# "I ain't reading all that": Exploration and evaluation of an abbreviated screening tool for depression, anxiety, and stress

Dav King

2023-12-14

## Introduction

Mental disorders are phenomena which are far more prevalent than many realize. Discussions of mental health have long been stigmatized, leading many with mental disorders to falsely believe they are alone - while one study calculated that the proportion of adults who at some point meet the diagnostic criteria for at least one mental disorder is as high as 83% (Schaefer et al., 2017). In combination with the lack of accessible mental healthcare around the world, a far lower percentage, then, ever actually get diagnosed - preventing them from accessing the care that they need and furthering the stigma around mental illness. One useful tool that has grown in popularity in recent years is psychometric diagnostics: generally survey-like questionnaires, administered by clinicians to classify the presence and extent of symptoms. These diagnostics provide a streamlined, standardized tool that can support the case for a diagnosis while requiring few resources to administer.

However, despite their advantages, there are still some shortcomings to psychometric diagnostic tools. A clear issue is the fact that many of these tools are very long, requiring extensive time and energy to fill out. This is a problem across all survey research: it has been shown that increased survey length predicts decreased response rates (Edwards et al., 2004) and may increase missing data (Rammstedt & Beierlein, 2014). This issue has been specifically addressed as a growing problem in psychometric research. Some researchers point to the fact that inter-item correlation tends to be over-emphasized as a criterion for validity in these scales, which lends itself to long, seemingly repetitive scales that may increase attrition among respondents (Stanton et al., 2002). Other researchers have suggested that shorter scales, especially in large-scale implementation settings, may reduce boredom and fatigue and alleviate the potential negative effect of lengthy questionnaires on participants' cognitive and motivational processes (Rammstedt & Beierlein, 2014). However, despite these clear drawbacks to lengthy psychometric scales and clear suggestions for ways to shorten them, psychometric surveys are only growing longer and little research that I can find has been conducted on the topic of shortening existing surveys. The purpose of this study, then, is to begin to explore the feasibility and efficacy of shortening psychometric diagnostics into abbreviated screening tools, with the eventual goal of making such diagnostics less taxing and more accessible around the world.

The specific diagnostic in question for this study are the Depression Anxiety Stress Scales (DASS; Lovibond & Lovibond, 1995). Anxiety and depression are two of the most frequently occurring disorders among US adults, with some estimates placing their prevalence as high as 12.5% and 5%, respectively (National Center for Health Statistics, 2023). These scales are designed to measure three different psychological constructs (depression, anxiety, and stress), but are represented in a single set of 42 Likert-type questions. While an abridged, 21-question version of these scales has been published, I believe that an effective screening tool can be implemented with substantially fewer items. Between the wealth of data provided by these lengthy scales and their relevance as diagnostic tools for highly prevalent disorders, these scales are a phenomenal candidate for exploring the reduction of psychometric scales into short screeners.

The primary goal in this project is to select a subset of variables from these scales that provide strong predictive performance over the scale to form an abbreviated screening tool, and subsequently to define a model over these variables that yields strong predictive accuracy. While any insights gleamed from the data

may be useful, the primary purpose of this project is prediction, with interpretability left as an afterthought. Specifically, the goals of this project are the following:

1) Explore the dataset and identify items with high cross-scale predictive power.
2) Select three questions from each subscale to be used as predictors in the abbreviated screening tool, as well as any other useful demographic predictors.

- 2.1) These questions will be selected based on predictive power over their own subscale, predictive power over the other two subscales, and a general assessment of their difference from one another (with the goal of selecting questions that do not simply rephrase one another).
- 2.2) Demographic and personality variables will also be evaluated as potential predictors.

3) Combine these questions into one short, easy-to-understand screening tool that can be completed in under two minutes and will accurately predict which participants may need further, more extensive testing.
4) Develop a predictive model for each of these three constructs, selected through 10-fold cross-validation performed on a 75% training subset of the data.
5) Evaluate the performance of the selected models on the remaining 25% of the data.
6) Draw conclusions regarding the efficacy of each model in predicting the corresponding DASS construct, and provide recommendations for how this screening tool might and might not be used.

There are a few important caveats in this study. First, as will be mentioned below, there are some concerns with the validity of these data as representative of the population as a whole. Thus, while this study can provide useful insights, it should not be used alone for the validation of an entirely new tool. Second, the computed DASS scales are simply summations of their items, and the variables we are selecting are these same items (which quite literally compose a part of their predicted response variable). While we mitigate this problem somewhat by creating a reduced version of each response that does not include the selected predictors, and showing that these do not suggest a different, better model specification, this is still a real concern with independence to take into account. Third, while the Likert-type participant responses are technically discrete, they are treated here as continuous predictors (though this approach is well-established in the literature). Fourth, no psychometric scale can be validated without empirical research data. While this diagnostic tool could be useful for its intended purpose of recommending certain users for further psychological evaluation, it is *not* intended to be a replacement for full-length diagnostic tools and evaluation by a licensed clinician. The purpose of this tool is to simply and effectively screen users to determine who may need further testing. Fifth, while I have some experience in psychology and generally understand the potential strengths and drawbacks of this approach, I am by no means a licensed, certified professional in this domain. Though I believe that an effective screening tool is developed in this study, it would not be wise to implement it in a practical setting without evaluation by psychology professionals, which goes beyond the scope of this study. However, despite these limitations, there is still much to be gained from the initial explorations of psychometric scale reductions that are evaluated in this study.

# Data

These data were originally sourced from a compiled dataset on Kaggle, which drew open-sourced raw data from Openpsychometrics.org (Greenwell, 2020). These data contain just shy of 40,000 responses to the DASS with zero missing data, including responses to all 42 items on the DASS, the order in which items were presented to each participant, and the time that it took for the participant to respond to each item The data also include responses to the full Ten-Item Personality Inventory (TIPI), which measures participants' Big 5 personality traits (openness, conscientiousness, extraversion, agreeableness, & neuroticism/emotional stability; Gosling et al., 2003), as well as several demographic variables. All of these variables are defined in the file `codebook.txt`; this codebook was compiled for the data on Kaggle and has not been modified.

Upon initial inspection, the data provide immediate motivation for the reduction of these scales. Demonstrated

in Figure 1 below, the median time for participants to respond to each item in the DASS decreases almost monotonically (and initially quite rapidly) as a function of the item's position in the survey. In other words, it took participants much longer to respond to the first couple of items in the survey, and the further they got in the survey the less time they spent on responding to each item. While there are obvious explanations for why response times may have been high for the first couple of items, this trend overall suggests that participants may have A) been able to settle on their responses quite quickly, and thus not need additional redundant questions to convey the constructs of depression, anxiety, and stress, and B) given less attention and consideration to later questions, reducing somewhat their validity. Figure B.1 in the appendix shows that this same pattern holds across all three scales.

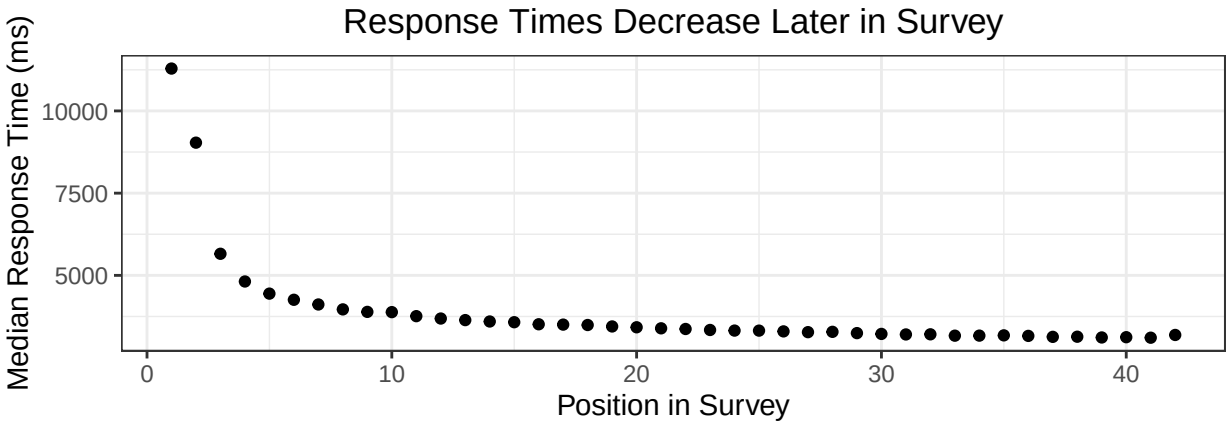



Figure 1

Constructing the depression, anxiety, and stress scales of the DASS is quite straightforward - the relevant questions for each scale (which can be found in the manual) are simply added together. The DASS measures responses from 0-3, but these data were reported on a scale of 1-4; thus, when each 14-item scale had been created, a value of 14 was subtracted from each score to correct for this. These served as the response variables in our modeling. Box-Cox transformations were employed to see whether each variable would perform better under some transformation towards normalcy, and exploratory modeling was performed for all three variables both with and without transformation. The scales for stress and depression did not suggest any meaningful improvement from transformation, but anxiety did and thus it was replaced by its own square root. Distributions of all three (untransformed) variables can be seen in Figure 2, and the transformation of anxiety is shown in appendix Figure B.2.

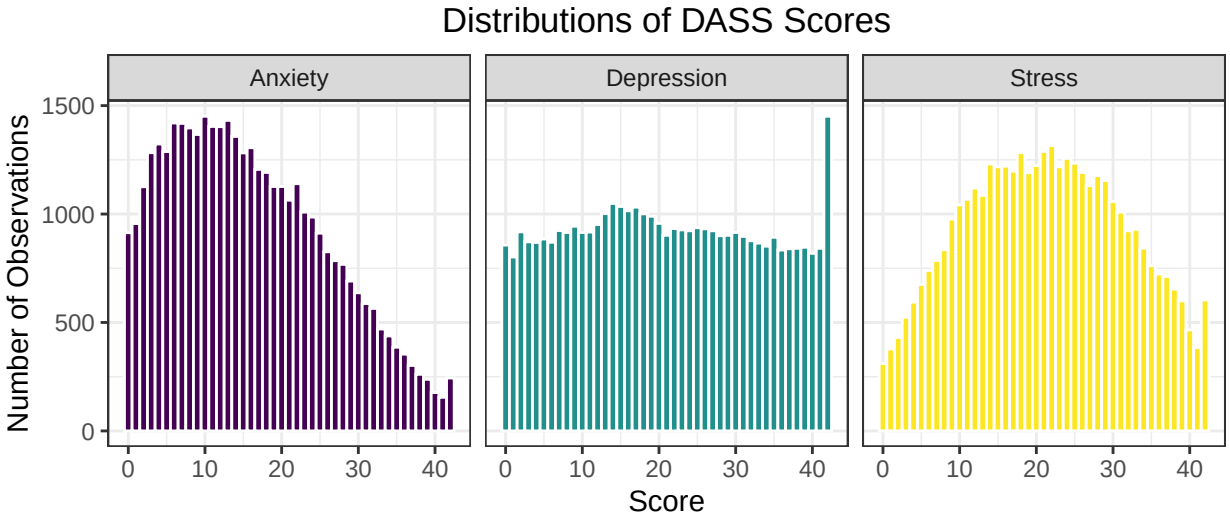



Figure 2

Additionally, using cutoffs that can be found in the DASS manual (and shown in Table 1 below), these variables can be cut into the categories "Normal," "Mild," "Moderate," "Severe," and "Extremely Severe." Although these categories were not used for prediction, they are useful both for data exploration and for model evaluation, and are printed here in Figure 3.

