maximizes adjusted

For stepwise feature selection, an approximation for the best set of features (best subset selection is required for the strictly best set of features) is given for every choice of numbers of features. As the model initially had 30 regressors, 30 different models were returned for forward and backward feature selection.

The next step is to select the number of features to use for the model. BIC (Bayesian Information Criterion)[[1]](#footnote-1) score is used for this procedure, as the BIC penalizes model inaccuracy and high number of features in the model. For both forward and backward selection BIC for each number of features were calculated (Figure 2). It turns out that minimizes the BIC for both forward and backwards selection. There has also been consideration to use adjusted as the criterion for choosing the model, but as Figure 2 shows, this would result in a model with more parameters and reduced interpretability. It also turns out that when , forward and backward stepwise selection offers the same linear regression model, which also saves the effort of choosing between the two models. (See Figure 3 for the model until this stage)

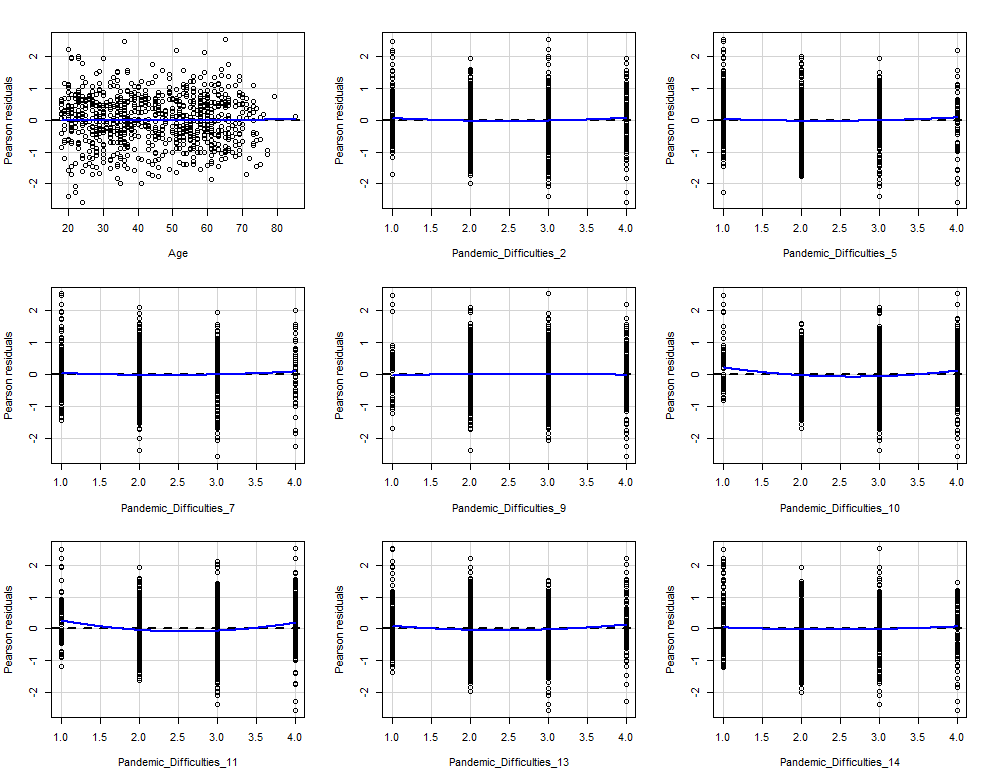
**Examining effects of covid risk on different age groups: adding interactions**

So far, the linear regression model does not involve any interaction terms of regressors, thus it has been assumed that relationship between key variables, such as age or covid risk, and generalized anxiety level, are independent of factors such as gender, education, and income continuity. By imposing a functional form that has interactions on the linear regression model, we allow association of regressors on generalized anxiety levels to be dependent on other regressors. Namely, whether the association between covid risk and generalized anxiety levels depend on a person’s age is of particular interest. Nonetheless, interactions of age with all other variables will be added to begin with. After this, feature selection will be performed again with the same methods to filter out interaction terms that do not help achieve better model fits significantly.

**Checking residual plots and adding second degree polynomial terms**

Residual plots were constructed to ensure the most suitable functional form is used in the linear regression model. (Appendix F) Upon observation of the plots, the linear functional form seemed suitable for all features with exception for Pandemic\_difficulties\_10 and Pandemic\_difficulties\_11, for which the quadratic functional form seemed more suitable. (figure 3) Residual plots for continuous variables (Age, Social\_support, and Covid\_risk) supports the previous assumption that a linear functional form for those variables is sufficient.

