**Neha Maddali**

**Question 1:**

The explanatory variables in my final logistic regression model to predict the probability of staying drug-free for one year after completion of the program is: AGE, IVHXPrevious, IVHXRecent, and NDTX.

The process of choosing this model looked like so:

1. Categorical variables were converted to factors to prepare the data for modeling in logistic regression
2. A full model was built using all potential explanatory variables: AGE, BECK, IVHX, NDTX and RACE.
3. The stepwise regression process was used to select the best model based on the AIC. The final model retained variables AGE, IVHX, and NDTX while excluding BECK and RACE.

**Question 2:**

The likelihood ratio test was performed to evaluate whether the model as a whole helps to explain the probability of staying drug-free in the population. The test resulted in a test statistic of 1.226 and p-value of 0.268.

With a p-value of 0.268 there is insufficient evidence to reject the null hypothesis. So, the model as a whole doesn’t significantly improve the fit compared to a null model (intercept-only model) in explaining the probability of staying drug-free in this population.

**Question 3:**

A screenshot of a computer

Description automatically generated

1. AGE
   1. Hypothesis test for the coefficient of AGE in the logistic regression model results in a test statistic of 2.943 and a corresponding p-value of 0.003.
   2. With a p-value of 0.003, there is sufficient evidence to reject the null hypothesis. Therefore, the variable AGE significantly helps to explain the probability of staying drug-free in this population, given the presence of other explanatory variables in the model.
2. IVHX
   1. IVHXPrevious
      1. The hypothesis test for the coefficient of IVHXPrevious results in a test statistic of -2.560 and a corresponding p-value of 0.010.
      2. With a p-value of 0.010, there is sufficient evidence to reject the null hypothesis. Therefore, the variable IVHXPrevious significantly helps to explain the probability of staying drug-free in this population, given the presence of other explanatory variables in the model.
   2. IVHXRecent
      1. The hypothesis test for the coefficient of IVHXRecent results in a test statistic of -3.868 and a corresponding p-value of 0.0001.
      2. With a p-value of 0.0001, there is sufficient evidence to reject the null hypothesis. Therefore, the variable IVHXRecent significantly helps to explain the probability of staying drug-free in this population, given the presence of other explanatory variables in the model.
3. NDTX
   1. The hypothesis test for the coefficient of NDTX results in a test statistic of -1.744 and a corresponding p-value of 0.081.
   2. With a p-value of 0.081, there is insufficient evidence to reject the null hypothesis. Therefore, the variable NDTX does not significantly help to explain the probability of staying drug-free in this population, given the presence of other explanatory variables in the model.

So, Age, History of Intravenous Drug Use (both "Previous" and "Recent" categories), but not Number of Previous Drug Treatments, are significant predictors of staying drug-free in this population when other explanatory variables are considered.

**Question 4:**

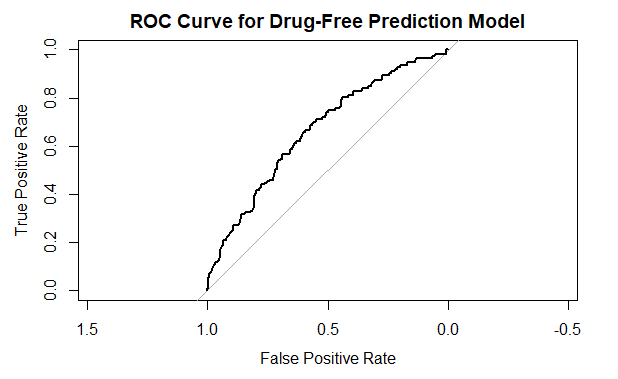
A screen shot of a computer

Description automatically generated

Summary of Statistics:

* The model achieved an agreement of approximately 0.7482, indicating moderate agreement.
* The model demonstrated low sensitivity, correctly identifying only about 5.41% (0.0541) of the drug users.
* The model exhibited high specificity, correctly identifying approximately 99.36% (0.9936) of the non-drug users
* The positive predictive value was 0.7500.
* The negative predictive value was 0.7482

**Question 5:**



The AUC for the drug-free prediction model is 0.6655. This suggests that the model has fair discriminative ability, meaning it performs better than random guessing but has room for improvement in accurately distinguishing between drug-free and non-drug-free individuals.