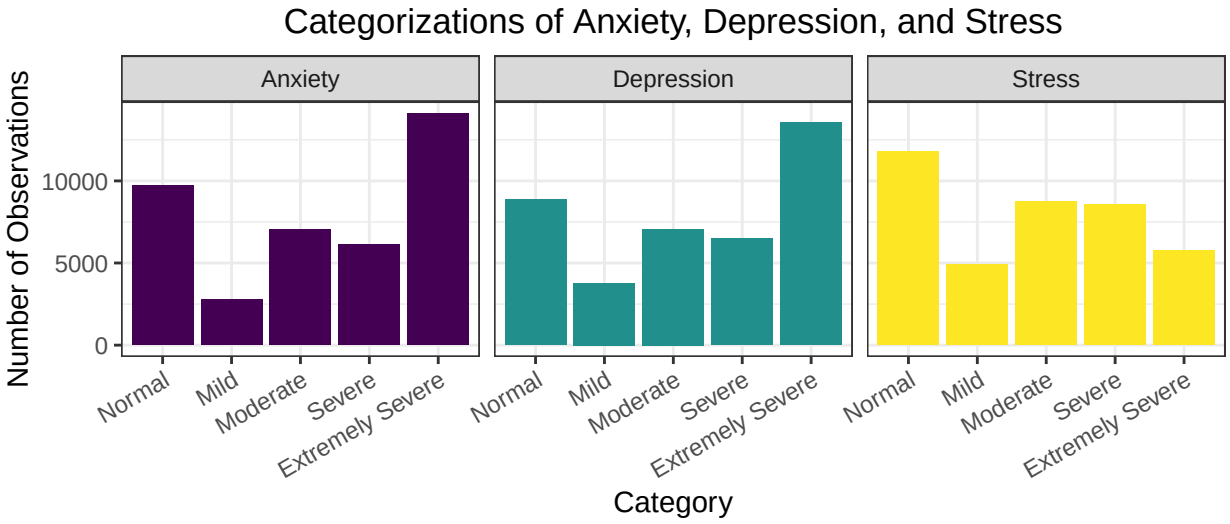



Figure 3

When Openpsychometrics collected these data, they asked participants if they had A) answered truthfully and B) were willing to have their data collected for research purposes, and only collected the data of participants who answered yes to both. However, Figures 2 and 3 show concerning trends that lead to suspicion over the accuracy of these data. It is worth noting that these data were collected via online convenience sample - participants had to seek out the Openpsychometrics website in order to fill out this survey, and thus there is a clear selection bias towards participants who may have higher symptoms of anxiety, depression, and stress in general. The first point of concern is in Figure 2, where all three scales have a strong uptick for the maximum possible score (i.e., participants who went through and selected "applied to me very much" for every single question). These are not necessarily valid data, and this speaks to the importance of reverse-scored items in any psychometric scale. However, these data were not removed from this dataset, as it is impossible to know how many of these responses were truly genuine. The second point of concern is in Figure 3. It is clear that very high proportions of these participants (indirectly) classified their symptoms as severe and extremely severe - far more than would be expected of the general population. However, while this may speak to the need for these models to be evaluated on a more accurate population, this is actually an asset of this dataset. By oversampling participants in categories of high severity, it is much easier to build a model that can predict for these extremes, rather than facing the usual difficulty in prediction when the population at risk is much smaller in number than the population not at risk.

The scores of these scales all have high correlation with one another (Stress-Anxiety = .802, Stress-Depression = .740, Anxiety-Depression = .670; see Figure B.3). Additionally, a principal component analysis was conducted on these DASS questions, under the assumption that three subscales should yield three meaningful principal components. However (as seen in Table 2 and Figure B.4), one principal component explained about 45.2% of the variance in the DASS. Taken together, these suggest that either the DASS subscales do not accurately capture their construct and their construct alone, or (more likely) that anxiety, depression, and stress are not entirely separable from one another. However, it is still possible to differentiate them somewhat - the second principal component (explaining about 6.8% of the variance) distinguishes anxiety and stress from depression, while the third (explaining about 3.9%) teases apart anxiety and stress. No other principal component explained more than 3% of the variance.

Demographic variables turned out to be very unimportant as predictors in this dataset, and so they are not discussed here beyond mentioning that i) the population generally skewed very young (i.e., people in their 20s) and ii) the majority of these data came from Malaysia. The final variable considered is the TIPI. Generally,

the TIPI results of these data aligned with those from other research - roughly normally distributed, and centered around the midpoint of the scale. There were some missing data on the TIPI. Thus, on item TIPI4 (the only variable from TIPI used as a predictor), missing data were imputed as a score of 4 rather than deleted from the dataset. This was done for a few reasons: 1) these made up a very small subset of the observations, and this allows us to not delete potentially useful observations from the dataset; 2) TIPI4 was already generally the weakest of the predictors, and 3) a score of 4 corresponds to the value "neither agree nor disagree," and this seems like a reasonable value to impute for missing data (in fact, there is an argument to be made that this is an equivalent response).

# Methodology

## Variable Selection

The first, and more important, aspect of this project was to select the variables for the abbreviated diagnostic tool. Variable selection was performed under several constraining criteria. First, in order to meet the goal of keeping this diagnostic tool under two minutes, I decided to pull the most predictive three items from each of the three scales. This would be paired with whichever demographic and personality variables were relevant and easy to collect. Second, since we were reducing these scales down to just three of their items, an important selection criterion was to select items that were strong predictors of the other two scales as well. Third, once a set of realistic candidate variables had been identified, care was taken to ensure that the items were substantially different, and not simply restatements of one another.

In order to determine which variables were the strongest predictors, several different models were run. Each model was trained on the 75% of the data randomly selected for training, and predicted the transformed and untransformed depression, anxiety, and stress scores. The highest weight in variable selection was placed on the findings from random forest models (i.e., decision trees grown on resampled data using a random subset of the predictors) and bagging models (i.e., random forest models using all 63 predictor variables) because these ensemble learning models provide more than just one view of the data and thus give a clearer image of variable importance. These expensive models were run for each response variable, using $m = 8$ predictors in the random forests, and their variable importance plots are shown in Figures C.1-3 below. However, these forest models were not alone. Linear regression models, shrinkage models (ridge & lasso; models designed to bias coefficients towards 0), and boosting models (trees grown sequentially on the residuals of the previous) were also calculated. In these regression models, larger coefficients (in absolute value, though almost all were positive) meant a more important variable. Different types of predictions were run for each scale: prediction from its own items, prediction from the items of other scales, prediction from the TIPI, and prediction from demographic variables.

This modeling process revealed several important findings. First, there was minimal evidence that any demographic variables were meaningful predictors. This finding supports the validity of the DASS, as it suggests that the scale's predictions are not largely affected by cultural, racial, national, or physical characteristics. Second, none of the TIPI items were significant, except for the two (TIPI4 and TIPI9) making up the neuroticism/emotional stability trait. This is consistent with previous findings that neuroticism is the only Big 5 personality trait which predicts mental health (e.g., Gale et al., 2016). As this was the only variable that consistently predicted all three scales outside of the DASS questions, TIPI4 was selected as a predictor for this diagnostic tool. Whether a single item from the TIPI can meaningfully be administered alone is a valid concern, but it is beyond the scope of this study.

Based on the findings from these models, items from the DASS were selected based on their predictive strength within both their own subscale and the other two subscales, as well as their ability to capture slightly different constructs (based on my own best judgement). The final variables selected were:

- Stress: `Q11A`, `Q27A`, `Q29A`
- Anxiety: `Q9A`, `Q28A`, `Q40A`
- Depression: `Q10A`, `Q13A`, `Q21A`

These variables, along with TIPI4, comprised the ten predictors chosen for this new diagnostic tool. To see

what that tool might look like, with written-out specifications of each of these items, see Appendix A. Adding up the response times for these questions (excluding outliers), we find that it in general takes participants less than 45 seconds to answer all nine questions. Even when accounting for the extra time to respond to TIPI4 and to read two sets of directions, this still seems highly likely to keep us under our goal of two minutes to complete the entire questionnaire.

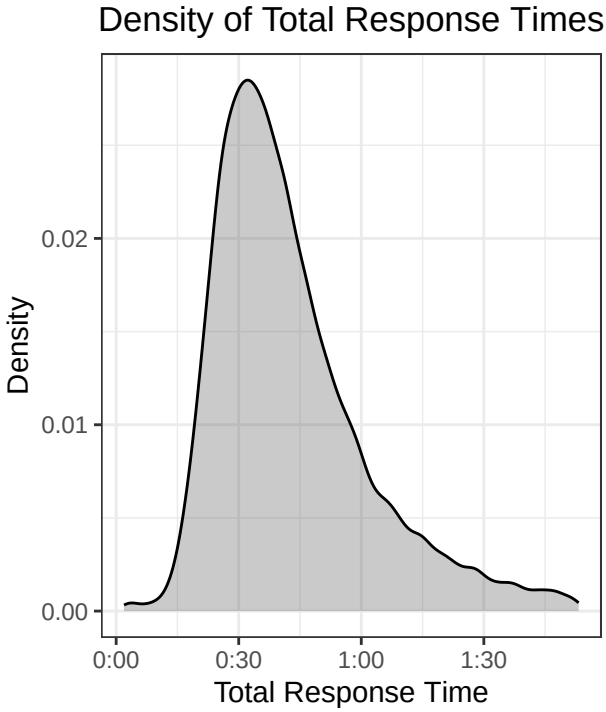



Figure 4.1

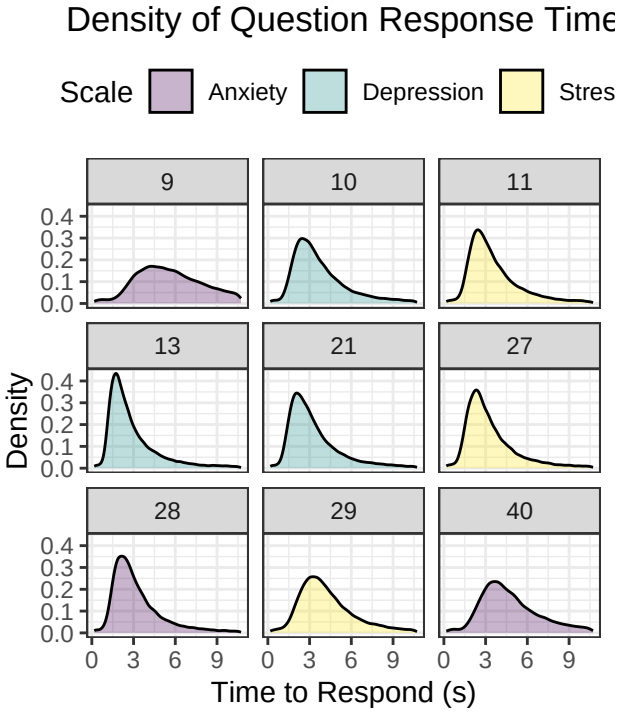



Figure 4.2

These predictors are all relatively uncorrelated with each other (considering their intention to capture the same constructs), with correlations ranging from .698 (Q21-Q10) to .312 (TIPI4-Q10; see Figure C.4).