Since both Pandemic\_difficulties\_10 and Pandemic\_difficulties\_11 have discrete values (survey responses between 1, 2, 3, and 4), arguments can be made for transforming those variables into categorical/dummy variables. However, after a dummy variable is created for each response, some of those variables may be of less statistical significance and be eliminated during stepwise feature selection, which leaves an incomplete set of response variables in the final model. Such a model may be better at generating predictions but will be less intuitive in giving interpretations.



**Figure 3 | Residual plots of linear regression model after feature selection.**

The residual plots show that the linear functional form is suitable for most variables except for Pandemic\_Difficulties\_10 and Pandemic\_Difficulties\_11. Second degree polynomial terms can be used for those variables to better capture the functional form for these variables. See Appendix F for full set of residual plots.

**Final model presentation**

A picture containing text, screenshot, number, menu

Description automatically generated

**Table 3 | Models reached after stepwise selection at different stages.**

The coefficients (standard error in parenthesis) suggest the direction and strength of the association between the variables and Gad\_score, the dependent variable. Asterisks indicate statistical significance, with more asterisks indicating greater significance. For models with all features see Appendix G.

The iterative building process of the linear regression model is shown in Figure 3. Model (1) is the model obtained after stepwise feature selection from OLS with a linear functional term of all variables and with no interaction terms. After adding interaction terms of Age with all other variables, model (2) and (3) are obtained through forward and backward feature selection respectively and yielded very distinct choices on features. That said, model (2) has relatively more intuitive interpretations in coefficients, thus terms of Pandemic\_Difficulties\_10 and Pandemic\_Difficulties\_11 are further added to arrive at model (4).

**Assessing model accuracy with cross validation**

A picture containing text, screenshot, font, number

Description automatically generated

\*Final linear model for inference

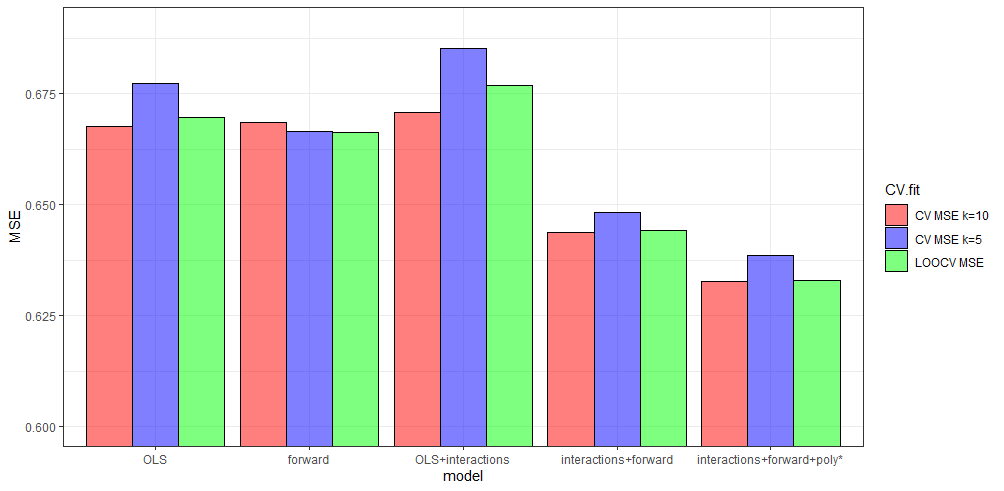
**Table 4 | Cross-validation mean squared errors for tested linear models.**

The coefficients (standard error in parenthesis) suggest the direction and strength of the association between the variables and Gad\_score, the dependent variable. Asterisks indicate statistical significance, with more asterisks indicating greater significance. For models with all features see Appendix G.

Prediction accuracies of the linear models are assessed to adopt a model that most accurately interprets the relationship between predictors and Gad\_score. Scores such as do not inform the model accuracy when used to generate predictions based on new data, as models’ accuracies are best evaluated based on accuracies of predictions on new data.

Cross-validation tests were therefore carried out to examine the prediction accuracy of linear models in the model building process. Mean squared error of cross-validation test sets are computed at , , and (LOOCV) respectively. (Table 4) The initial linear model with all predictors (OLS) has marginally higher cross-validation MSEs across the board when compared to the feature selection model based on the same set of variables. When new interaction terms with the Age predictor were added (‘OLS+interaction terms’, Table 4) the MSE increased for CV, suggesting overfitting of the test data. Once feature selection was used again, there is a larger drop in MSE, which implies significance of adding the interactions and perhaps collinearities in the interacted terms.

Figure 4 provides visual comparison between cross-validation performance of the linear models. Models with less predictors generally outperformed those with more predictors. The feature selected model with interaction and polynomial terms (model(4) in Table 3) has the best test performance.

