**J. Jedediah Smith**

**BIFX 503: Biostatistics in R**

**Homework Set #3**

**Due: November 2, 2021**

*Instructions:*

*Use R to complete this assignment. Assignment is to be submitted via Blackboard.*

The R dataset **Pima.tr** (from the MASS package) contains data on 200 women of Pima Indian heritage living near Phoenix, Arizona. All women were tested for diabetes according to World Health Organization criteria. The dataset contains whether or not they met WHO criteria for diabetes (variable type=Yes for diabetes, No otherwise), as well as data on risk factors for diabetes.

1. Fit a series of bivariate logistic regression models having whether they met WHO criteria for diabetes (type) as the dependent variable and each of these independent variables: npreg, bp, skin, bmi, and age (5 models total). Report the results in terms of regression coefficients and 95% confidence intervals. Which variables are significant predictors of diabetes?

npreg: The regression coefficient is 1.1838717. The 95% confidence interval is 1.0839791 to 1.2992759. Since the p-value was 0.000242, npreg is a significant predictor of diabetes.

bp: The regression coefficient is 1.04088122. The 95% confidence interval is 1.013538015 to 1.0707704. Since the p-value was 0.004053, bp is a significant predictor of diabetes.

skin: The regression coefficient is 1.0473224. The 95% confidence interval is 1.0192763 to 1.0786565. Since the p-value was 0.00136, skin is a significant predictor of diabetes.

bmi: The regression coefficient is 1.11050925. The 95% confidence interval is 1.054302323 to 1.1743257. Since the p-value was 0.000129, bmi is a significant predictor of diabetes.

age: The regression coefficient is 1.07482437. The 95% confidence interval 1.04466978 to 1.1083821. Since the p-value was 1.58e-06, age is a significant predictor of diabetes.

Text

Description automatically generated

(Included the output just for age since there would otherwise be tons of screenshots)

1. Fit a multiple logistic regression model of predictors of diabetes:
   1. In which order should IVs enter the model?

They should enter in order of significance with the most significant first. In our case, this would be: age, bmi, npreg, skin, bp.

* 1. Are any IVs removed? If so, why?

Yes. We would remove npreg and bp because they do not significantly improve model fit when added to the type-age model.

* 1. What is the final model? Report the results in terms of regression coefficients and 95% confidence intervals.

The final model uses age, bmi, and skin. For age, the regression coefficient was 1.073794882 and the 95% confidence interval was 1.0421930682 to 1.10892247. For bmi, the regression coefficient was 1.112213221 and the 95% confidence interval was 1.0327747253 to 1.20052543. For skin, the regression coefficient was 0.999059608 and the 95% confidence interval was 0.9627536112 to 1.04004093.



* 1. Interpret the meaning of each regression coefficient in your final model.

The regression coefficients are interpreted as an odds ratio. Age has a regression coefficient of 1.073794882. Since it is bigger than 1, it has a positive association with type. A unit increase in age makes the diabetic type outcome 1.073794882 times more likely. Bmi has a regression coefficient of 1.112213221. Since it is bigger than 1, it has a positive association with type. A unit increase in bmi makes the diabetic type outcome 1.1112213221 times more likely. Skin has a regression coefficient of 0.999059608. Since it is smaller than 1, it has a negative association with type. A unit increase in skin makes the diabetic type outcome 0.999059608 times less likely.

* 1. Using a ROC curve, determine how well the final model predicts diabetes. Report the area under the curve as well as the ROC curve.

Final model predicts diabetes type slightly better than just age by itself. Age by itself has an AUC of 0.7333 while the final model has an AUC of 0.7633. In either case, the model fit is fair.

Chart, line chart

Description automatically generated

1. Now use recursive partitioning to develop a model to predict diabetes:
   1. Fit the initial model using all five predictors, and examine the cptable. Does the model need pruning? Explain why or why not. If yes, re-fit the pruned model.

Yes, the model needs pruning. At a glance, the unpruned plot looks way too cluttered. Further analysis confirms that the xerror is lowest for an nsplit of 2.

Diagram

Description automatically generated

Diagram

Description automatically generated

* 1. Using your final model from part a, predict the values of diabetes, and compare them with the values observed in the data. What is the overall accuracy (% agreement)?

20 + 56 = 76 correct predictions

(76 / 200) \* 100 = 38% accuracy



The dataset **diabetic** (from the survival package) contains results from a clinical trial of laser coagulation for treatment of diabetic retinopathy (the Diabetic Retinopathy Study, or DRS). Each patient had one eye randomized to laser treatment while the other eye received no treatment. The study outcome is visual acuity dropping below 5/200 for two consecutive visits. The time-to-event variable is time (in months), the censoring variable is status (0=censored, 1=visual loss).

1. Visualize the relationship between laser treatment and visual loss using a Kaplan-Meier plot.

There were two treatment types. No treatment (trt=0) is represented by the solid line. Laser treatment (trt=1) is represented by the dotted line. The laser treatment seemed to have a much higher probability of non-blindness than no treatment.

Chart, line chart

Description automatically generated

1. Calculate the median survival time for each treatment group and compare. Does laser treatment appear to delay visual loss?

As shown below, there is no median survival time for the laser treatment group, because their non-blindness probability does not drop past 50%. This is reflected in the Kaplan-Meier plot, with no treatment dropping below 50%, while laser treatment manages to stay above it. It certainly appears that laser treatment is able to delay blindness.

Table

Description automatically generated with medium confidence

1. Use a log-rank test to test the hypothesis that laser treatment affects time to visual loss.

The p-value is 2.584e-06, far below the 0.05 threshold for significance. This would indicate that laser treatment has a significant effect on time to visual loss.

Text

Description automatically generated

1. Create a variable called juvenile that is equal to 1 for patients under the age of 20, and equal to zero otherwise. Run a Cox proportional hazards model to determine if there is an interaction between juvenile diabetes and laser treatment. Does the effect of laser treatment differ for juveniles vs. adults (*i.e.,* does interaction exist)?

The p-value for the interaction between juvenile and treatment is 0.0159, below the 0.05 threshold of significance. We can therefore conclude that their interaction both exists and is significant. This is further confirmed by side-by-side graphs of adult versus juvenile, which show that the treatments effects do certainly differ between age groups.

A screenshot of a computer

Description automatically generated with low confidence

Chart, line chart

Description automatically generated