Statistical tests

Which statistical test should we choose?

The answer will depend on answering three questions:

1. What type of data are we using?

   Our data set contains predominantly quantitative data (measurements, counts, pre-established scales), but some categorical data as well (labels).

   We will need different statistical tests for different data types in our sample (would ask for Ben/Zein's advice on how do best do this).

2. How many different sets of data do we have?

   I.e. how many different measurements do we have for a single sample set (the sample set being the young adult cohort in our case), or how many different sample sets do we have (if we decide to include MCS2-5, as opposed to only MCS6).

   In our case, we have 1000s of different measuements.

3. What's the purpose of the tests?

   We want understand the predictive power of x for y.

In short, we are building a logistic regression models which contains a mixture of categorical and numerical variables (x) to predict a binary outcome (y). The key objective for our statistical tests are to understand model performance and the importance of the predictor variables. Specifically, we want to use at least four types of statistical tests to evaluate our model:

1. **Goodness of fit tests**
    1. Assess whether the model adequately fits the data.
    2. The null hypothesis is that the fitted model is correct, and the output is a p-value (a number between 0 and 1 with higher values indicating a better fit).
    3. Common tests:
        1. Hosmer-Lemeshow test: Assesses the goodness of fit by comparing observed and expected event rates across different groups or deciles of predicted probabilities. It divides the data into a specified number of groups based on predicted probabilities (often 10 groups or deciles) and compares the observed and expected frequencies of outcomes within each group using a chi-square statistic. A nonsignificant p-value (usually greater than 0.05) suggests that there is no significant difference between observed and expected event rates, indicating that the model fits the data well. Compared to the Pearson chi-square test, the Hosmer-Lemeshow model allows you to understand how the model performs across different levels of predicted probabilities or risk groups, which is why I think it would be a better choice.

2. Pearson chi-square test: Assesses the overall goodness of fit by comparing the observed frequencies of outcomes with the frequencies predicted by the logistic regression model. It calculates a chi-square statistic based on the discrepancy between observed and expected frequencies. A nonsignificant p-value indicates that there is no significant difference between observed and expected frequencies, suggesting a good fit.
3. Pearson Residuals: Pearson residuals are calculated by taking the difference between observed and predicted probabilities, divided by the standard error of the predicted probability. These residuals are used to detect lack of fit or potential outliers in the model. Large Pearson residuals may indicate observations where the model performs poorly in predicting the outcome.
4. Deviance Residuals: Deviance residuals are based on the difference between the observed log-likelihood and the log-likelihood expected under the fitted model. Like Pearson residuals, deviance residuals are used to identify observations that are poorly predicted by the model. Deviance residuals tend to be more sensitive to outliers and extreme values than Pearson residuals.

2. **Measures of predictive power**
    1. Assess how well the independent variables predict the outcome variable.
    2. Typically vary between 0 and 1, with 0 meaning no predictive power and 1 meaning perfect predictions. The higher the better, but there is no fixed cut-off for an 'adequate' model.
    3. Common tests:
        1. R-square: Measures the proportion of variance in y that is explained by x. Unlike linear regression, logistic regression does not have a single R-square measure, such as the coefficient of determination (R^2), because the likelihood function used in logistic regression does not lend itself to direct interpretation in terms of variance explained. However, several pseudo-R-square measures have been proposed to provide a similar interpretation, such as:
            - Cox and Snell R-square: This pseudo-R-square is defined as the proportional increase in likelihood for the full model compared to the null model (model with no predictors). It ranges from 0 to 1, with higher values indicating better model fit.
            - Nagelkerke R-square: This is an adjusted version of the Cox and Snell R-square that adjusts for the maximum possible value of the likelihood function. It provides a better approximation to the proportion of variance explained by the model and can range from 0 to 1.

