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
$health$	categorical	state of general health: 1 (very good), 2 (good), 3 (average), 4 (poor), 5 (very poor)
age	continuous	age of individual (years)
cvd	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 cvd as a function of age.

The population model is

$$CVD = p + \epsilon = \frac{1}{1 + \exp\left(-\beta_0 - \beta_1 * age\right)} + \epsilon$$

where

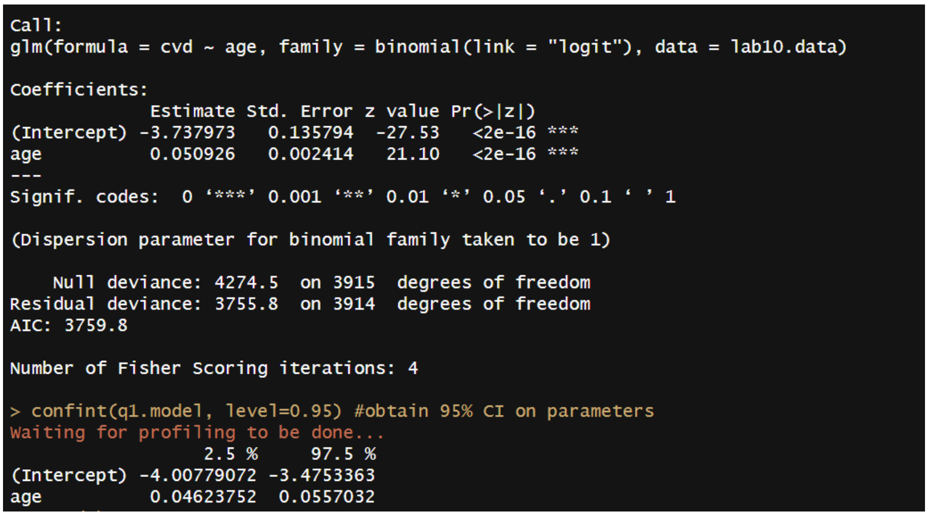
$$p = \text{Prob}(CVD = 1).$$

or

$$p(CVD) = \text{Prob}(CVD = 1|age).$$

if we wish to make the dependence on age explicit.

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



```
Call:
glm(formula = cvd ~ age, family = binomial(link = "logit"), data = lab10.data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.737973   0.135794  -27.53   <2e-16 ***
age          0.050926   0.002414   21.10   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4274.5  on 3915  degrees of freedom
Residual deviance: 3755.8  on 3914  degrees of freedom
AIC: 3759.8

Number of Fisher Scoring iterations: 4

> confint(q1.model, level=0.95) #obtain 95% CI on parameters
Waiting for profiling to be done...
                  2.5 %     97.5 %
(Intercept) -4.00779072 -3.4753363
age          0.04623752  0.0557032
```

Log-odds scale:

$$\text{Log(p)} = \beta_o + \beta_1 * \text{age} = -3.737973 + 0.050926 * age$$

Odds scale:

$$\frac{p}{1-p} = e^{(\beta_o + \beta_1 * \text{age})} = e^{(-3.737973+0.050926*age)}$$

Probability scale:

$$p = \frac{1}{1 + e^{(3.737973-0.050926*age)}}$$

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

Intercept: $\beta_0 = -3.73$

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: $\beta_1 = 0.050926$

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

(c) Provide interpretations of the estimated parameters on the odds scale **[3 marks]**.

Intercept: $e^{(-3.737973)} \approx\ 0.0238023$

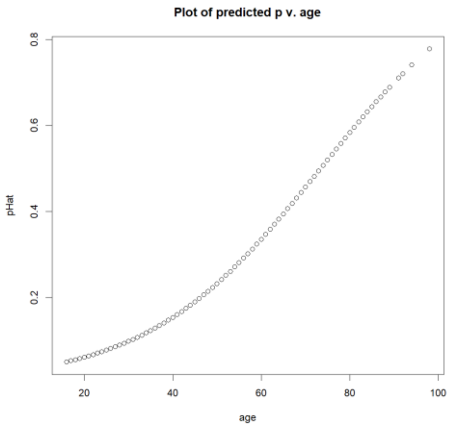
This represents the odds of having a CVD at age 0, indicating a low likelihood of having a CVD at young ages, establishing a baseline of the CVD occurrence at young ages.

