Chapter 8

Interpreting a Logistic Regression Model

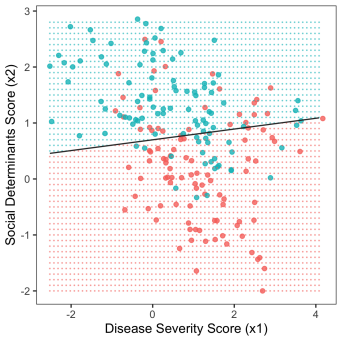
This chapter is similar to Chapter 7 but focuses on logistic regression models. As we saw in Chapter 7, linear regression models are used in situations where the outcome of a supervised learning problem, y, follows a normal distribution, conditional on the values of the predictors. **Logistic regression** models, in contrast, handle situations where the outcome, y, is binary: either 0 or 1. We first encountered these models as examples of classification algorithms in Chapter 1. Because of their popularity in the clinical domain, it's important to understand how these models are fit and how to interpret the summary output produced by software.

Unfortunately, a full understanding of logistic regression requires knowledge of maximum likelihood estimation. We will, therefore, skip over some of the details until we've had more time to explore this topic.

8.1 ER Readmissions Example from Chapter 1

In Chapter 1, we saw an example where information about two predictors – a disease severity score (x_1) and a social determinants score (x_2) – was used to predict a binary outcome: whether a patient would be readmitted to the ER within 30 days of discharge. We tried three different supervised learning

algorithms, one of which was a logistic regression model (Section 1.3.1). The output from that model is repeated below.




```
Call:
glm(formula = y ~ x1 + x2, family = "binomial", data = df)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.88232  -0.90614  -0.05965   0.86579   2.28489

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.9780     0.2945   3.321 0.000897 ***
x1            0.1344     0.1372   0.980 0.327272
x2           -1.3981     0.2316  -6.035 1.59e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 277.26  on 199  degrees of freedom
Residual deviance: 209.54  on 197  degrees of freedom
AIC: 215.54

Number of Fisher Scoring iterations: 4
```

8.2 Understanding the Model Summary

A logistic regression model looks like this (see also Section 1.3.1):

$$\log \frac{\mu}{1-\mu} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{8.1}$$

where μ is the mean of the Bernoulli distribution (see Section 3.3) governing our binary outcome, y; in other words, it is the probability that $y = 1$.

You will note that there is no independent error term here as there was in linear regression. That's because the variance and mean of a Bernoulli distribution are coupled and depend only on μ (again, see Section 3.3).

Question 8.1

In logistic regression, μ itself is not equal the sum of the predictors; instead, the **logit** of μ is their sum. Based on what you know about μ, why is a logistic

regression model not of the form

$$\mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p?$$

We will explore this further in Chapter 12.

Question 8.2

The decision boundary in logistic regression (see picture above) occurs where the sum of the linear predictors, $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$, is zero. What value of μ does this correspond to? Why does this make sense, intuitively?

8.2.1 The Call

The first line of the output repeats the call you made to the `glm` function in R to fit the model. The `glm` package fits a variety of different generalized linear models using maximum likelihood estimation (see Chapter 4; this will also be discussed in more detail in Chapter 12). The `family = "binomial"` argument tells the function to fit a logistic regression model.

8.2.2 Coefficients and Standard Errors

Logistic regression models, like other GLMs, are fit using maximum likelihood (see Chapter 4 for a brief introduction). We will skip most of the details for now, but you can gain intuition by staring at Equation 9.1. This equation says that the model's predicted value of μ, the probability that the outcome will be positive ($y = 1$), is controlled by the values of the predictors and their coefficients $\beta_0, \ldots, \beta_p$.

By adjusting the values of the βs, the model causes μ to be high in regions of the feature space where $y = 1$ and low where $y = 0$. The values of the βs that do this the best are called the **maximum likelihood estimates**, and they are the coefficients shown in the model output.

As with linear regression, a full understanding of the standard errors requires matrix multiplication. However, they are related to the same factors that drive the standard errors in linear regression: (1) the spread of the values

of the corresponding covariate about its mean (more spread will decrease the standard error) and (2) correlations between that covariate and other covariates in the model (tighter correlations will increase the standard error).