## Model Selection

With our variables selected, we now turn to the second part of this project: predictive modeling. The goal of this model is prediction alone; thus, interpretability is not a major concern. This allowed more complicated models, including nonlinear and interaction terms, to be considered. One of the biggest concerns with nonlinear predictors (especially polynomials) is end behavior - that is, predictions for data points that fall outside (or at the extremes) of the training data. However, in this case, our scales are restricted to integers from 0-3 for DASS questions and integers 1-7 for TIPI. No inputs can ever be given beyond these values; thus, we do not need to be too concerned about constraining the end behavior of our models, and can utilize more flexible predictors if the data support this decision.

The goal of this modeling was to select three different models: one each for predicting scores for stress, anxiety, and depression. Although models for classification into categories of "high" and "low" risk based on some arbitrary cutoff point were considered, these classification models did not significantly outperform the equivalent predictions of a regression model with a designated cutoff, and they provide substantially less information than a regression model does. Thus, they were not retained for final consideration.

To evaluate and compare models, I computed mean squared error through 10-fold cross-validation to estimate the testing error of each model specification. Mean squared error was used because it is a direct calculation of testing error (which we are trying to minimize), and because it can be calculated without any respect to the parameters of the model, making it a more flexible and adaptable measure. This 10-fold cross-validation

was conducted across the 75% training subset, so that the selected models could then be evaluated on the remaining, unseen test set.

For each scale, several different types of models were considered. These specifications included (with various interactions and nonlinear predictors): linear regression models, shrinkage (ridge & lasso) models, generalized additive models focused on polynomial terms, bagging & random forest models, and non-parametric $k$-nearest neighbors regression models. Because of computational complexity issues, the bagging & random forest models did not undergo 10-fold CV; instead, their out-of-bag error estimates were computed, and they were compared to the best-performing models from the CV approach on their test set predictive performance. In all cases, these forest models performed substantially worse than other regression models; thus, this use of reserved data was not an issue.

For each scale, models were also considered on the "reduced" version of the scales, that is, the sum of all items in each DASS scale excluding the three used as predictors in this abbreviated tool. This was done to ensure that the problem of dependent variables predicting a response variable that they partially compose did not impair model selection. However, the correlations between the full and reduced versions of all three scales were higher than .98, and in no cases did the MSE calculated on the reduced scale favor a different model than the MSE calculated on the full scale. Thus, these models will not be discussed further.

For all three scales, some form of generalized additive model was selected, with polynomial expressions of some DASS items and either polynomial or cubic spline expressions of TIPI4 used in the model specification. In R, the `poly()` function generates orthogonal polynomials from the dataset to use in prediction. While this is useful for a number of reasons, it does mean that these models cannot be meaningfully expressed in a written equation. Additionally, these are very lengthy models, and would be tedious to write here. Thus, for each of the following selected models, we list the specification here but will leave the details, with full coefficients, error, and probabilities, as well as a written out "equation," listed in Appendix D. Similarly, while model diagnostics are addressed here, the plots are left for the appendix. For future use, the models for stress, anxiety, and depression are stored here, for convenience, in `Stress Regression Model.RData`, `Anxiety Regression Model.RData`, and `Depression Regression Model.RData`, respectively.

**Stress**

$$\begin{aligned}\text{stress score} \sim \text{Q11A} + \text{Q27A} + \text{poly(Q29A, 2)} + \text{poly(Q9A, 3)} + \text{poly(Q28A, 3)} + \text{poly(Q40A, 2)} \\ + \text{Q10A} + \text{Q13A} + \text{poly(Q21A, 2)} + \text{poly(TIPI4, 3)}\end{aligned}$$

This model includes specifications of several predictors in terms of polynomials, and predicts TIPI4 with a degree-3 polynomial as well. In this model, the only non-significant variables are the second-order effects of Q9A and Q28A, both of which have significant third-order effects. This model is overall significant, $F(19, 29811) = 10680$, $p < .001$. It has an $R^2$ value of 0.8719, meaning that this model can explain about 87.19% of the variance in the training `stress_score`, and its adjusted $R^2$ value is the same, which suggests that none of these variables are particularly redundant or meaningless because they all provide some additional information to the prediction. This same specification (with different coefficients) also explains about 78.84% of the variance in the reduced scale, which is still very high predictive power.

This model generally meets the required assumptions for regression. There are slight differences in the residuals based on fitted values, but these differences are very slim. The Q-Q plot looks very good, helping us meet the normalcy assumption of regression. While the semi-discrete, Likert-type predictors create some weird patterns in the scale-location plot, it is generally not concerning; further, there are no data points with outsized leverage skewing this model. Thus, the model overall seems largely sound.

**Anxiety**

$$\begin{aligned}\text{anxiety trans} \sim \text{poly(Q11A, 2)} + \text{poly(Q27A, 2)} + \text{poly(Q29A, 2)} + \text{poly(Q9A, 3)} + \text{poly(Q28A, 3)} \\ + \text{poly(Q40A, 2)} + \text{poly(Q10A, 3)} + \text{poly(Q13A, 3)} + \text{poly(Q21A, 3)} + \text{bs(TIPI4, df} = 5)\end{aligned}$$

This model includes a polynomial specification over every single DASS item, as well as a cubic spline with five degrees of freedom for TIPI4. This model is somewhat more sparse in terms of significant variables, but it is still overall significant, $F(28, 29802) = 4356$, $p < .001$. It has an $R^2$ of 0.8037, meaning that this model can explain 80.37% of the variance in `anxiety_trans`, and its adjusted $R^2$ of 0.8035 suggests that, despite some lack of significance, each predictor generally provides some additional information. As a reminder, this model is specified for `anxiety_trans`, or the square root of `anxiety_score`. This specification also explains about 66.54% of the variance in the reduced scale - not as strong of a performance, but still a lot of explanation nonetheless.

The assumptions of this model largely mirror those of the stress model. While the square-root transformation causes slightly more messiness in the residuals and the scale=location of this plot, these differences are generally not that large, and this model, too, seems to be generally sound.

**Depression**

$$\text{depression score} \sim \text{Q11A} + \text{Q27A} + \text{Q29A} + \text{poly(Q9A, 3)} + \text{Q28A} + \text{poly(Q40A, 3)} + \text{Q10A} + \text{Q13A} + \text{poly(Q21A, 3)} + \text{bs(TIPI4, df} = 5)$$

This model is, once again, defined with polynomials over around half of the DASS items, and a cubic spline for TIPI4. Almost all variables are significant; however, TIPI4 is mostly not - across all models tested, TIPI4 seems to be much less predictive of depression than it is of anxiety and stress. This model is actually more well-defined than either that of stress or anxiety. It is overall significant, $F(20, 29810) = 14150$, $p < .001$, and its $R^2$ (with identical adjusted $R^2$) of 0.9047 suggests that over nine-tenths of the variance in `depression_score` can be explained by this model from just 10 predictors. This model also explains about 84.17% of the variance in the reduced scale - a very high amount, and a finding that means this model is better at predicting a scale it does not make up than the model for anxiety is at predicting a scale from *some of the variables that make it up*. While I was not expecting depression to be the easiest to model given that anxiety and stress seem more directly related, this is a good finding to have, and likely speaks to the accuracy of the DASS in capturing depression.

This model, once again, has similar diagnostics to the other two. There is a little bit of trend in the residuals at different fitted values, but not enough to truly violate the assumption of constant variance; the Q-Q plot is also a little messier, but not so much as to violate the assumption of normalcy. This model appears to be, once again, quite sound overall.

# Results

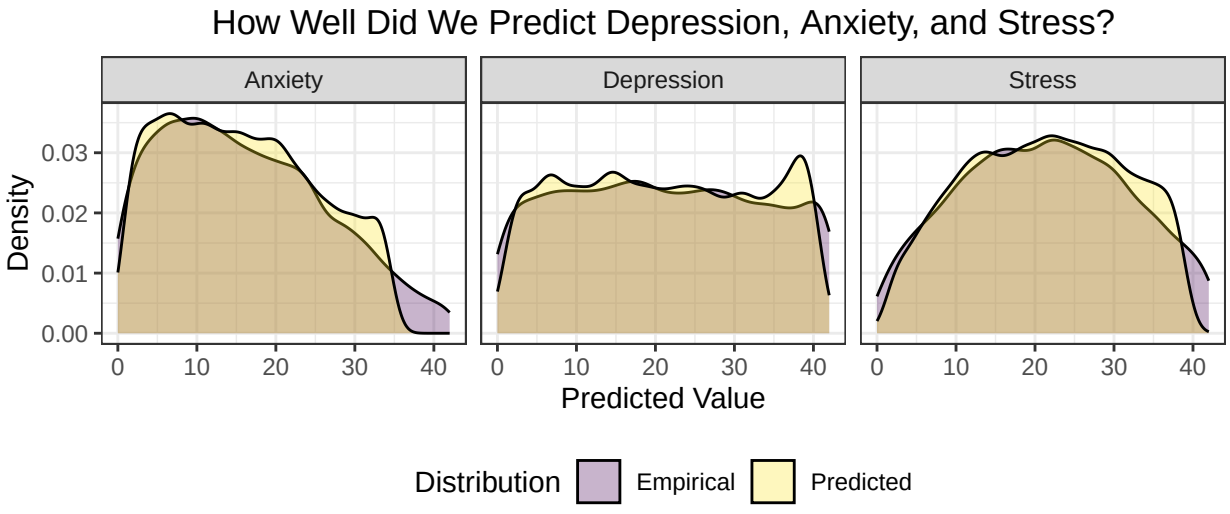



Figure 5

Overall, each model performed quite well over the testing data. The models for stress, anxiety, and depression had test MSE values of 14.339, 21.634 (0.389 non-adjusted), and 14.310, respectively, with comparable test MSE over the reduced scales. What this means, if we take square roots, is that on average this model mispredicts stress, anxiety, and depression by 3.787, 4.651, and 3.723, respectively. While these may seem like large values, in the grand scheme of things they are not. We knew going into this process that this tool would not be able to exactly emulate the true outcome variables, nor was the goal to do so. Instead, our goal is to generally approximate and screen users for risk of depression, anxiety, and stress, and make recommendations for further testing. In this instance, these error rates are generally small, and we can take them into account when assessing our prediction for a given user and any thresholds we may later define.

Looking at the densities of each scale's predicted values over the test data versus the empirical, true distribution in Figure 5, we see that we have very closely approximated the true distributions. Across all three variables, the biggest issue is underpredicting very high values - remember from our exploratory data analysis that this survey appears to have an overrepresentation of these values to begin with. It looks as though our predictions are slightly more "centered" than are the empirical distributions, but only by a miniscule amount. The only major issue is underpredicting especially high and low response values; otherwise, these approximations seem quite true to the empirical distributions.

Looking at Figure 6 below, we can see that we also categorized these variables with very high accuracy compared to the empirical distribution. In fact, across all three scales we generally under-predicted "Normal" and "Mild" while over-predicting the more severe categories - and this is generally something that we would like to do with such a model. In the development of a screening tool for mental disorders, it is far more costly to falsely label someone as "low risk" when they indeed need further testing than it is to falsely label someone as "high risk" when they do not need to be labeled as such. Overall, then, it appears that these models are serving quite well at our purpose of screening patients for the possibility of depression, anxiety, and stress, and sending those at high risk to undertake further, more extensive testing.

