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
h	categorical	state of health: 1 (very good), 2 (good), 3 (average), 4 (poor), 5 (very poor)
age	continuous	age of individual
bmi	continuous	body mass index of individual
bp	integer	presence of high blood pressure: 0 (no), 1 (yes)

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

The total number of observations is $N = 4745$.

The data is available in lab12.csv.

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

$$(h_2, h_3, h_4, h_5) = \begin{cases} (0,0,0,0) & h = 1 \\ (1,0,0,0) & h = 2 \\ (0,1,0,0) & h = 3. \\ (0,0,1,0) & h = 4 \\ (0,0,0,1) & h = 5 \end{cases}$$

QUESTION 1 [12 marks].

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

$$\ln\frac{p}{1-p} = \beta_0 + \gamma_2 h_2 + \gamma_3 h_3 + \gamma_4 h_4 + \gamma_5 h_5 + \beta_b bmi + \beta_a age$$

where

$$p = \text{Prob}(BP = 1)$$

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

$$p(bmi, age, h_2, h_3, h_4, h_5) = \text{Prob}(BP = 1 | bmi, age, h_2, h_3, h_4, h_5).$$

Fitting this model produced the following summary information.

```
Call:
glm(formula = bp ~ bmi + h + age, family = binomial(link = "logit"),
    data = lab12.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7470  -0.4516  -0.2320  -0.1284   3.1320

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.571870   0.390632 -21.944  < 2e-16 ***
bmi          0.067191   0.010867   6.183 6.28e-10 ***
h2           0.111093   0.126429   0.879   0.3796
h3          -0.032439   0.145795  -0.222   0.8239
h4          -0.119156   0.237372  -0.502   0.6157
h5          -0.923572   0.482471  -1.914   0.0556 .
age          0.086311   0.003684  23.429  < 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: 3543.8  on 4744  degrees of freedom
Residual deviance: 2610.9  on 4738  degrees of freedom
AIC: 2624.9

Number of Fisher Scoring iterations: 6
```

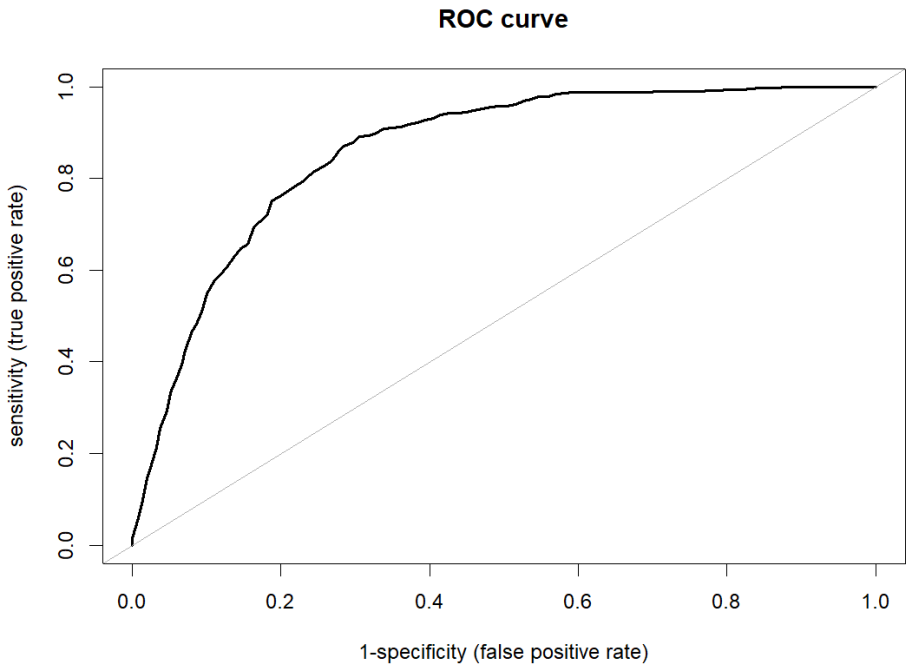
(a) By generating an appropriate cross-tabulation and using 0.5 cut-off probability, calculate the overall prediction accuracy, the true positive rate (sensitivity) and true negative rate (specificity) of the fitted model's prediction of *BP* **[3 marks]**.

```
> conf_matrix1 <- table(Predicted = lab12.data$bpHat1, Actual = lab12.data$bp)
>
> # Calcular la precisión
> accuracy1 <- sum(diag(conf_matrix1)) / sum(conf_matrix1)
>
> # Calcular la sensibilidad (True Positive Rate)
> sensitivity1 <- conf_matrix1[2,2] / sum(conf_matrix1[,2])
>
> # Calcular la especificidad (True Negative Rate)
> specificity1 <- conf_matrix1[1,1] / sum(conf_matrix1[,1])
>
> # Mostrar resultados
> accuracy1
[1] 0.8756586
> sensitivity1
[1] 0.157265
> specificity1
[1] 0.9766827
```