Question 8.3

Looking at the form of the logistic regression model

$$\log \frac{\mu}{1-\mu} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

what does the value of each of the βs mean? What is β_j telling us about how y varies with the predictor j, all else being equal?

Question 8.4

The **odds** of something happening are defined as $\mu/(1-\mu)$, where μ is the probability that the thing occurs. In our example model, we are interested in the odds that $y = 1$ (the patient is readmitted). Does a unit increase in x_1 (disease severity score) increase or decrease the odds that a patient will be readmitted? What about x_2 (social determinants score)?

Question 8.5

What are the odds of readmission for a patient with:

(a) $x_1 = 0.1$ and $x_2 = 0.3$?

(b) $x_1 = 0.1$ and $x_2 = -1.3$?

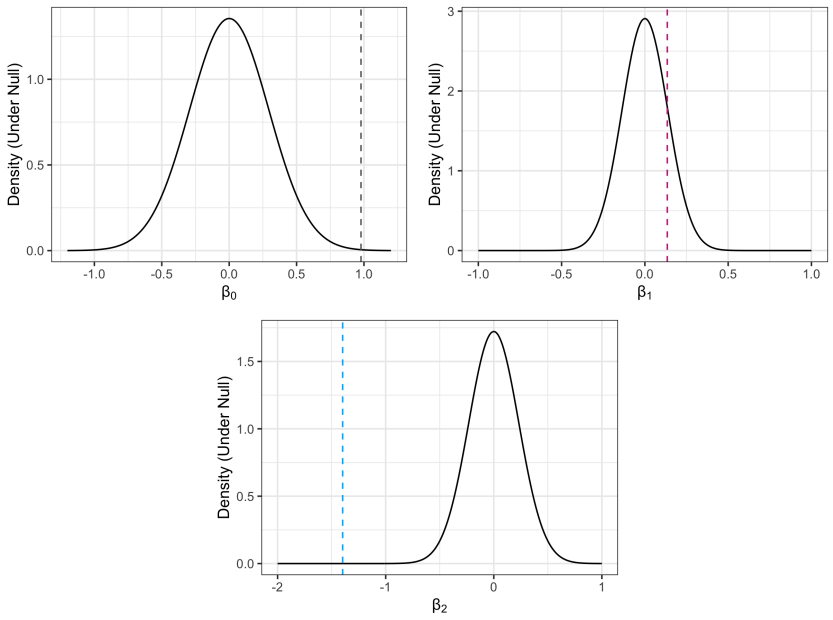
(c) $x_1 = 1.1$ and $x_2 = 0.3$?

8.2.3 Hypothesis Tests of Coefficients

Just as in linear regression, we can use our coefficients and standard errors to perform a hypothesis test on each regression coefficient, β_j, against the null hypothesis that $\beta_j = 0$ (the predictor x_j has no effect on the outcome). In logistic regression, the quantity $\hat{\beta}_j / \text{se}(\hat{\beta}_j)$ will follow a normal distribution under the null.

Question 8.6

Below are the null distributions for the hypothesis tests of our three regression coefficients, β_0, β_1, and β_2. In each graph, the maximum likelihood estimate of the coefficient is shown as a vertical dashed line. Based on these graphs, can you tell why the p-values for β_0 and β_2 are low and the one for β_1 is high? What is the intuition behind this?



8.2.4 Deviance and Deviance Residuals

The **deviance** (called **residual deviance** in the model output) plays a role in GLMs akin to that of the residual standard error in linear regression; it is a measure of the residual variation in the outcome not explained by the model. The **null deviance** is the deviance for a model that only includes an intercept. Under the null hypothesis that all of the βs are zero except the intercept (i.e, a model with no predictors explains the data as well as our model), the difference

$$\text{null deviance} - \text{residual deviance}$$

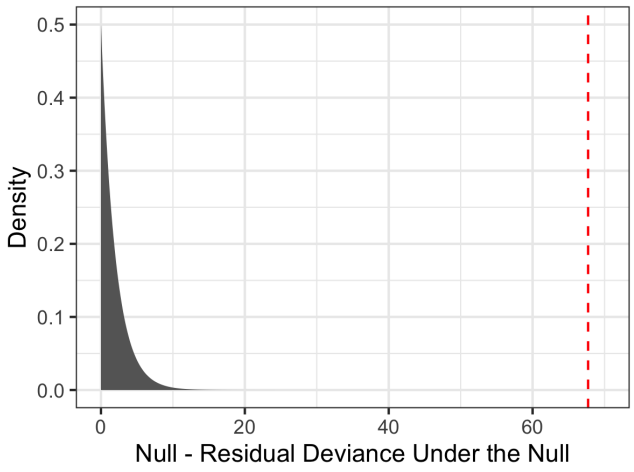
is distributed as χ^2_p, a chi-squared distribution (see Section 3.8) with p degrees of freedom, where p is the number of predictors.

