**Statistics 402 – Fall 24 – Homework Three**

**Professor Esfandiari**

**Problem one**

Using the liver.csv data given in the homework three folder given on week nine, answer the following questions:

**Background of the study**

This data set results from consulting with a liver transplant surgeon, professor Zarrinnia, who wanted to predict the odds of graft survival from multiple predictors. As Dr. Zarrinpar explained, they do not need to transplant the whole liver, they could do partial transplantation and they call it graft transplantation. If graft transplanted survives for a month in the body of the patient, then it is considered successful and it will continue to work. However, if it fails before a month, then it is considered a failure.

**Question asked by the client?**

Is graft failure (kidney transplant was not successful) related to height, BMI, history of hypertension, and whether the donor died of coronary heart disease. The definition of the outcome might look counter-intuitive meaning that usually the outcome is defined as odds of success. But this is how the client wanted it defined as graft failure.

**b. Clarification of the variables involved and how they were measured**

**Outcome variable**

* Did kidney transplant fail?

Variable label = “graft survival” (tx\_fail)

1 = failure

0 = success

**Predictors**

* Height of the donor (numerical)

Variable label = (hgt\_cm\_don\_calc.x)

* Body mass index (numerical)

Variable label = (bmi\_don\_calc.x)

* Whether the donor died as a function of coronary heart disease (categorical)

Variable label = coroncoronary\_angio\_donor

* If the donor had a history of hypertension (categorical) – fluctuation in blood pressure

Variable label = hist\_hypertens\_don.x

**Given the above information, answer the following questions:**

1. Conduct the relevant EDA including looking at relevant histograms, contingency tables, etc.
2. Run a logistic model to predict graft failure as a function of the predictors of interest.
3. Exponentiate the coefficients of the model and interpret them within context.
4. Exponentiate the relevant confidence intervals.
5. Using the results in part three and four, complete table one.

*Table one: The odds, 95% confidence interval for the odds, and p-value of kidney transplant failure for different predictors in the model*

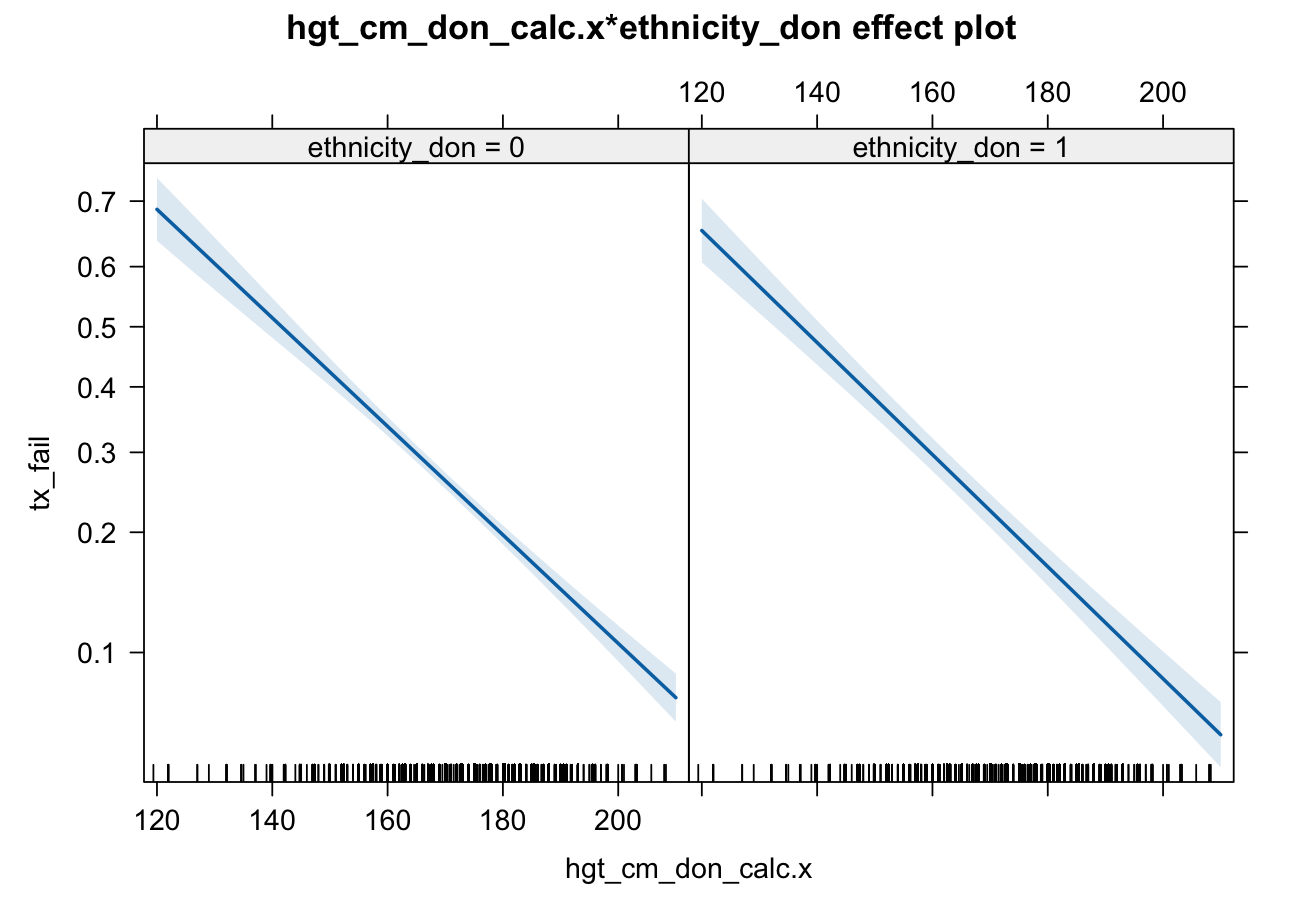
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | *Confidence interval for the odds* | | *p-value* |
| *Predictor* | *odds* | *2.5%* | *97.5%* |
| *height* |  |  |  |  |
| *BMI* |  |  |  |  |
| *Coronary angio* |  |  |  |  |
| *hypertension* |  |  |  |  |

