**Understanding Data and Statistical Design (60117)**

**Lab 12: Logistic Regression Model Fit**

This lab is marked from 24.

Please submit via Canvas.

**Due by the conclusion of the lab class**

This week we analyse the presence of high blood pressure in individuals using a multiple logistic regression model. The variables we consider are summarised in the table below.

|  |  |  |
| --- | --- | --- |
| **Name** | **Type** | **Description** |
|  | categorical | state of health: 1 (very good), 2 (good), 3 (average), 4 (poor), 5 (very poor) |
|  | continuous | age of individual |
|  | continuous | body mass index of individual |
|  | integer | presence of high blood pressure: 0 (no), 1 (yes) |

We will use to refer to the RV describing the population from which the sample was taken.

The total number of observations is .

The data is available in lab12.csv.

To represent in our model, we need 4 binary dummy variables that we will code as

**QUESTION 1 [12 marks].**

The first model we consider on the log-odds scale is

where

or if we wish to make the dependence on the predictors explicit

Fitting this model produced the following summary information.

Table

Description automatically generated

1. By generating an appropriate cross-tabulation and using 0.5 cut-off probability, calculate the overall prediction accuracy, the true positive rate (sensitivity) and true negative rate (specificity) of the fitted model’s prediction of **[3 marks]**.

A computer screen shot of a computer code

Description automatically generated

1. Explain how an ROC curve is constructed **[not assessed]**.

To construct a ROC curve, you first train a binary classification model and obtain the predicted probabilities. Then, various cut-off thresholds are set from 0 to 1. For each threshold, each observation is classified as positive or negative depending on whether the predicted probability is higher or lower than the threshold. The True Positive Rate (Sensitivity) and False Positive Rate are calculated for each threshold. Finally, the Sensitivity (Y-axis) is plotted against the False Positive Rate (X-axis) for all thresholds, forming the ROC curve.

1. Use R to produce an ROC curve for the fitted model and using this, classify the fit of the model using the criteria set by Hosmer and Lemeshow **[3 marks]**.

A graph of a curve

Description automatically generated

AUC: 0.858

According to the Hosmer and Lemeshow criteria, the fit of the model is excellent.

1. Calculate the pseudo Statistic based on the proportional change in deviance from the null model **[3 marks]**.

A number on a black background

Description automatically generated

Approximately 26.32% of the variability in the presence of high blood pressure is explained by the model.

1. Using significance level , document a test to determine if the fitted probabilities do not match the observed probabilities. Write down the hypotheses, the test statistic and p-value, the result of the test with reason and a conclusion in non-mathematical language **[3 marks]**.

A black background with white text

Description automatically generated

**Ho:** No difference between observed and predicted probabilities. The model fits the data well.

**Ha:** There is a significant difference between observed and predicted probabilities. The model does not fit the data well.

**Statistic:** 19.568

**P\_value:** 0.0121

**Decision:** Since the p\_value < , we rejec the null hypothesys.

**Conclusion:** we have sufficient evidence to conclude that the probabilities predicted by the model are significantly different from those observed. Therefore, we conclude that the model does not adequately fit the data.

**QUESTION 2 [12 marks].**

Now we extend the model from Q1 by including interaction between and . On the log-odds scale we fit

where

or if we wish to make the dependence on the predictors explicit

R produced the following summary information for the fitted model.

Table

Description automatically generated

1. By generating an appropriate cross-tabulation, calculate the overall prediction accuracy and use this criteria to determine if the fit of the Q2 model is superior or inferior to that of Q1 **[3 marks]**.

A screenshot of a computer

Description automatically generated

**Accuracy:** Model 2 is slightly better.

**Specificity:** Model 2 is slightly better.

**Sensitivity:** Model 1 is slightly better.

The model with interaction offers a slight improvement in some metrics, but the difference is not substantial.

1. Generate an ROC curve for the Q2 model and use this criteria to determine if the fit of the Q2 model is superior or inferior to that of Q1 **[3 marks]**.

A graph of a curve

Description automatically generated

AUC: 0.859, according to the Hosmer and Lemeshow criteria, the fit of the model is excellent. However, compared to model 1, there is no significant improvement.

1. For the Q2 model, calculate the pseudo statistic based on the change in deviance from the null model and use this criteria to determine if the fit of the Q2 model is superior or inferior to that of Q1 **[3 marks]**.

: 0.2651391

The Q2 model fits the data slightly better than the Q1 model based on the pseudo R². But it is not a significant improvement

1. Using significance level , document a test to determine if the fitted probabilities do not match the observed probabilities. Write down the hypotheses, the test statistic and p-value, the result of the test with reason and a conclusion in non-mathematical language **[3 marks]**.

A black background with white text

Description automatically generated

**Ho:** No difference between observed and predicted probabilities. The model fits the data well.

**Ha:** There is a significant difference between observed and predicted probabilities. The model does not fit the data well.

**Statistic:** 10.184

**P\_value:** 0.2524

**Decision:** Since the p\_value > , we sustain the null hypothesys.

**Conclusion:** The probabilities predicted by the model are not significantly different from the observed probabilities, suggesting that the model fits the data adequately.

1. Using significance level , document a test to determine if the interaction term is significant. Write down the hypotheses, the test statistic and p-value, the result of the test with reason and a conclusion in non-mathematical language **[not assessed]**.