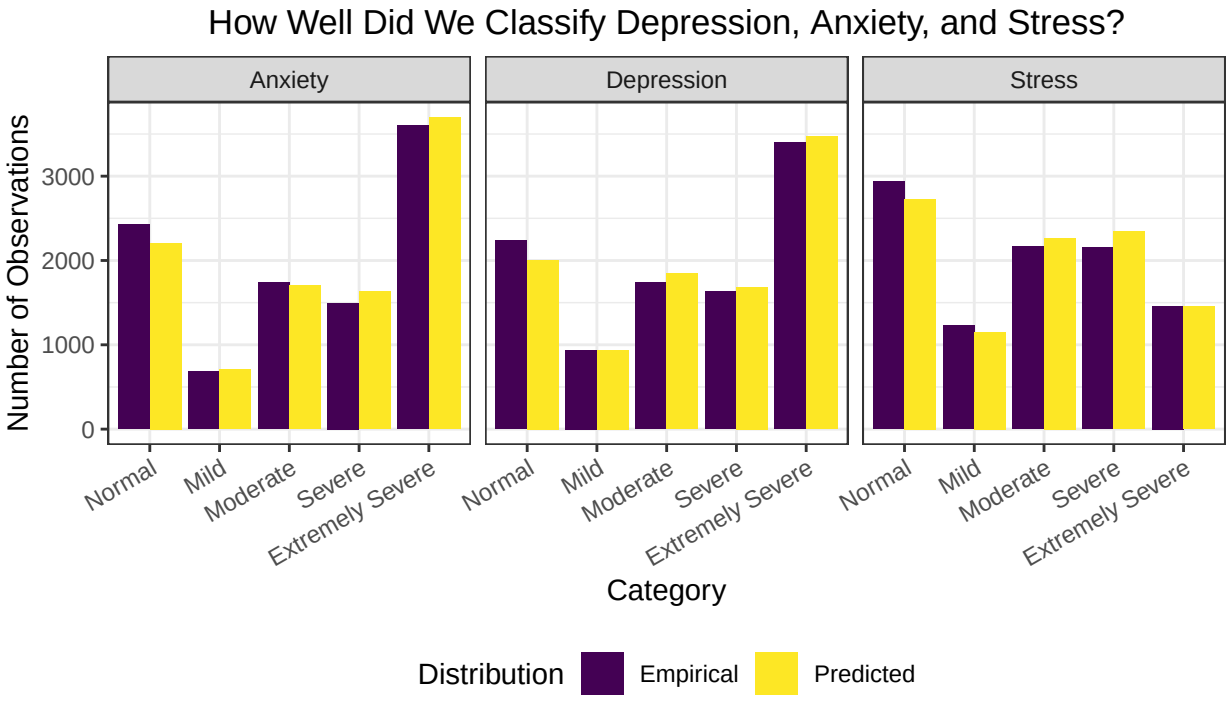



Figure 6

Some checks for robustness were also performed, though they are not listed here. In general, across all three scales, residuals did not differ much by country, age, race, gender, or education levels. This suggests a high level of cross-cultural validity among these models - generally speaking, the decisions to leave out demographic variables seem justified, and these models seem robust against these characteristic interpersonal differences.

## Conclusion

In this brief study, we have taken nine variables from the DASS and one from the TIPI, and combined them to develop a potential screening tool for depression, anxiety, and stress with a reasonable level of accuracy. Depression, anxiety, and stress all seem highly related to one another, and the DASS can predominately be explained in terms of a single principal component. However, these scales are different enough from one another that three different model specifications are defined, and our ability to predict depression is much higher than our ability to predict anxiety and somewhat higher than our ability to predict stress. Consistent with prior findings in the literature, neuroticism is predictive of mental disorders while the other four Big 5 personality traits are not. Participants put a lot less time (and presumably thought) into questions later in the DASS, but this tool, which should be able to be filled out in under two minutes (with true estimates of the time to complete the DASS portion as low as 45 seconds), will hopefully alleviate that problem. Overall, this tool is able to very closely approximate the true scales of the DASS, explaining most of their variance without mispredicting responses by too wide of a margin. One important takeaway is this: if we are able to predict a 14-question scale from only 3 of its questions and seven other (theoretically unrelated) items to a high degree of accuracy, the potential for models that dramatically shorten existing psychometric tools is extremely salient.

There are some things that this tool cannot do. For one, it can do nothing at all as of now: a tool like this should never be implemented without first validation in a controlled setting. There are some reasons to doubt in the validity of these data; thus, these findings should be taken with a grain of salt. For another, I am a statistician with a penchant for psychology but not a psychometrician; my own (statistically-aided) judgement on which variables to include in this tool is not necessarily the best opinion that an expert can give. Third, and perhaps most importantly, though I describe this tool at times as a "diagnostic" it indeed cannot diagnose. Diagnoses are not to be made by any psychometric scale without the aid of a licensed clinician, and this test is not designed to diagnose at all - merely to recommend some users for further testing.

However, there are some things that (if confirmed as valid) it would be able to do quite well. This tool provides a very short, low-effort, accessible screening tool that could be implemented in disadvantaged communities all around the world. With help from a psychometrician, a theoretical cutoff value could be set, above which users should be recommended for further testing, and this would dramatically increase the number of people getting the diagnoses that they need.

This tool also paves the way for similar research to follow. In this study, we have only scratched the surface of the potential for future research in this field. Even with this set of DASS responses, more could be explored. While response times for each question were not considered as a predictor, it is entirely possible that spending more time contemplating a question may be just as important of a predictor as the actual response to that question, especially if participants are unwilling to be truthful. Additionally, while the models seemed not to differ much by demographics, early results highlighted a few such domains to explore - gender for stress, race for anxiety, education for depression. DASS modeling would also benefit from being run on other, perhaps more valid data, and this tool might be validated with empirical testing on a controlled population. But we are far from limited to just the DASS. Any number of other diagnoses, including learning disabilities, could benefit heavily from the exact type of short screening tools that this approach may be able to develop. This tool, and others like it to follow, could revolutionize psychology by making psychometric testing and psychological diagnosis more accessible than ever before. This study is just the start - but if these findings are any indication at all, there is almost unfathomable untapped potential to explore in the domain of psychological statistics.

# References


Edwards, P., Roberts, I., Sandercock, P., & Frost, C. (2004). Follow-up by mail in clinical trials: Does questionnaire length matter? *Controlled Clinical Trials, 25*, 31-52. 10.1016/j.cct.2003.08.013

Gale, C.R., Hagenaars, S.P., Davies, G., Hill, W.D., Liewald, D.C.M., Cullen, B., Penninx, B.W., International Consortium for Blood Pressure GWAS, CHARGE Consortium Aging and Longevity Group, Boomsma, D.I., Pell, J., McIntosh, A.M., Smith, D.J., Dreary, I.J., & Harris, S.E. (2016). Pleiotropy between neuroticism and physical and mental health: Findings from 108,038 men and women in UK Biobank. *Translational Psychiatry, 6*(e791). 10.1038/tp.2016.56

Gosling, S.D., Rentfrow, P.J., & Swann, W.B., Jr. (2003). A very brief measure of the Big Five personality domains. *Journal of Research in Personality, 37*, 504-528.10.1016/S0092-6566(03)00046-1

Greenwell, L. (2020, June 3). *Depression Anxiety Stress Scale responses.* Kaggle. https://www.kaggle.com/datasets/lucasgreenwell/depression-anxiety-stress-scales-responses/data

Lovibond, S.H., & Lovibond, P.F. (1995). *Manual for the Depression Anxiety Stress Scales.* (2nd. Ed.) Psychology Foundation. www.psy.unsw.edu.au/dass/

National Center for Health Statistics. (2023, September 19). *Mental health.* Centers for Disease Control and Prevention. https://www.cdc.gov/nchs/fastats/mental-health.htm

Rammstedt, B., & Beierlein, C. (2014). Can't we make it any shorter? The limits of personality assessment and ways to overcome them. *Journal of Individual Differences, 35*(4), 212-220. 10.1027/1614-0001/a000141

Schaefer, J.D., Caspi, A., Belsky, D.W., Harrington, H., Houts, R., Horwood, L.J., Hussong, A., Ramrakha, S., Poulton, R., & Moffitt, T.E. (2017). Enduring mental health: Prevalence and prediction. *Journal of Abnormal Psychology, 126*(2), 212-224. 10.1037/abn0000232

Stanton, J.M., Sinar, E.F., Balzer, W.K., & Smith, P.C. (2002). Issues and strategies for reducing the length of self-report scales. *Personnel Psychology, 55*(1), 167-194. 10.1111/j.1744-6570.2002.tb00108.x

# Appendix A: View of Diagnostic Questionnaire

Please read each statement and circle (select) a number 0, 1, 2, or 3 which indicates how much the statement applied to you *over the past week*. There are no right or wrong answers. Do not spend too much time on any statement.

*The rating scale is as follows:*

0 Did not apply to me at all

1 Applied to me to some degree, or some of the time

2 Applied to me to a considerable degree, or a good part of the time

3 Applied to me very much, or most of the time

---

1) I found myself getting upset rather easily.
2) I found that I was very irritable.
3) I found it hard to calm down after something upset me.
4) I found myself in situations that made me so anxious I was most relieved when they ended.
5) I felt I was close to panic.
6) I was worried about situations in which I might panic and make a fool of myself.
7) I felt that I had nothing to look forward to.
8) I felt sad and depressed.
9) I felt that life wasn't worthwhile.

---

The following personality traits may or may not apply to you. Please write a number next to the statement to indicate the extent to which you agree or disagree with that statement. You should rate the extent to which the pair of traits applies to you, even if one characteristic applies more strongly than the other.

1 = Disagree strongly

2 = Disagree moderately

3 = Disagree a little

4 = Neither agree nor disagree

5 = Agree a little

6 = Agree moderately

7 = Agree strongly

---

I see myself as:

___ Anxious, easily upset.

# Appendix B: Data Visualizations

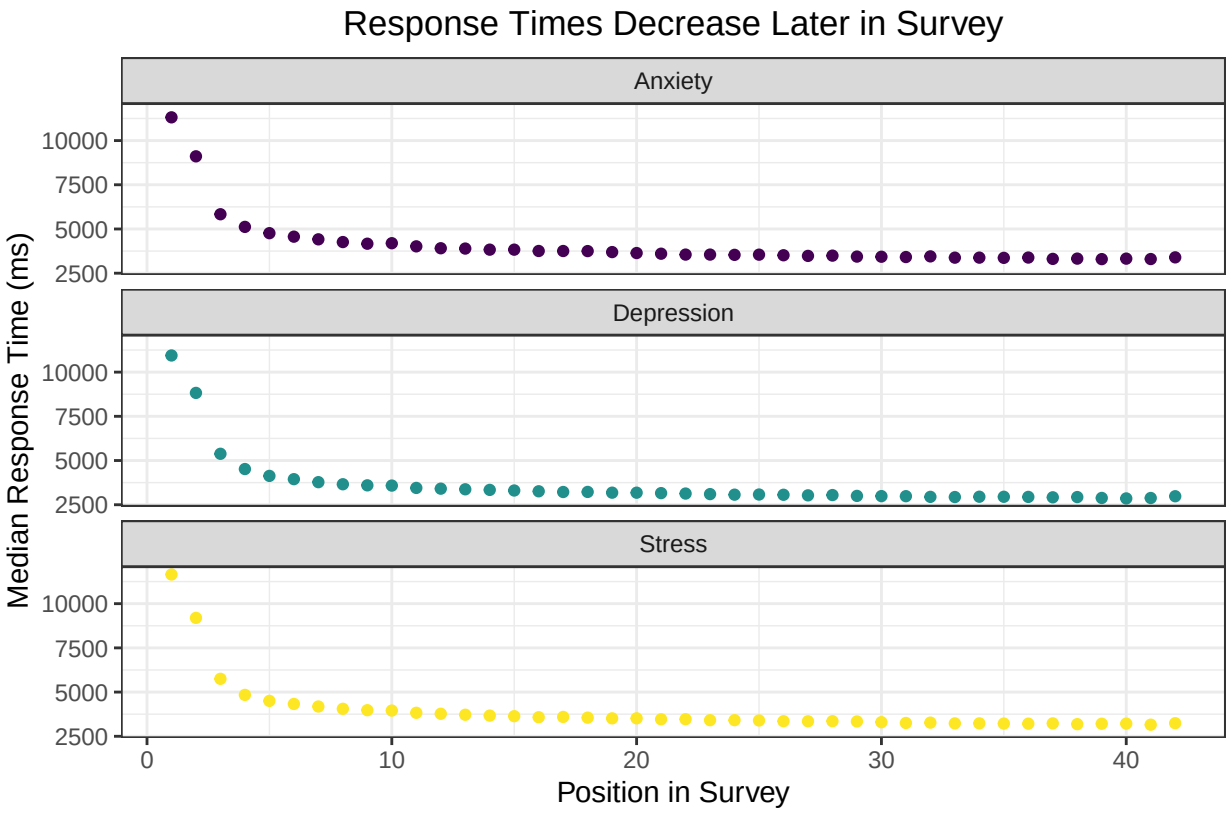



Figure B.1

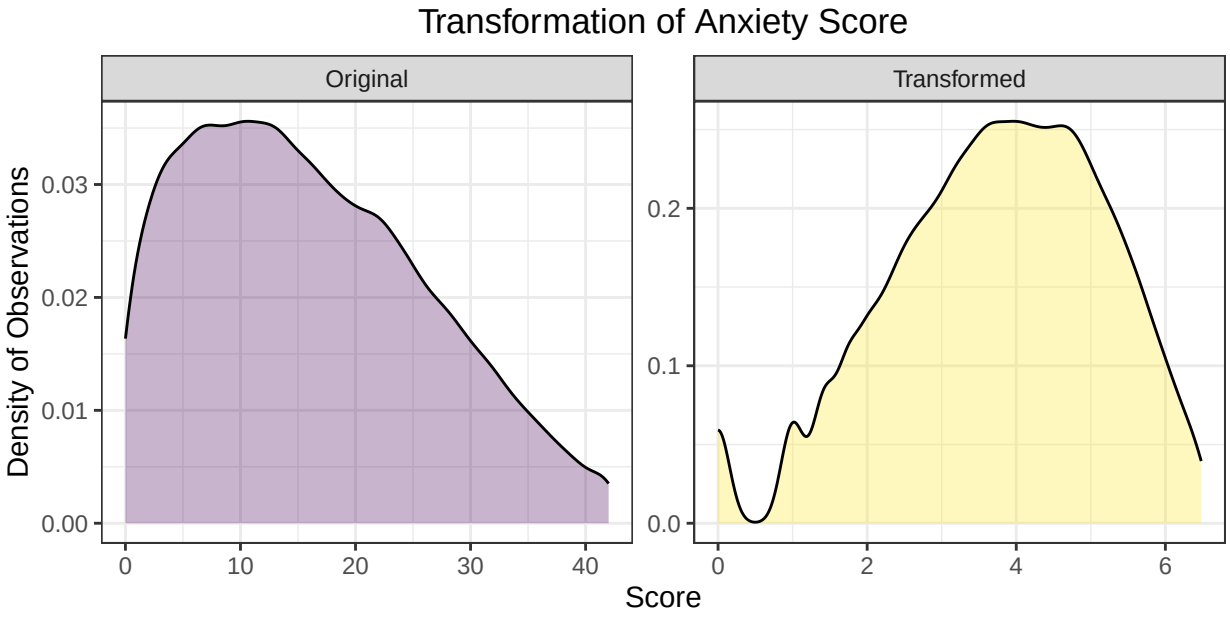



Figure B.2

Table 1: Cutoff Values

| Category | Stress | Anxiety | Depression |
|---|---|---|---|
| Normal | 0-14 | 0-7 | 0-9 |
| Mild | 15-18 | 8-9 | 10-13 |
| Moderate | 19-25 | 10-14 | 14-20 |
| Severe | 26-33 | 15-19 | 21-27 |
| Extremely Severe | 34+ | 20+ | 28+ |

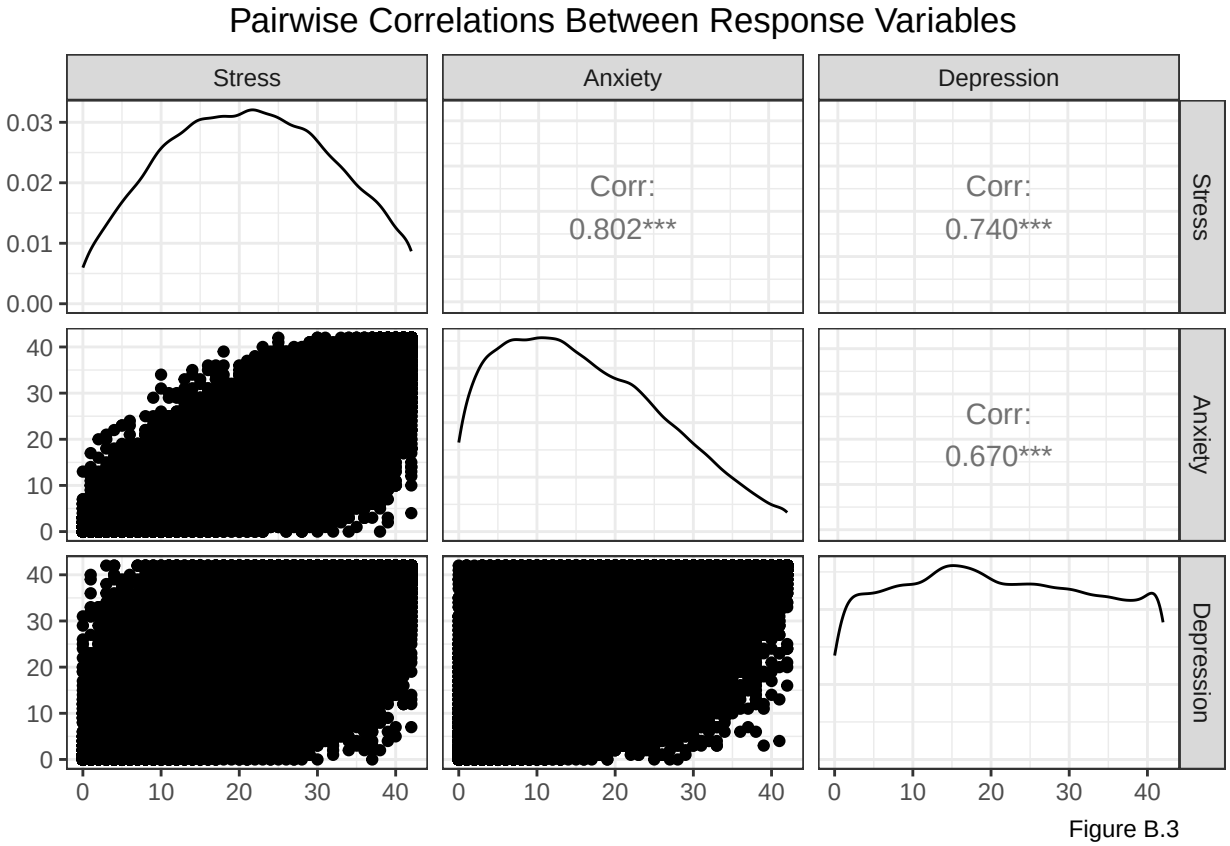


Figure B.3

Table 2: Principal Component Variance

| Component | Variance |
|---|---|
| 1 | 0.4515 |
| 2 | 0.0681 |
| 3 | 0.0391 |

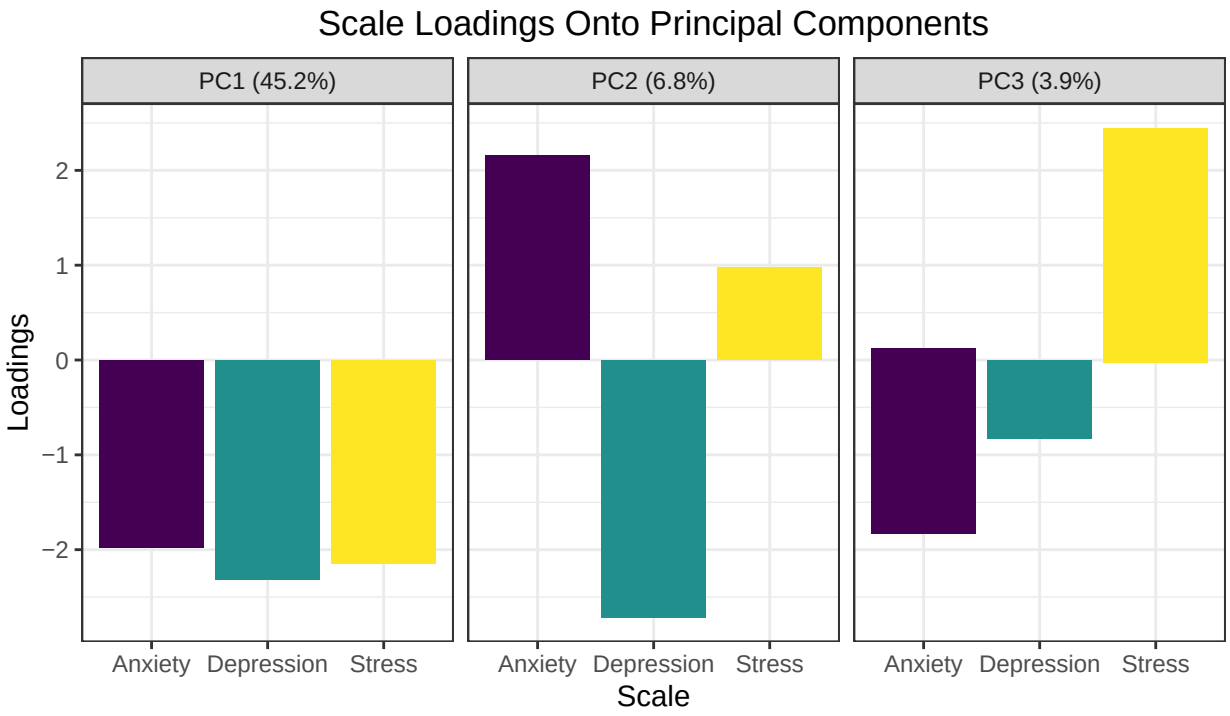

Scale Loadings Onto Principal Components
PC1 (45.2%)
PC2 (6.8%)
PC3 (3.9%)
Loadings
2
1
0
−1
−2
Anxiety
Depression
Stress
Scale