(b) 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.

(c) 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]**.



AUC: 0.858

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

(d) Calculate the pseudo R^2Statistic based on the proportional change in deviance from the null model **[3 marks]**.

```
>
> (3543.8 - 2610.9) / 3543.8 #pseudo R^2
[1] 0.2632485
```

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

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

```
        Hosmer and Lemeshow goodness of fit (GOF) test

data:  lab12.data$bp, lab12.data$pHat1
X-squared = 19.568, df = 8, p-value = 0.0121
```

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 < α, 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 age and h. On the log-odds scale we fit

$$\ln\frac{p}{1-p} = \beta_0 + \gamma_2 h_2 + \gamma_3 h_3 + \gamma_4 h_4 + \gamma_5 h_5 + \beta_b bmi$$
$$+\beta_a age + (\delta_2 h_2 + \delta_3 h_3 + \delta_4 h_4 + \delta_5 h_5) age$$

where

$$p = \text{Prob}(BP = 1)$$

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

$$p(bmi, age, h_2, h_3, h_4, h_5) = \text{Prob}(BP = 1 | bmi, age, h_2, h_3, h_4, h_5).$$

R produced the following summary information for the fitted model.

```
Call:
glm(formula = bp ~ bmi + h * age, family = binomial(link = "logit"),
    data = lab12.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7160  -0.4546  -0.2298  -0.1241   3.1037

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.224050   0.540717 -17.059  < 2e-16 ***
bmi          0.066630   0.010874   6.127 8.94e-10 ***
h2           0.914402   0.560460   1.632   0.1028
h3           1.270287   0.678352   1.873   0.0611 .
h4           1.301252   1.325832   0.981   0.3264
h5          -7.338796   6.034035  -1.216   0.2239
age          0.097457   0.007214  13.509  < 2e-16 ***
h2:age      -0.013345   0.008965  -1.489   0.1366
h3:age      -0.020739   0.010463  -1.982   0.0475 *
h4:age      -0.022326   0.019752  -1.130   0.2583
h5:age       0.083277   0.078836   1.056   0.2908
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3543.8  on 4744  degrees of freedom
Residual deviance: 2604.2  on 4734  degrees of freedom
AIC: 2626.2

Number of Fisher Scoring iterations: 7
```

(a) 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]**.

```
Total Observations in Table:  4745


             | lab12.data$bpHat2
lab12.data$bp |         0 |         1 | Row Total |
--------------|-----------|-----------|-----------|
            0 |      4074 |        86 |      4160 |
              |     0.979 |     0.021 |     0.877 |
              |     0.859 |     0.018 |           |
--------------|-----------|-----------|-----------|
            1 |       499 |        86 |       585 |
              |     0.853 |     0.147 |     0.123 |
              |     0.105 |     0.018 |           |
--------------|-----------|-----------|-----------|
 Column Total |      4573 |       172 |      4745 |
--------------|-----------|-----------|-----------|


> conf_matrix2 <- table(Predicted = lab12.data$bpHat2, Actual = lab12.data$bp)
> accuracy2 <- sum(diag(conf_matrix2)) / sum(conf_matrix2)
> sensitivity2 <- conf_matrix2[2,2] / sum(conf_matrix2[,2])
> specificity2 <- conf_matrix2[1,1] / sum(conf_matrix2[,1])
> accuracy2
[1] 0.8767123
> sensitivity2
[1] 0.1470085
> specificity2
[1] 0.9793269
```

