Final Exam Review

Format of the Exam

This final exam is cumulative. You will be tested on material from Unit 1-Unit 20 (up until section 2). Like the last two exams, you will be expected to read a dataset and be able to answer the following questions involving this dataset. There may be some conceptual questions as well. See the attached documents for specific question examples of this nature. The attached Jupyter notebook is not comprehensive, in that it does not cover ALL things talked about in this document and what might be on the test.

1. Cumulative Final Exam: Be sure to review questions/items in the Midterm 1 Review and Midterm 2 Reviews!

2. More About Logistic Regression

2.1. What objective function are optimizing when you find the "optimal" values of the slopes and intercept in a logistic regression model.

2.1.1. Conducting Inference on Logistic Regression Slopes

2.1.1.1. Using p-values and confidence intervals.

2.1.2. **Log-likelihood function**

2.1.2.1. What is the log likelihood function? Aka, write out the function that you are trying to optimize that gives you the "best" values of the slopes and intercept in a logistic regression model.

2.1.2.2. How to extract the optimal log-likelihood function value of a model in Python.

3. Making Predictions with a Classifier: Making 0/1 Predictions with a Threshold and Assessing the Accuracy of the Predictions

3.1. Definitions/Calculations

3.1.1. Predictive Probability Threshold

3.1.1.1. How to define.

3.1.1.2. How to classify a set of explanatory variables values given their corresponding predictive probabilities and a predictive probability threshold.

3.1.2. True positives, true negatives, false positives, false negatives

3.1.2.1. How to define

3.1.2.2. How to calculate by hand (given a set of true labels and predicted labels).

3.1.2.3. How to calculate with Python function.

3.1.3. **Confusion matrix**

3.1.3.1. How to define

3.1.3.2. How to calculate with Python function.

3.1.4. Sensitivity, Specificity, True Positive Rate, False Positive Rate, and Accuracy

3.1.4.1. How to define

3.1.4.2. How to calculate by hand (given a set of true labels and predicted labels).

3.1.4.3. The relationship between each of these terms.

3.1.4.4. How to calculate with Python function.

3.1.5. ROC Curve and AUC (Area under Curve)

3.1.5.1. How to define each.

3.1.5.2. The relationship between each of these terms.

3.1.5.3. How to plot the ROC curve with Python (given a set of true labels and predicted labels).

3.1.5.4. How to calculate the AUC with Python (given a set of true labels and predicted labels).

3.1.5.5. How to use the ROC curve to pick out the "ideal" predictive probability threshold.

3.1.5.6. How to interpret the ROC curve and AUC. Given two models, be able to determine what has a better ROC and AUC.

3.1.5.7. What is the best possible ROC and AUC a model can have, and why is this the best?

4. Variance-Bias Tradeoff in Model Building

4.1. Problem of Overfitting a Model (High Variance)

4.1.1. **Overfitting**

4.1.1.1. What does it mean when a model has been "overfit?" Would we expect a model that has been "overfit" with a training dataset to perform well with a test dataset?

4.1.1.2. Is a model more likely to be overfit if it has more or less explanatory variables?

4.2. Problem of Poor Predictive Performance of a Model (High Bias)

4.2.1. **Predictive Power**

4.2.1.1. Is a model more likely to have better predictions if it has more or less explanatory variables.

4.3. Balancing the Variance-Bias Tradeoff

4.3.1. **Parsimonious Model**

4.3.1.1. What does it mean for a model to model to be parsimonious?

4.3.1.2. What two model statistics did we learn to assess how parsimonious a model is?

5. Building, Selecting, and Assessing a Models Ability to Predict *New Data*

5.1. Using a training dataset and a test dataset

5.1.1. **Training Data vs. Test Data**

5.1.1.1. What is the purpose of the training dataset?

5.1.1.2. What is the purpose of the test dataset?

5.1.1.3. How to randomly create a single training dataset and test dataset in Python.

5.1.1.4. Why would assessing predictive accuracy on the test dataset provide a better idea as to how the model will perform with *new data,* than if you had assessed predictive accuracy with the training dataset?

6. Selecting the "Best" Explanatory Variables to Use in a Regression Model

6.1. Inference-Based Method (used to compare a **full model** and a **reduced model).**

6.1.1. **Log-Likelihood Ratio Test**

6.1.1.1. How to set up a **full model** and a **reduced model** with respect to this test.

6.1.1.2. How to set up the null and alternative hypothesis for this test.

6.1.1.3. How to calculate the test statistic for this test in Python (given the output of the two models).

6.1.1.4. What distribution (with what degrees of freedom) is this test statistic an observation from?

6.1.1.5. How to calculate the p-value for this test.

6.1.1.6. How to interpret the p-value for this test.

6.2. Model Statistic-Based Methods (used to compare **multiple different models).**

6.2.1. **AIC**

6.2.1.1. How to calculate the AIC by hand (given a logistic regression model output table).

6.2.1.2. How to find the AIC of a model using Python functions.

6.2.1.3. What does AIC measure?

6.2.1.4. What kind of AIC scores do we prefer in a model?

6.2.2. **BIC**

6.2.2.1. How to calculate the BIC by hand (given a logistic regression model output table).

6.2.2.2. How to find the BIC of a model using Python functions.

6.2.2.3. What does BIC measure?

6.2.2.4. What kind of BIC scores do we prefer in a model?

6.2.3. **Comparison**

6.2.3.1. For the same model, (which uses a sample size n>8), which statistic will penalize a larger number of slopes more: the AIC or the BIC of the model?

6.3. Backwards Elimination Algorithm

6.3.1. How to use it to try to find the model with the best BIC or AIC score.