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

**Lab 10: Simple Logistic Regression**

This lab is marked from 24.

Please submit via Canvas.

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

This week we continue the analysis of the history of cardiovascular disease in individuals with simple logistic regression models. The variables we consider are summarised in the table below.

|  |  |  |
| --- | --- | --- |
| **Name** | **Type** | **Description** |
|  | categorical | state of general health: 1 (very good), 2 (good), 3 (average), 4 (poor), 5 (very poor) |
|  | continuous | age of individual (years) |
|  | categorical | history of cardiovascular disease: 0 (no), 1 (yes) |

The data is from a British Panel Survey of consisting of the responses of 3916 individuals (available in lab10.csv on Canvas).

**QUESTION 1 [12 marks]. Continuous predictor**

We begin by modelling as a function of .

The population model is

where

or

if we wish to make the dependence on explicit.

1. Fit the logistic regression model and write down the fitted model equation in log-odds scale, odds scale and probability scale **[3 marks]**.

A computer screen with numbers and symbols

Description automatically generated

**Log-odds scale:**

**Odds scale:**

=

**Probability scale:**

1. Provide interpretations of the estimated parameters on the log-odds scale **[3 marks]**.

**Intercept:** β0 = -3.73

This is the representation of the baseline log-odds, of having a CVD at age 0, indicating a low likelihood of having a CVD at young ages.

**Age coefficient:** β1 =

This indicates that with each additional year of age, the logs-odds of having a CVD increase by 0.050926.

1. Provide interpretations of the estimated parameters on the odds scale **[3 marks]**.

**Intercept:**

This represents the odds of having a CVD at age 0, indicating a low likelihood of having a CVD at young ages, establishing a baseline of the CVD occurrence at young ages.

**Age coefficient:**

This represents the odds of having CVD, which increase by a factor of about 1.052 each year.

1. Using significance level , document a test to determine if the fitted logistic regression model 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]**.
2. Use a scatterplot to plot against and describe the relationship **[not assessed]**.

A graph of a graph

Description automatically generated with medium confidence

1. Using the rule

determine if the fitted regression model predicts a history of cardiovascular disease for those with and for those with **[3 marks]**.

A black screen with white text

Description automatically generated

**Age 35:** p=0.1239498 ≤ 0.5, the prediction is: there is no history of CVD.

**Age 75:** p=0.520366 ≥ 0.5, the prediction is: there is history of CVD.

**QUESTION 2 [12 marks]. Categorical predictor**

Now we look at modelling as a function of .  
  
To represent this in our model, we need 4 binary dummy variables that we will code as

The population model is

where

or

if we wish to make the dependence on explicit.

1. Write down the fitted logistic regression model in log-odds scale, odds scale and probability scale **[3 marks]**.

A screenshot of a computer

Description automatically generated

**Log-odds scale:**

**Odds scale:**

**Probability scale:**

1. Using the fitted logistic regression model, calculate the odds of having a history of cardiovascular disease () for those with average health (), very good health () and the odds ratio of these with very good health as reference. Compare these to the calculations carried out using the cross tabulation in Lab 9 **[3 marks]**.

A computer screen with text

Description automatically generated

1. Provide interpretations of the estimated parameters and on the log-odds scale **[3 marks]**.

**Intercept:** β0 =

This represents the log odds of having a CVD when all the dummy variables are equal to 0, representing the base category (healt1, Very good health). This value indicates that the log-odds of having CVD are low.

**Age coefficient:** β3 =

This represents the log odds of having a CVD when having a average health (health3). The value suggests that the log-odds of developing CVD are higher for those with average health.

1. Provide interpretations of the estimated parameters and on the odds scale **[3 marks]**.

**Intercept:** β0 =

This represents the log odds of having a CVD when all the dummy variables are equal to 0, representing the base category (healt1, Very good health). This value indicates that the odds of having CVD for those in very good health are relatively low, about 0.1437.

**Age coefficient:** β3 = = 4.313

This represents the log odds of having a CVD when having a average health (health3). The value suggests that the odds of having CVD for those with average health are 4.313 times higher than for those in very good health.

1. Using the rule

determine if the fitted regression model predicts a history of cardiovascular disease for those with very good health and for those with average health **[not assessed]**.