****

**Figure 4 | Cross-validation mean squared errors for tested linear models at k=5, k=10, and k=1115**

From left to right are CV MSE of iterations of the linear regression model. The vertical axis does not begin at zero but 0.6 to show better comparisons between model performance. The true variation in MSE of the models are less significant than the figure may suggest.

**Discussion and Limitations**

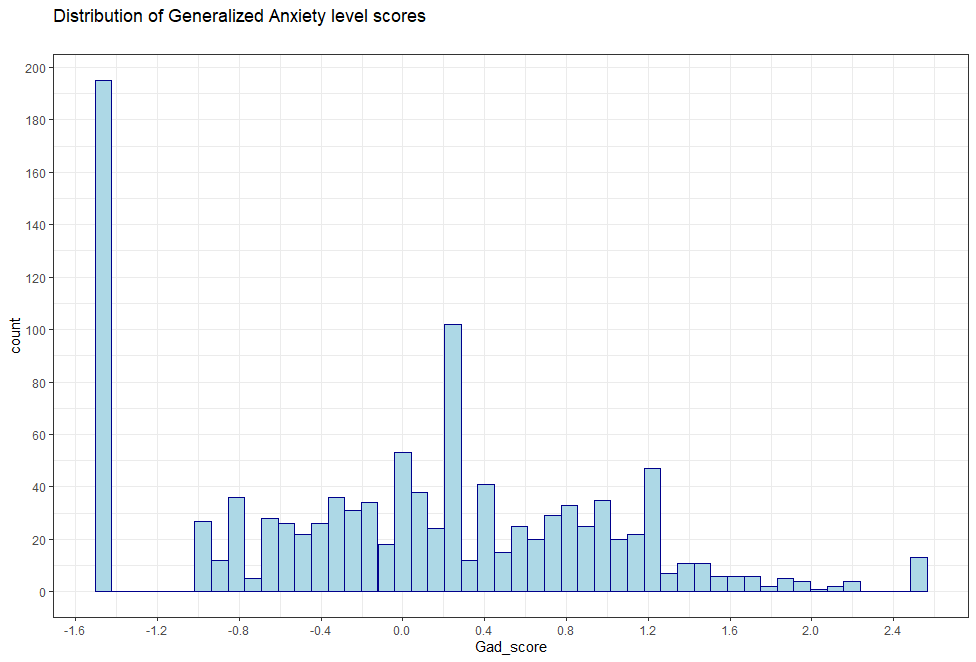
**Interpretation & Conclusions for a Lay Audience**

**Appendix A.** Correlation matrix of continuous variables in the raw dataset

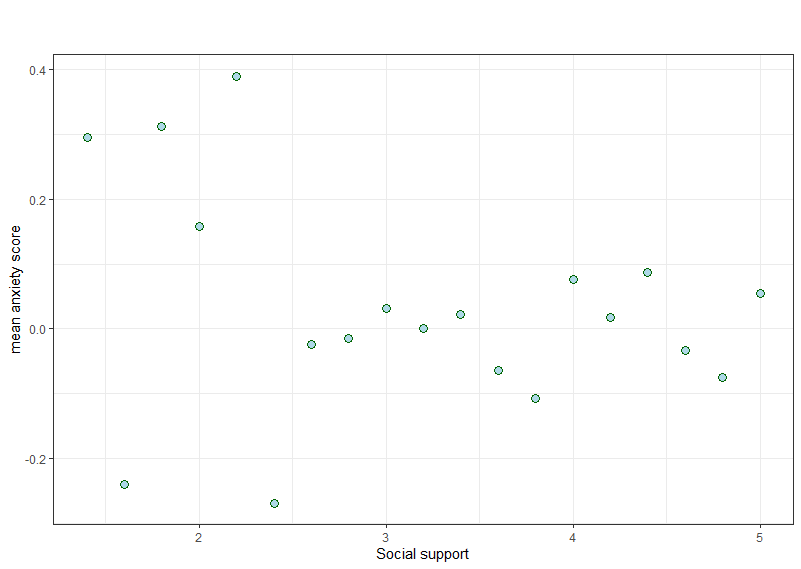
Table

Description automatically generated

**Appendix B.** Distribution histogram of Gad\_score



**Appendix C.** Mean anxiety score level plotted against each level of social support received.



**Appendix D.** Mean anxiety score at different levels of pandemic difficulties

Appendix E. First linear regression model with all variables included

**Appendix F.** Residual Plotsafter feature selection

1. , where is the estimate for the variance of the error term, is the number of observations, and is the number of features. [↑](#footnote-ref-1)