            Pseudo-R-square measures in logistic regression provide an indication of how well the model fits the data and the amount of variation in the outcome variable that is accounted for by the predictor variables. However, they should be interpreted with caution as they are not directly comparable to R-square in linear regression and do not have the same intuitive interpretation.
        2. Area under the ROC curve (AUC-ROC): Is a graphical representation of the trade-off between sensitivity (true positive rate) and specificity (true negative rate) for different threshold values of a binary classifier. The AUC-ROC quantifies the overall discriminative ability or predictive performance of a logistic regression model. AUC-ROC values range from 0 to 1, where a value of 0.5 indicates random prediction (no discriminative ability), and a value of 1 indicates perfect prediction. Typically, higher AUC-ROC values indicate better model

performance in distinguishing between the two classes (e.g., diseased vs. non-diseased). An AUC-ROC value above 0.7 or 0.8 is generally considered acceptable, but the interpretation may vary depending on the specific application and context.

3. Rank-order correlations, such as Spearman's rank correlation coefficient or Kendall's tau coefficient, measure the strength and direction of the association between two ranked variables. In logistic regression, rank-order correlations can be used to assess the concordance between predicted probabilities from the model and observed outcomes. High rank-order correlations indicate that the predicted probabilities from the logistic regression model are in agreement with the observed outcomes, suggesting good predictive performance. Conversely, low rank-order correlations may indicate poor model calibration or discrimination. However, rank-order correlations do not provide information about the magnitude of prediction errors or the overall model fit.

3. **Signficance of individual predictor variables**
    1. Assess the importance of individual independent variables to predicting the outcoming variable.
    2. Common tests:
        1. Wald test: This test assesses the significance of individual variable, by testing whether each coefficient is significantly different from zero.
        2. Likelihood ratio test: If you're comparing nested models like we are (e.g., adding or removing variables from a model), you can use the likelihood ratio test to assess whether the added variables significantly improve model fit.
        3. Odds ratio and confidence intervals: For each predictor variable, the odds ratio and its confidence interval assess the magnitude and direction of the effect.
4. **Multicollinearity**
    1. Assess whether predictor variables are highly correlated with each other.
    2. Common tests:
        1. Correlation matrix: Calculate the correlation coefficients between all pairs of independent variables. High correlations (typically above 0.7 or 0.8) indicate potential multicollinearity issues. However, note that correlation alone does not capture all aspects of multicollinearity, especially if the relationships are nonlinear.
        2. Variance Inflation Factor (VIF): VIF measures how much the variance of an estimated regression coefficient is increased due to multicollinearity. A VIF value greater than 5 or 10 is often considered indicative of multicollinearity.
        3. Tolerance: Tolerance is the reciprocal of VIF and indicates the proportion of variance in a predictor variable that is not explained by other predictors. A tolerance value less than 0.1 suggests significant multicollinearity.

Outstanding question: Should we use stepwise regression for our model?

Stepwise regression is a method used to select the most important predictors from a large set of potential predictor variables. It's particularly useful when dealing with datasets where there are numerous independent variables and you want to identify a subset of predictors that best explain the variation in the dependent variable.

There are two main approaches to stepwise regression:

1. **Forward Selection**: In forward selection, the process begins with an empty model (no predictors included). Predictors are added to the model one at a time, and at each step, the predictor that most improves the fit of the model is added. This process continues until no additional predictors significantly improve the model fit, based on predefined criteria (e.g., p-value, AIC, BIC).
2. **Backward Elimination**: In backward elimination, the process starts with a model that includes all potential predictors. At each step, the predictor that contributes the least to the model (e.g., has the highest p-value) is removed. This process continues until removing additional predictors does not significantly improve the model fit, based on predefined criteria.

The choice between forward selection and backward elimination often depends on practical considerations and the specific goals of the analysis.

Here's a general outline of how stepwise regression works:

1. **Start**: Begin with an empty model (for forward selection) or a model including all potential predictors (for backward elimination).
2. **Step 1**: Perform either forward selection or backward elimination to add or remove predictors from the model.
3. **Step 2**: At each step, assess the fit of the model using a criterion such as the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), or p-values. These criteria balance the goodness of fit with model complexity.
4. **Step 3**: Continue the stepwise process until the addition or removal of predictors no longer significantly improves the model fit or until predefined stopping criteria are met.
5. **Final Model**: The final model consists of the predictors selected through the stepwise procedure.