Figure B.4

## Appendix C: Variable Selection

Figure C.1: Stress Random Forest Variable Importance

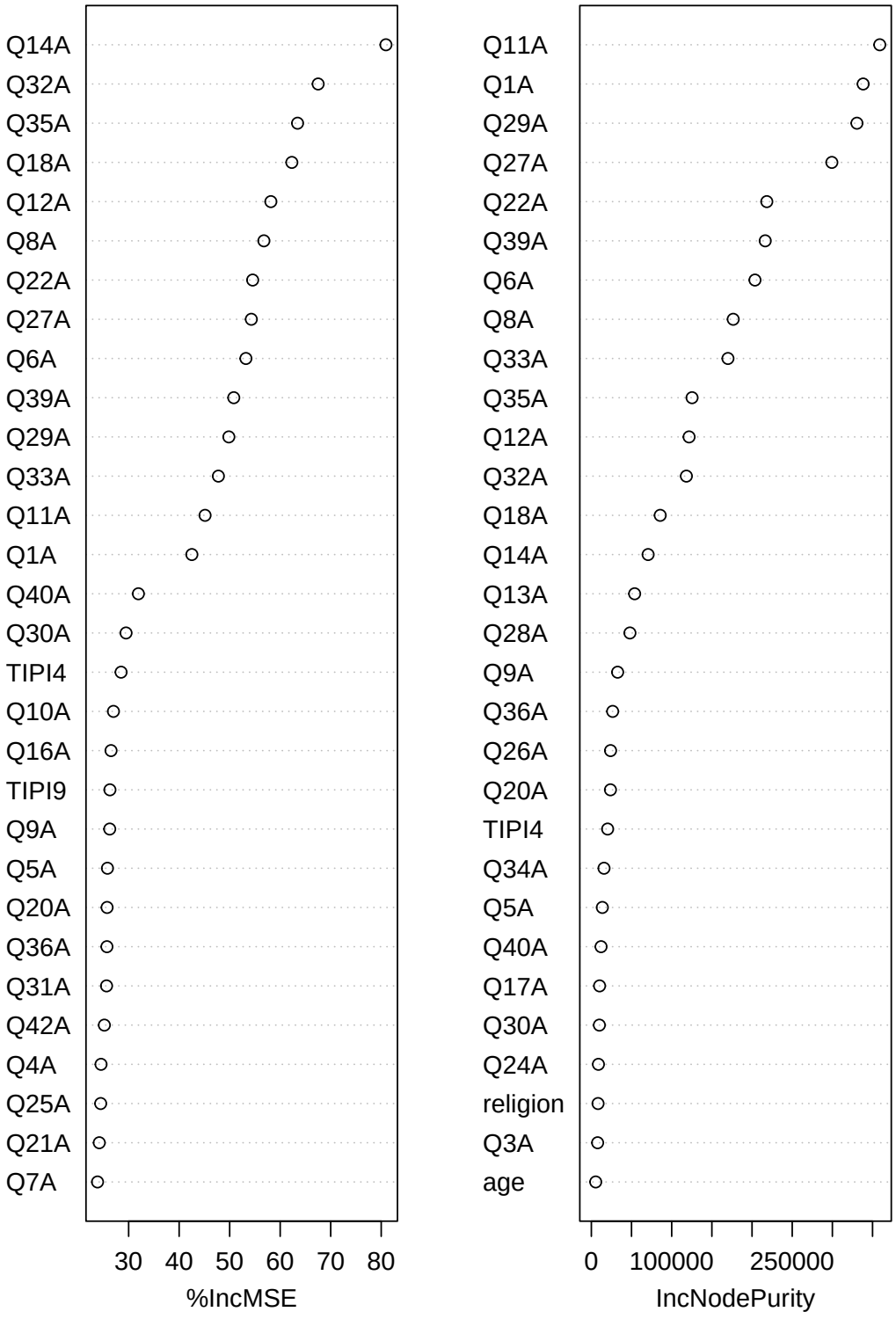


Figure C.2: Anxiety Random Forest Variable Importance

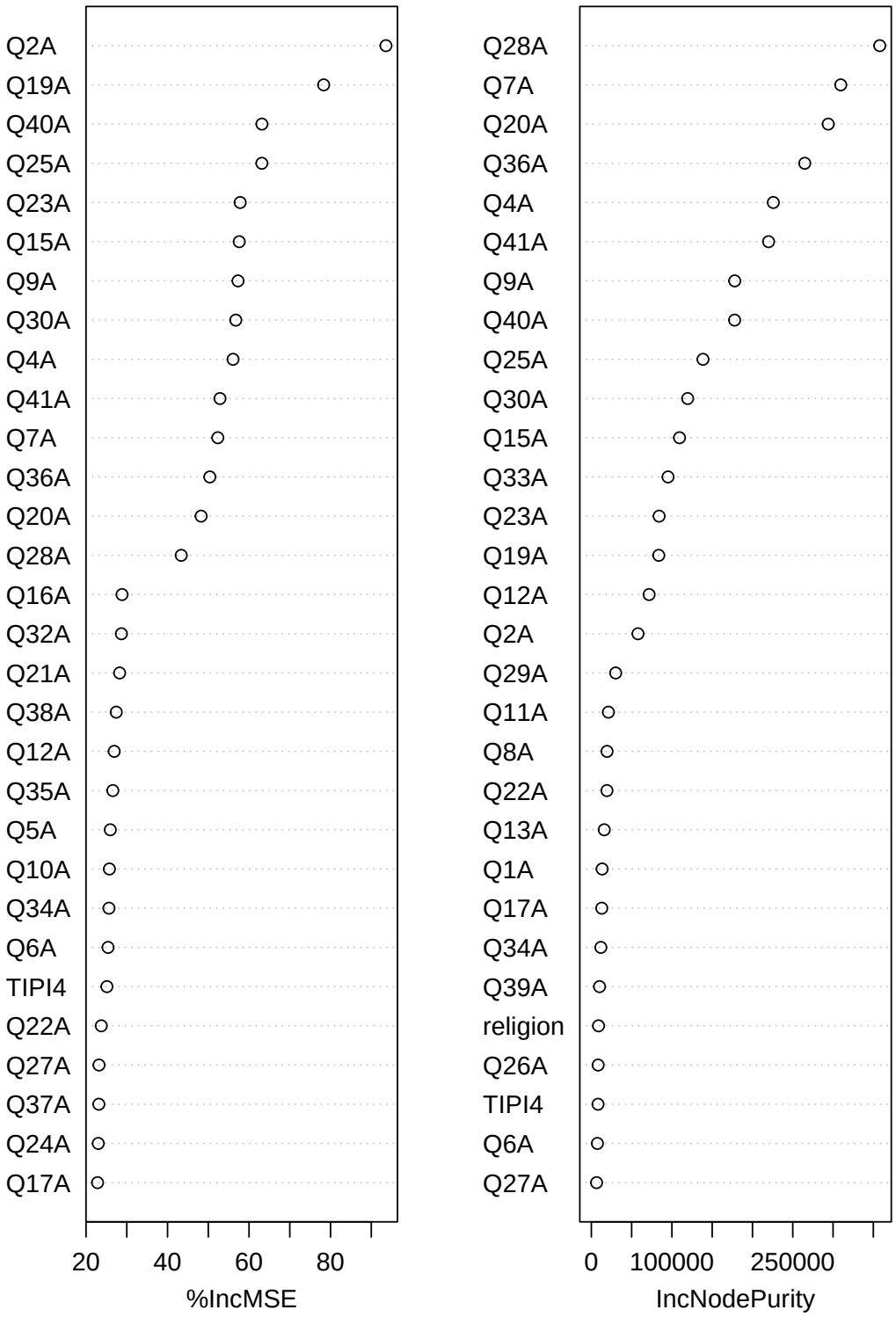


Figure C.3: Depression Random Forest Variable Importance

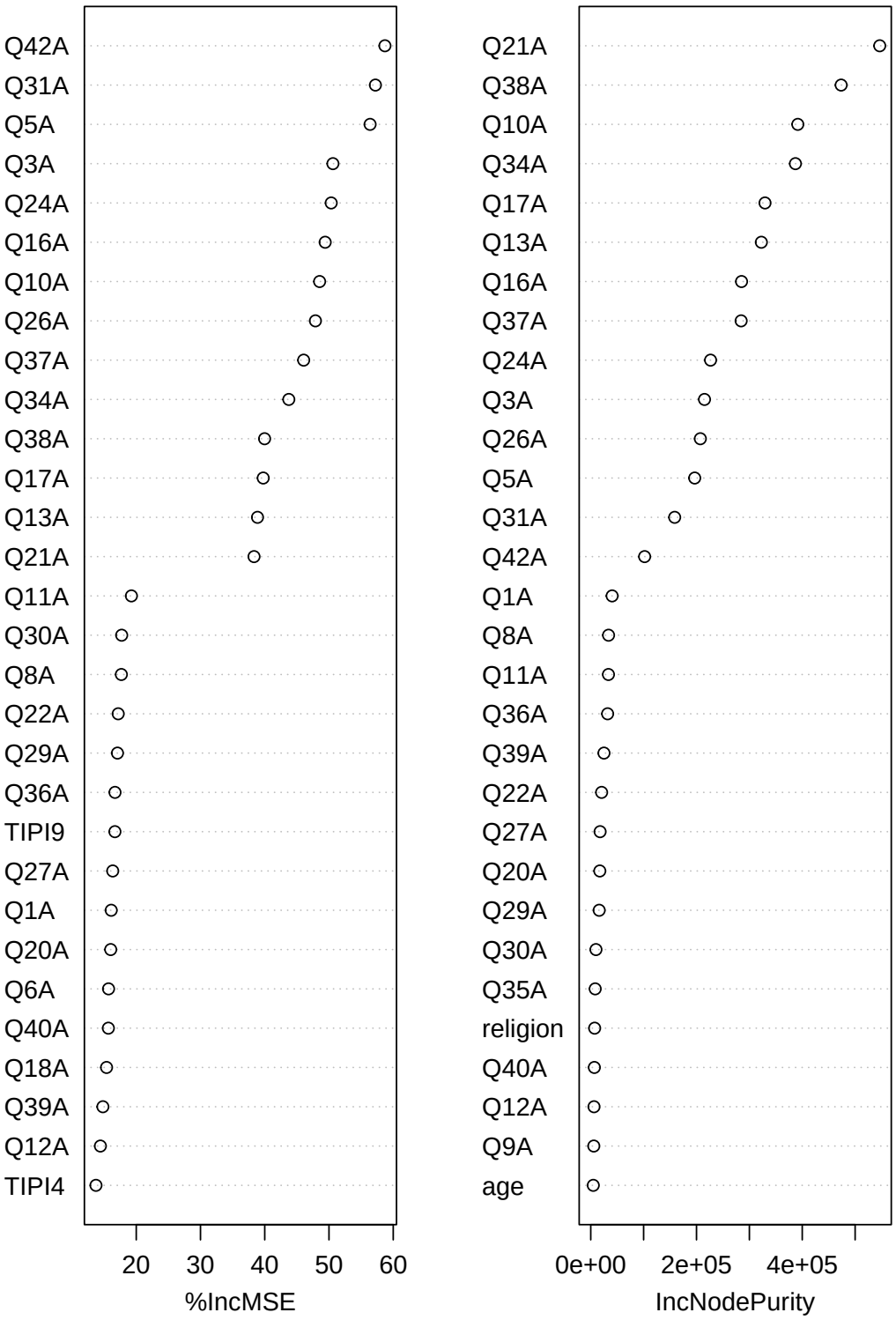
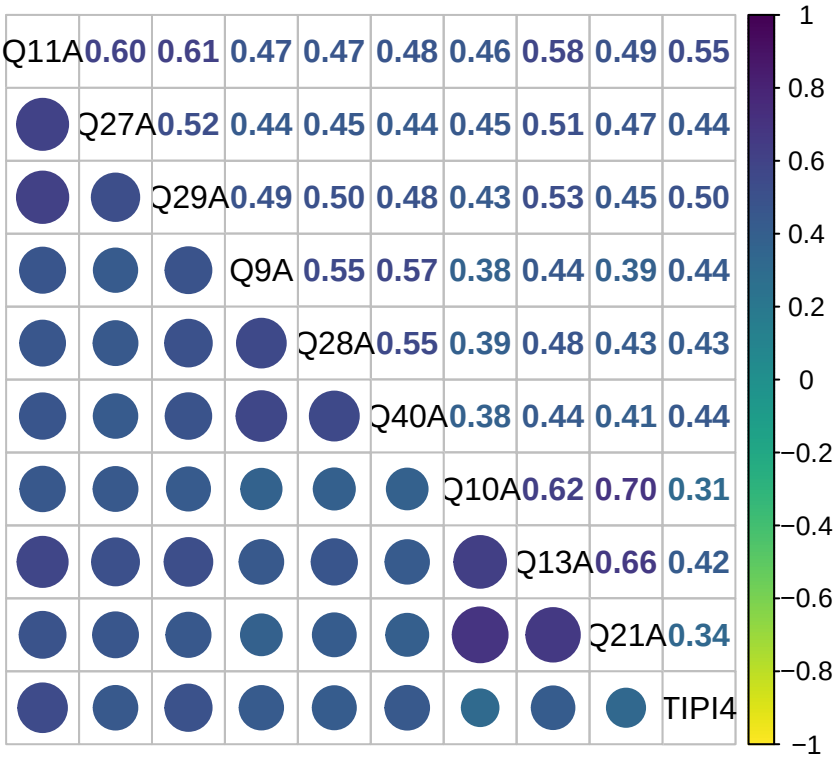