Question 8.7

This test is a hypothesis test of the null hypothesis that a model with no predictors fits our data as well as our model, where goodness of fit is measured by the deviance (lower is better). What is this hypothesis test akin to in the linear regression model output?

Question 8.8

The difference in null and residual deviances in this case is 67.72. It follows a χ^2_2 distribution under the null. A plot of the χ^2_2 distribution and our test statistic is shown below. What do these findings indicate about the p-value of this goodness of fit test and what does it mean?



In the GLM context, there are multiple types of residual (more on this later). **Deviance residuals** quantify the contributions of the individual samples to the deviance. Unfortunately, the output from `glm` is confusing because what `glm` calls a deviance residual in the model summary is actually something called a **working residual**. We will ignore this part of the output until we understand more about the inner workings of GLMs.

8.3 Example: Low Birthweight Dataset

The goal of this study was to identify risk factors associated with giving birth to a low birth weight baby (a baby weighing less than 2500 grams). Infant mortality rates and birth defect rates are very high for low birth weight babies. A woman's behavior during pregnancy (including diet, smoking habits, and receiving prenatal care) can greatly alter the chances of carrying the baby to term and, consequently, of delivering a baby of normal birth weight.

Data were collected on 189 women, 59 of which had low birth weight babies and 130 of which had normal birth weight babies[1].

LOW	Low birth weight (0 = birth weight $\geq$ 2500 g; 1 = birth weight $<$ 2500 g)
AGE	Age of mother in years
LWT	Mother's weight in pounds at last menstrual period
RACE	Race (1 = white, 2 = black, 3 = other)
SMOKE	Smoking status during pregnancy (1 = yes, 0 = no)
PTL	History of premature labor (0 = none, 1 = one, etc.)
HT	History of hypertension (0 = no, 1 = yes)
UI	Presence of uterine irritability (0 = no, 1 = yes)
FTV	Number of physician visits during the first trimester
BWT	Birth weight in grams

We will build a model that predicts the value of LOW based on all of the other covariates except, of course, BWT. (Why not use BWT?)

[1]SOURCE: Hosmer and Lemeshow (2000) *Applied Logistic Regression: Second Edition*. Data were collected at Baystate Medical Center, Springfield, Massachusetts during 1986.

```
Call:
glm(formula = LOW ~ AGE + LWT + RACE + SMOKE + PTL + HT + UI +
    FTV, family = "binomial", data = d)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8946  -0.8212  -0.5316   0.9818   2.2125

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.480623   1.196888   0.402  0.68801
AGE         -0.029549   0.037031  -0.798  0.42489
LWT         -0.015424   0.006919  -2.229  0.02580 *
RACE2        1.272260   0.527357   2.413  0.01584 *
RACE3        0.880496   0.440778   1.998  0.04576 *
SMOKE        0.938846   0.402147   2.335  0.01957 *
PTL          0.543337   0.345403   1.573  0.11571
HT           1.863303   0.697533   2.671  0.00756 **
UI           0.767648   0.459318   1.671  0.09467 .
FTV          0.065302   0.172394   0.379  0.70484
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 201.28  on 179  degrees of freedom
AIC: 221.28

Number of Fisher Scoring iterations: 4
```

Question 8.9

In this model, is the effect of one predictor (say, `AGE`) impacted by the value(s) of any of the other predictor(s)? How does this differ from the other classification algorithms we've seen (KNN and decision trees)? What are the advantages and disadvantages of this choice?

Question 8.10

Comment on how the variable `RACE` enters into the model here. Does this make sense in light of what that variable means and how it potentially interacts with the other study variables?

Question 8.11

Interpret the values of each of these coefficients. Based on the coefficient values and their standard errors, which predictor(s) do you think have the greatest impact on whether or not a woman has a low birthweight baby?