Age coefficient: $e^{(0.050926\ )} \approx\ 1.05224$

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

(d) Using significance level $\alpha = 0.05$, 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]**.

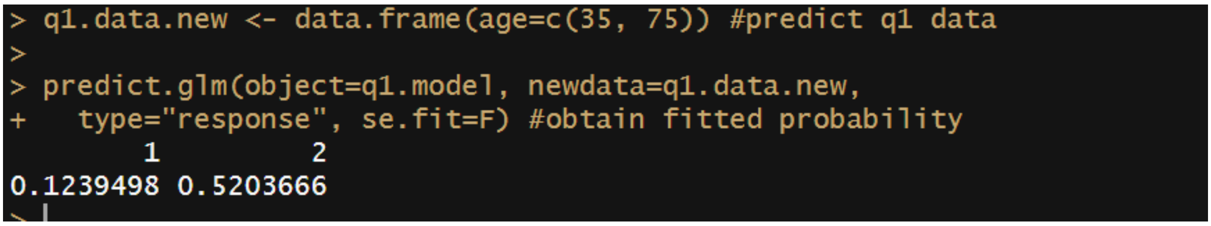
(e) Use a scatterplot to plot $\hat{p}$ against age and describe the relationship **[not assessed]**.



(f) Using the rule

$$\widehat{cvd} = \begin{cases} 0, & \hat{p} \leq 0.5 \\ 1, & \hat{p} > 0.5' \end{cases}$$

determine if the fitted regression model predicts a history of cardiovascular disease for those with $age = 35$ and for those with $age = 75$ **[3 marks]**.



```
> q1.data.new <- data.frame(age=c(35, 75)) #predict q1 data
>
> predict.glm(object=q1.model, newdata=q1.data.new,
+   type="response", se.fit=F) #obtain fitted probability
        1         2 
0.1239498 0.5203666 
>
```

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 cvd as a function of $health$.

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

$$(health2, health3, health4, health5) = \begin{cases} (0,0,0,0) & health = 1 \\ (1,0,0,0) & health = 2 \\ (0,1,0,0) & health = 3. \\ (0,0,1,0) & health = 4 \\ (0,0,0,1) & health = 5 \end{cases}$$

The population model is

$$CVD = p + \epsilon$$

$$= \frac{1}{1 + \exp\left(-\beta_0 - \beta_2 * health2 - \beta_3 * health3 - \beta_4 * health4 - \beta_5 * health5\right)} + \epsilon$$