**Figure C.4: Correlations Between Predictors**

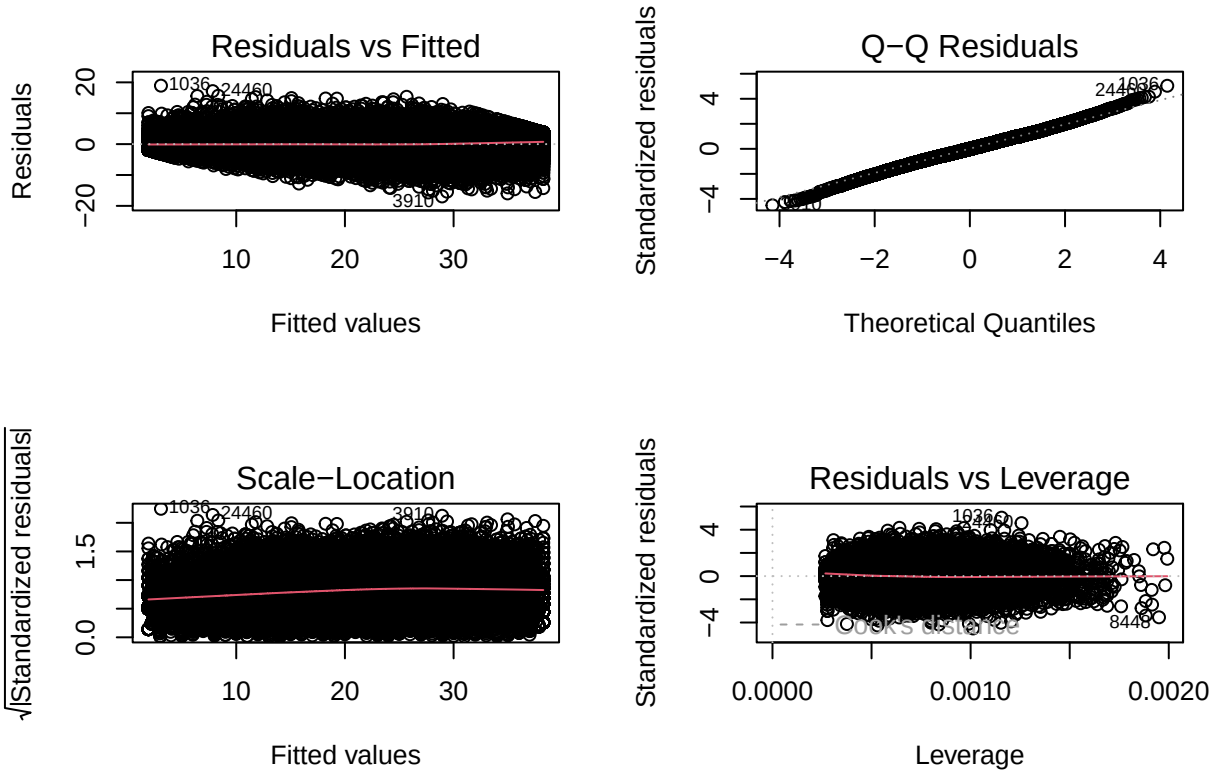


# Appendix D: Model Specification & Diagnostics

## Stress

```
##
## Call:
## lm(formula = stress_score ~ Q11A + Q27A + poly(Q29A, 2) + poly(Q9A,
##     3) + poly(Q28A, 3) + poly(Q40A, 2) + Q10A + Q13A + poly(Q21A,
##     2) + poly(TIPI4, 3), data = trainData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.9430  -2.3929  -0.0326   2.5183  18.9261
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        5.00033    0.12428  40.235  < 2e-16 ***
## Q11A               2.22818    0.03161  70.500  < 2e-16 ***
## Q27A               2.80960    0.02800 100.336  < 2e-16 ***
## poly(Q29A, 2)1   433.38534    5.29064  81.915  < 2e-16 ***
## poly(Q29A, 2)2    13.52736    4.01716   3.367  0.00076 ***
## poly(Q9A, 3)1    168.99328    5.07512  33.298  < 2e-16 ***
## poly(Q9A, 3)2      3.47120    4.04736   0.858  0.39109
## poly(Q9A, 3)3     -9.18648    3.78259  -2.429  0.01516 *
## poly(Q28A, 3)1   256.91431    5.11996  50.179  < 2e-16 ***
## poly(Q28A, 3)2     5.19109    4.02391   1.290  0.19704
## poly(Q28A, 3)3    12.17405    3.77624   3.224  0.00127 **
## poly(Q40A, 2)1   104.46818    5.09938  20.486  < 2e-16 ***
## poly(Q40A, 2)2    20.59881    4.04003   5.099 3.44e-07 ***
## Q10A               0.38350    0.02819  13.605  < 2e-16 ***
## Q13A               0.58072    0.03124  18.590  < 2e-16 ***
## poly(Q21A, 2)1    58.36120    5.84513   9.985  < 2e-16 ***
## poly(Q21A, 2)2    12.12693    3.87999   3.126  0.00178 **
## poly(TIPI4, 3)1   94.91304    4.87273  19.478  < 2e-16 ***
## poly(TIPI4, 3)2   34.53324    4.01466   8.602  < 2e-16 ***
## poly(TIPI4, 3)3   17.89833    3.78177   4.733 2.22e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.758 on 29811 degrees of freedom
## Multiple R-squared:  0.8719, Adjusted R-squared:  0.8719
## F-statistic: 1.068e+04 on 19 and 29811 DF,  p-value: < 2.2e-16
```

$$\begin{aligned}\text{stress\^{}score} = 5 + 2.228 \times \text{Q11A} + 2.81 \times \text{Q27A} + 33.385 \times \text{poly(Q29A, 2)1} + 13.527 \times \text{poly(A29A, 2)2} \\ +168.993 \times \text{poly(Q9A, 3)1} + 3.471 \times \text{poly(Q9A, 3)2} - 9.186 \times \text{poly(Q9A, 3)3} + 256.914 \times \text{poly(Q28A, 3)1} \\ +5.191 \times \text{poly(Q28A, 3)2} + 12.174 \times \text{poly(Q28A, 3)3} + 104.468 \times \text{poly(Q40A, 2)1} + 20.599 \times \text{poly(Q40A, 2)2} \\ +0.384 \times \text{Q10A} + 0.581 \times \text{Q13A} + 58.361 \times \text{poly(Q21A, 2)1} + 12.127 \times \text{poly(Q21A, 2)2} \\ +94.913 \times \text{poly(TIPI4, 3)1} + 34.533 \times \text{poly(TIPI4, 3)2} + 17.898 \times \text{poly(TIPI4, 3)3}\end{aligned}$$


Residuals vs Fitted
Residuals
Fitted values
1036
24460
3910
Q−Q Residuals
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Cook's distance
Leverage
8448

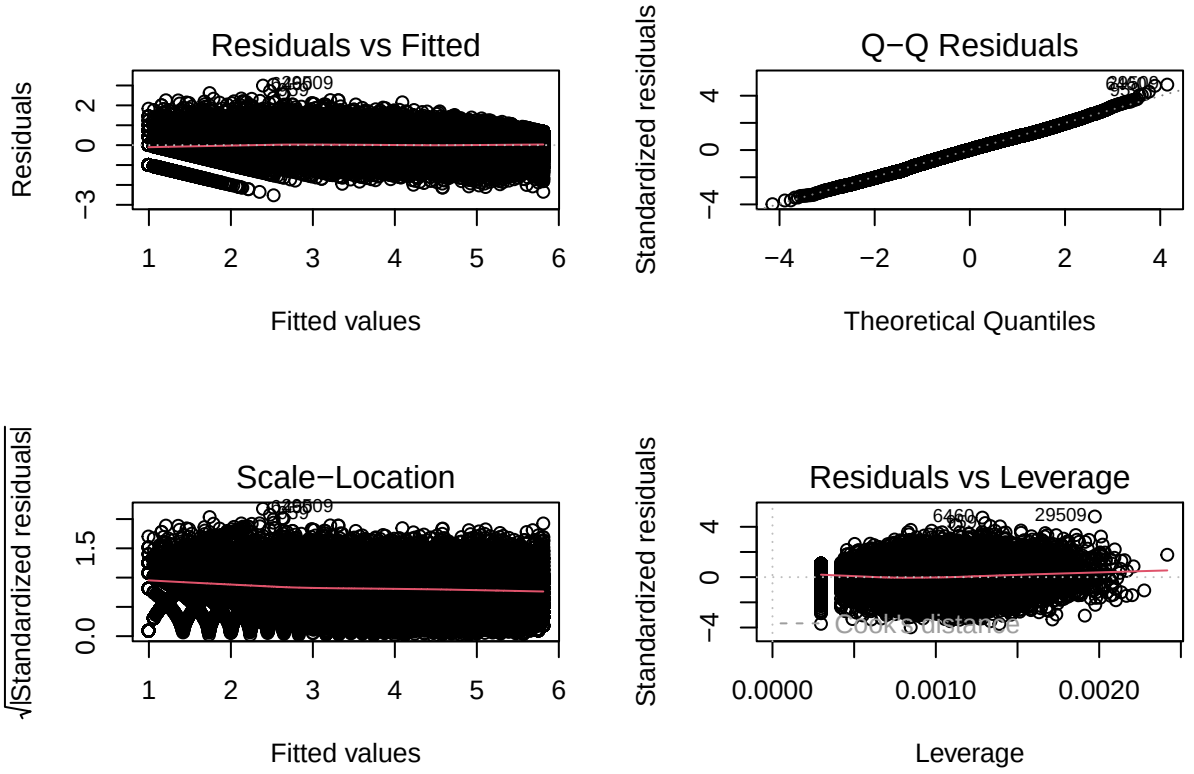
### Anxiety

```
##
## Call:
## lm(formula = anxiety_trans ~ poly(Q11A, 2) + poly(Q27A, 2) +
##     poly(Q29A, 2) + poly(Q9A, 3) + poly(Q28A, 3) + poly(Q40A,
##     2) + poly(Q10A, 3) + poly(Q13A, 3) + poly(Q21A, 3) + bs(TIPI4,
##     df = 5), data = trainData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.52117 -0.41612  0.01612  0.43184  3.04743
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3.60398    0.01775 203.054  < 2e-16 ***
## poly(Q11A, 2)1        8.78217    0.96388   9.111  < 2e-16 ***
## poly(Q11A, 2)2       -1.29194    0.72164  -1.790   0.0734 .
## poly(Q27A, 2)1       11.84901    0.85705  13.825  < 2e-16 ***
## poly(Q27A, 2)2       -2.00229    0.69184  -2.894   0.0038 **
## poly(Q29A, 2)1       21.21724    0.89168  23.795  < 2e-16 ***
## poly(Q29A, 2)2       -1.58343    0.69668  -2.273   0.0230 *
## poly(Q9A, 3)1        62.12923    0.85561  72.614  < 2e-16 ***
## poly(Q9A, 3)2       -12.47651    0.68669 -18.169  < 2e-16 ***
## poly(Q9A, 3)3         3.99622    0.63985   6.246 4.28e-10 ***
## poly(Q28A, 3)1       76.43603    0.86315  88.555  < 2e-16 ***
## poly(Q28A, 3)2       -8.46975    0.68291 -12.402  < 2e-16 ***
## poly(Q28A, 3)3        3.08624    0.63772   4.839 1.31e-06 ***
## poly(Q40A, 2)1       60.12716    0.85966  69.943  < 2e-16 ***
## poly(Q40A, 2)2      -12.00720    0.68350 -17.567  < 2e-16 ***
## poly(Q10A, 3)1        9.75123    0.93555  10.423  < 2e-16 ***
## poly(Q10A, 3)2       -0.35116    0.70360  -0.499   0.6177
## poly(Q10A, 3)3        1.23693    0.64576   1.915   0.0554 .
## poly(Q13A, 3)1       14.38073    0.97707  14.718  < 2e-16 ***
## poly(Q13A, 3)2       -1.36808    0.70938  -1.929   0.0538 .
## poly(Q13A, 3)3        1.00807    0.64375   1.566   0.1174
## poly(Q21A, 3)1       15.82056    0.98745  16.022  < 2e-16 ***
## poly(Q21A, 3)2       -0.61317    0.71291  -0.860   0.3897
## poly(Q21A, 3)3        1.12497    0.64650   1.740   0.0819 .
## bs(TIPI4, df = 5)1    0.06643    0.05025   1.322   0.1862
## bs(TIPI4, df = 5)2    0.28876    0.05417   5.331 9.86e-08 ***
## bs(TIPI4, df = 5)3   -0.01606    0.05301  -0.303   0.7620
## bs(TIPI4, df = 5)4    0.52777    0.11084   4.762 1.93e-06 ***
## bs(TIPI4, df = 5)5    0.17535    0.02036   8.611  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6333 on 29802 degrees of freedom
## Multiple R-squared:  0.8037, Adjusted R-squared:  0.8035
## F-statistic:  4356 on 28 and 29802 DF,  p-value: < 2.2e-16
```