1. Draw the plot of odd
2. Summarize the results of the plot of odds and the relevant 95% confidence intervals within context
3. What does null and residual deviance show?
4. Calculate pesudo R-squared.
5. Calculate accuracy using the confusion matrix.
6. Carry out five-fold cross validation and report the results.
7. Given the following model and the resulting plot (plot one), answer questions a and b.

m1<-glm(tx\_fail~hgt\_cm\_don\_calc.x:ethnicity\_don,family="binomial")

1. What is the question underlying the following plot.
2. What is the answer to this question.

Plot one



1. Now create a new model and add the above factor (hgt\_cm\_don\_calc.x:ethnicity\_don) to it.
2. Carry out the relevant ANOVA test to show if the two models are different and decide which model you would recommend.
3. Create the ROC curve for the model that you recommend and interpret it.

**Problem two.**

Using the diabetic data set in the homework folder three on week nine, perform the following.

1. Do the following data manipulation
2. For the variable of Totalriskfactor, recode so that 0 = none, 1 = one, and classify as “2 or more”. So, you will have three levels and reference will be none.
3. Recode diabetes.new (0,1) to “0 = no” - does not have diabetic type II – and “1= yes” – has diabetic type II.
4. Smoking ((SmokingStatus\_NISTCode) has three levels including false (non-smoker will be the base) former smokers and smokers.
5. Relevel age.new (low, medium, high) so that low will be the reference age category.
6. Do the relevant EDA by doing the relevant contingency tables.
7. Run a multinomial model with age.new (three levels as the outcome) and Totalriskfactors, diabetes.new, hypertensionDX , and smoking status as predictors.
8. Exponentiate the relevant coefficients and confidence intervals.
9. Using the results in parts two and three, complete the following tables one and two.

**Table one – log of odds, odds ratios, 95% confidence interval for odds ratios, and p-value for comparing participants in the “medium” age group for those in the “low” age group.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| variable | Log of odds | odds | 95% CI for odds | P-value |
| One risk vs none |  |  |  |  |
| Two or more risks vs none |  |  |  |  |
| Has diabetic (yes vs no) |  |  |  |  |
| HypertensionDX (Yes vs. No) |  |  |  |  |
| Former smoker vs non-smoker |  |  |  |  |
| Smoker vs smoker |  |  |  |  |

**Table two – log of odds, odds ratios, 95% confidence interval for odds ratios, and p-value for comparing participants in the “high” age group for those in the “low” age group.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| variable | Log of odds | odds | 95% CI for odds | P-value |
| One risk vs none |  |  |  |  |
| Two or more risks vs none |  |  |  |  |
| Has diabetic (yes vs no) |  |  |  |  |
| HypertensionDX (Yes vs. No) |  |  |  |  |
| Former smoker vs non-smoker |  |  |  |  |
| Smoker vs smoker |  |  |  |  |

1. For table two, interpret the following within context.
2. The odds of high risk factor.
3. The odds of diabetic type II.
4. The odds of smoker
5. The 95% confidence interval for the odds of diabetic type II.
6. The 95% confidence interval for smoker.
7. Using the results reported in table two, write the relevant equation for the predictive model created.
8. Create the plot of odds for the multinomial model
9. Summarize the plot of odds for high vs low age; elaborating the effect of predictors on the outcome without interpretation the numerical value of odds or confidence intervals (For instance factor one increases the odds of hypertension for people with high age group compared to low age group etc.)
10. Create the confusion matrix for the multinomial model you created and calculate accuracy.
11. Explain the areas (accuracy, sensitivity, specificity) in which your mode performed well and the areas in which it performed poorly.