where

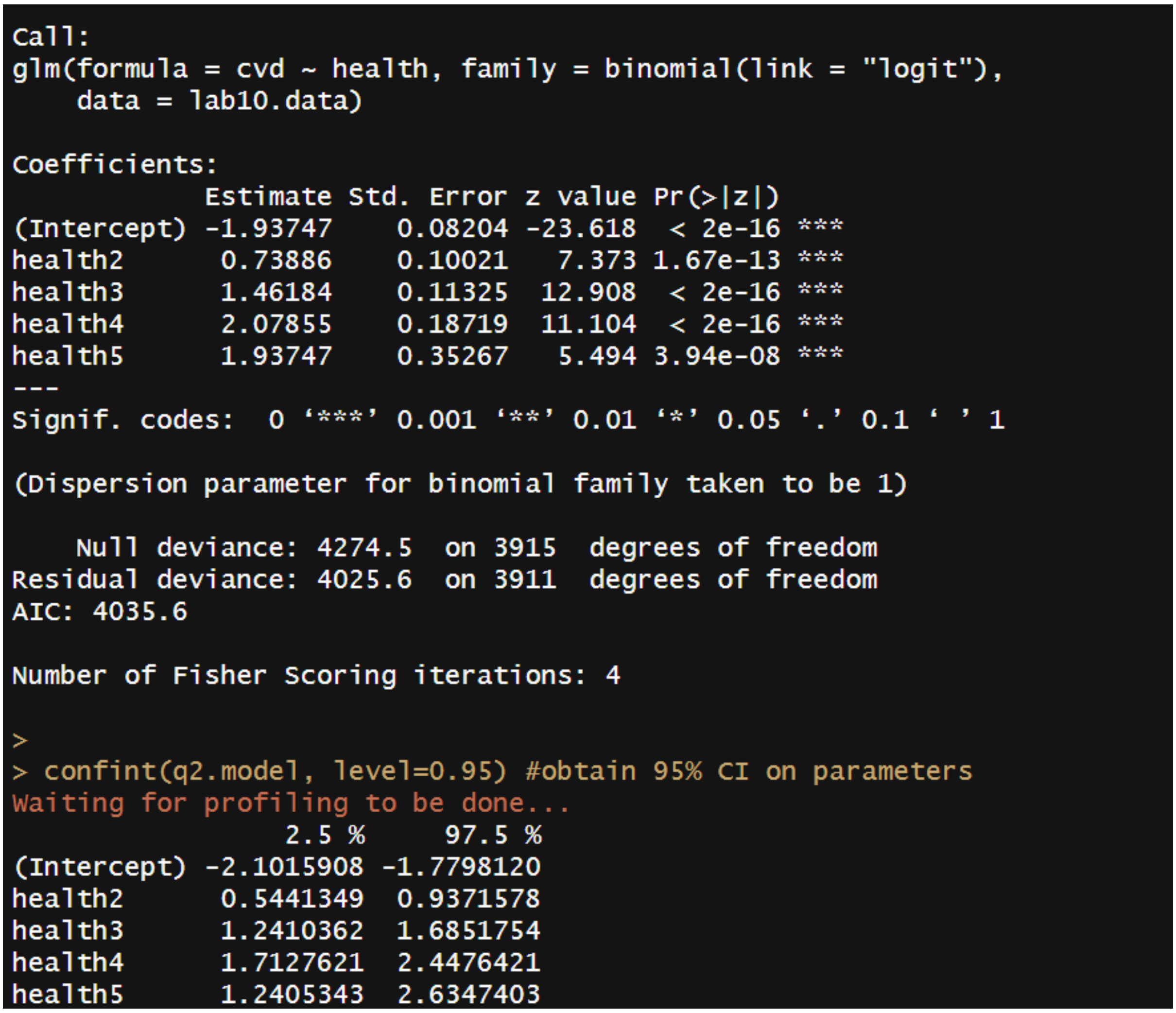
$$p = \text{Prob}(CVD = 1).$$

or

$$p(health) = \text{Prob}(CVD = 1|health).$$

if we wish to make the dependence on $health$ explicit.

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



```
Call:
glm(formula = cvd ~ health, family = binomial(link = "logit"),
    data = lab10.data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.93747    0.08204 -23.618  < 2e-16 ***
health2      0.73886    0.10021   7.373 1.67e-13 ***
health3      1.46184    0.11325  12.908  < 2e-16 ***
health4      2.07855    0.18719  11.104  < 2e-16 ***
health5      1.93747    0.35267   5.494 3.94e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4274.5  on 3915  degrees of freedom
Residual deviance: 4025.6  on 3911  degrees of freedom
AIC: 4035.6

Number of Fisher Scoring iterations: 4

>
> confint(q2.model, level=0.95) #obtain 95% CI on parameters
Waiting for profiling to be done...
                 2.5 %     97.5 %
(Intercept) -2.1015908 -1.7798120
health2      0.5441349  0.9371578
health3      1.2410362  1.6851754
health4      1.7127621  2.4476421
health5      1.2405343  2.6347403
```

Log-odds scale:

$$\mathrm{Log(p)} = \beta_0 + \beta_2 * health2 + \beta_3 * health3 + \beta_4 * health4 + \beta_5 * health$$
$$= -1.93747 + 0.73886 * health2 + 1.46184 * health3 + 2.07855 * health4 + 1.93747 * health5$$