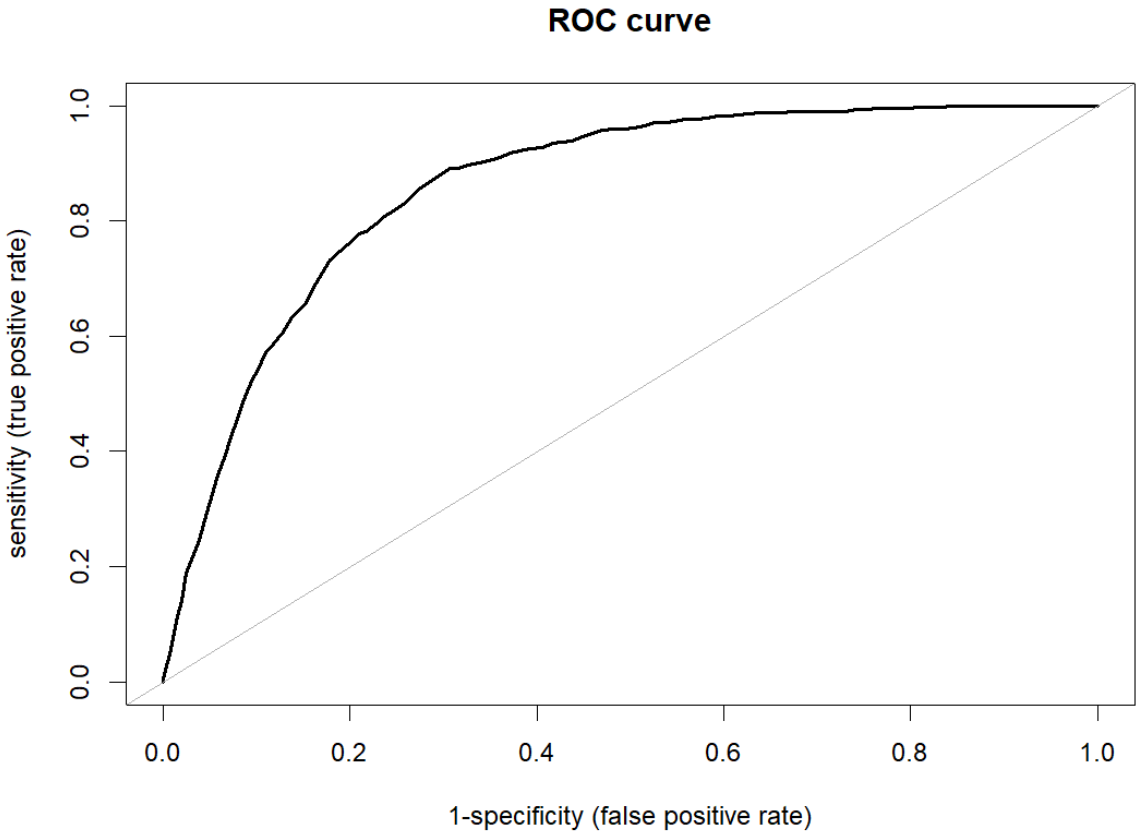
Accuracy: Model 2 is slightly better.

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.

(b) 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]**.



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.

(c) For the Q2 model, calculate the pseudo R^2 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]**.

R^2: 0.2651391

The Q2 model fits the data slightly better than the Q1 model based on the pseudo R^2. But it is not a significant improvement

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

```
        Hosmer and Lemeshow goodness of fit (GOF) test

data:  lab12.data$bp, lab12.data$pHat2
X-squared = 10.184, df = 8, p-value = 0.2524
```

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 > α, 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.

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