$$\text{anxiety\^{}trans} = 3.604 + 8.782 \times \text{poly(Q11A, 2)1} - 1.292 \times \text{poly(Q11A, 2)2} + 11.849 \times \text{poly(Q27A, 2)1}$$

$$-2.002 \times \text{poly(Q27A, 2)2} + 21.217 \times \text{poly(Q29A, 2)1} - 1.583 \times \text{poly(Q29A, 2)2} + 62.129 \times \text{poly(Q9A, 3)1}$$

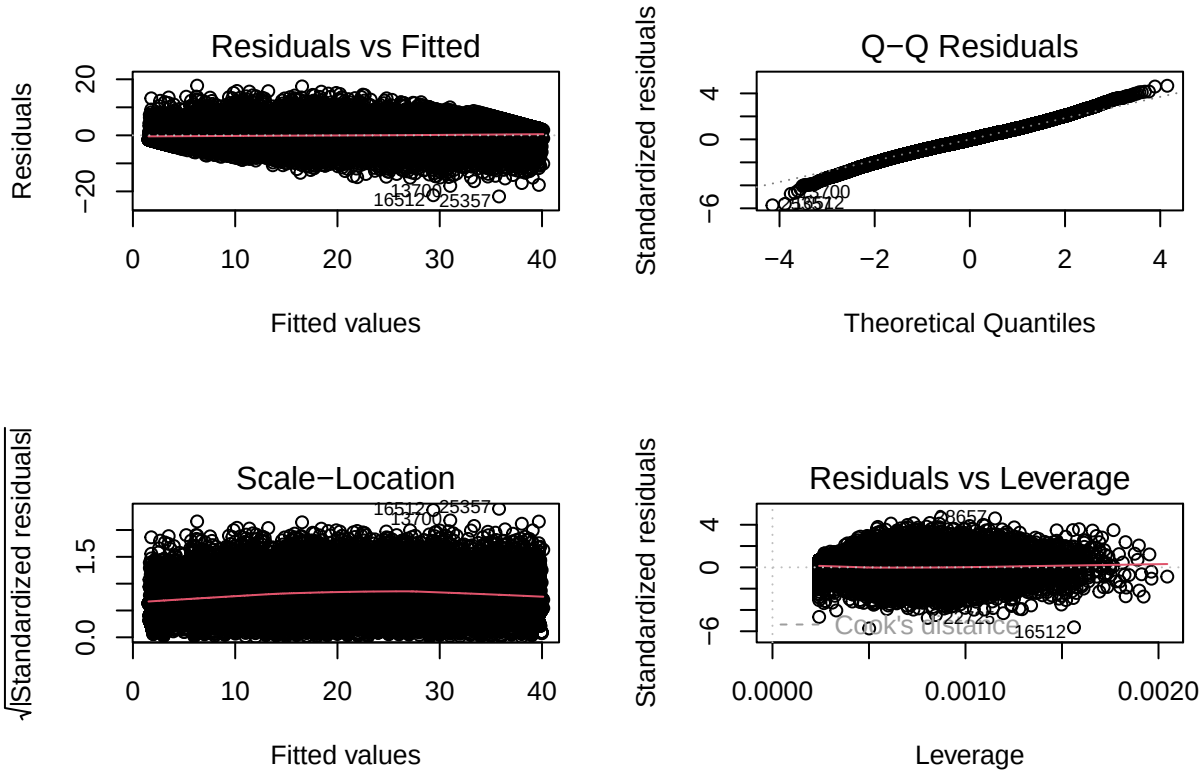
$$-12.477 \times \text{poly(Q9A, 3)2} + 3.996 \times \text{poly(Q9A, 3)3} + 76.436 \times \text{poly(Q28A, 3)1} - 8.47 \times \text{poly(Q28A, 3)2}$$
$$+3.086 \times \text{poly(Q28A, 3)3} + 60.127 \times \text{poly(Q40A, 2)1} - 12.007 \times \text{poly(Q40A, 2)2} + 9.751 \times \text{poly(Q10A, 3)1}$$
$$-0.351 \times \text{poly(Q10A, 3)2} + 1.237 \times \text{poly(Q10A, 3)3} + 14.381 \times \text{poly(Q13A, 3)1} - 1.368 \times \text{poly(Q13A, 3)2}$$
$$+1.008 \times \text{poly(Q13A, 3)3} + 15.821 \times \text{poly(Q21A, 3)1} - 0.613 \times \text{poly(Q21A, 3)2} + 1.125 \times \text{poly(Q21A, 3)3}$$
$$+0.066 \times \text{bs(TIPI4, df} = 5)1 + 0.289 \times \text{bs(TIPI4, df} = 5)2 - 0.016 \times \text{bs(TIPI4, df} = 5)3$$
$$+0.528 \times \text{bs(TIPI4, df} = 5)4 + 0.175 \times \text{bs(TIPI4, df} = 5)5$$



### Depression

```
##
## Call:
## lm(formula = depression_score ~ Q11A + Q27A + Q29A + poly(Q9A,
##     3) + Q28A + poly(Q40A, 3) + Q10A + Q13A + poly(Q21A, 3) +
##     bs(TIPI4, df = 5), data = trainData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.8009  -2.3847  -0.0393   2.3604  17.7210
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -1.04905    0.14896  -7.043 1.93e-12 ***
## Q11A                 0.28799    0.03197   9.007  < 2e-16 ***
## Q27A                 0.66434    0.02832  23.458  < 2e-16 ***
## Q29A                 0.42302    0.02929  14.441  < 2e-16 ***
## poly(Q9A, 3)1       66.11393    5.13121  12.885  < 2e-16 ***
## poly(Q9A, 3)2        9.57005    4.01721   2.382  0.01721 *
## poly(Q9A, 3)3       -9.77804    3.83632  -2.549  0.01081 *
## Q28A                 0.24589    0.02792   8.808  < 2e-16 ***
## poly(Q40A, 3)1      57.59132    5.15614  11.169  < 2e-16 ***
## poly(Q40A, 3)2      13.21327    4.01506   3.291  0.00100 ***
## poly(Q40A, 3)3      -9.74662    3.83315  -2.543  0.01100 *
## Q10A                 3.61373    0.02851 126.736  < 2e-16 ***
## Q13A                 3.30201    0.03160 104.505  < 2e-16 ***
## poly(Q21A, 3)1     720.27828    5.91302 121.812  < 2e-16 ***
## poly(Q21A, 3)2      -1.71420    3.89136  -0.441  0.65957
## poly(Q21A, 3)3      10.92734    3.81235   2.866  0.00416 **
## bs(TIPI4, df = 5)1   0.53187    0.30038   1.771  0.07663 .
## bs(TIPI4, df = 5)2  -0.69850    0.32466  -2.152  0.03144 *
## bs(TIPI4, df = 5)3   0.26189    0.31555   0.830  0.40657
## bs(TIPI4, df = 5)4  -1.10003    0.66515  -1.654  0.09818 .
## bs(TIPI4, df = 5)5  -0.27480    0.12081  -2.275  0.02294 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.801 on 29810 degrees of freedom
## Multiple R-squared:  0.9047, Adjusted R-squared:  0.9047
## F-statistic: 1.415e+04 on 20 and 29810 DF,  p-value: < 2.2e-16
```

$$\begin{aligned}\text{depression\^{}score} = &-1.049 + 0.288 \times \text{Q11A} + 0.664 \times \text{Q27A} + 0.423 \times \text{Q29A} + 66.114 \times \text{poly(Q9A, 3)1} \\ &+9.57 \times \text{poly(Q9A, 3)2} - 9.778 \times \text{poly(Q9A, 3)3} + 0.246 \times \text{Q28A} + 57.591 \times \text{poly(Q40A, 3)1} \\ &+13.213 \times \text{poly(Q40A, 3)2} - 9.747 \times \text{poly(Q40A, 3)3} + 3.614 \times \text{Q10A} + 3.302 \times \text{Q13A} \\ &+720.278 \times \text{poly(Q21A, 3)1} - 1.714 \times \text{poly(Q21A, 3)2} + 10.927 \times \text{poly(Q21A, 3)3} \\ &+0.532 \times \text{bs(TIPI4, df = 5)1} - 0.699 \times \text{bs(TIPI4, df = 5)2} + 0.262 \times \text{bs(TIPI4, df = 5)3} \\ &-1.1 \times \text{bs(TIPI4, df = 5)4} - 0.275 \times \text{bs(TIPI4, df = 5)5}\end{aligned}$$


Residuals vs Fitted
Residuals
Fitted values
Q−Q Residuals
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage
16512
25357
13700