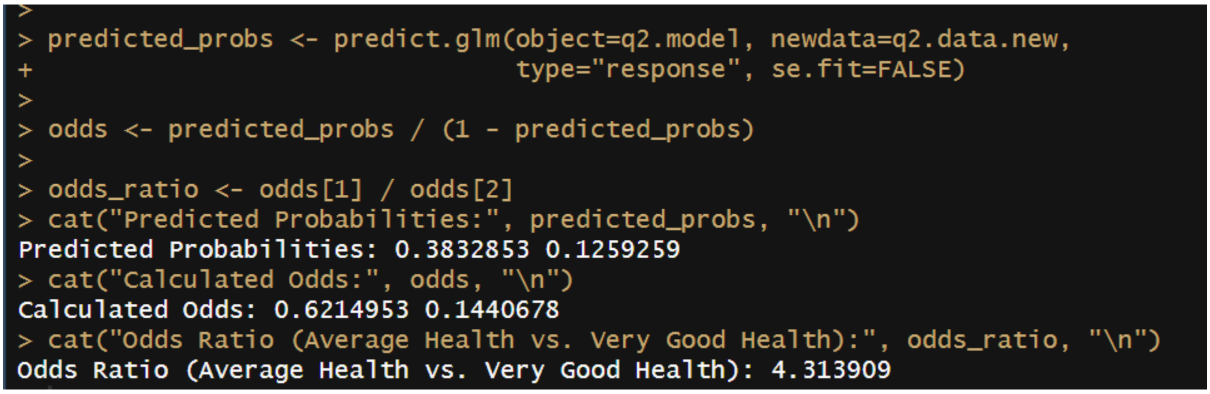
Odds scale:

$$\frac{p}{1-p} = e^{(\beta_0+\beta_2*health2+\beta_3*health3+\beta_4*health4+\beta_5*health)} = e^{(-1.93747+0.73886*\mathrm{health2}+1.46184*\mathrm{health3}+2.07855*\mathrm{health4}+1.93747*\mathrm{health5})}$$

Probability scale:

$$p = \frac{1}{1 + e^{(1.93747-0.73886*\mathrm{health2}-1.46184*\mathrm{health3}-2.07855*\mathrm{health4}-1.93747*\mathrm{health5})}}$$

(b) Using the fitted logistic regression model, calculate the odds of having a history of cardiovascular disease ($cvd = 1$) for those with average health ($health = 3$), very good health ($health = 1$) 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]**.



```
>
> predicted_probs <- predict.glm(object=q2.model, newdata=q2.data.new,
+                                type="response", se.fit=FALSE)
>
> odds <- predicted_probs / (1 - predicted_probs)
>
> odds_ratio <- odds[1] / odds[2]
> cat("Predicted Probabilities:", predicted_probs, "\n")
Predicted Probabilities: 0.3832853 0.1259259
> cat("Calculated Odds:", odds, "\n")
Calculated Odds: 0.6214953 0.1440678
> cat("Odds Ratio (Average Health vs. Very Good Health):", odds_ratio, "\n")
Odds Ratio (Average Health vs. Very Good Health): 4.313909
```

(c) Provide interpretations of the estimated parameters $\hat{\beta}_0$ and $\hat{\beta}_3$ on the log-odds scale **[3 marks]**.

Intercept: $\beta_0 = -1.93747$

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: $\beta_3 = 1.46184$

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.

(d) Provide interpretations of the estimated parameters $\hat{\beta}_0$ and $\hat{\beta}_3$ on the odds scale **[3 marks]**.

Intercept: $\beta_0 = e^{-1.93747} = 0.1437$

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: $\beta_3 = e^{1.46184} = 4.313$

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.

(e) Using the rule

$$\widehat{cvd} = \begin{cases} 0, & \hat{p} \leq 0.5 \\ 1, & \hat{p} > 0.5' \end{cases}$$

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]**.