4.3 Logistic Regression

Dr. Lauren Perry

Logistic Regression

Logistic regression will allow us to examine the case where the outcome variable has two distinct categories.

- The outcome Y will be constructed as a dummy variable.
- Rather than modeling the response directly, logistic regression models the *probability* of falling into a particular category.
- Then we might use, say, $\hat{Y} = 0.5$ as the cutoff point for predicting one category or the other.

The Logistic Model

How do we model the relationship between this *probability* and the outcome X?

$$p(X) = \beta_0 + \beta_1 X$$

- We want to avoid the potential to predict probabilities outside of $[0, 1]$.
- In logistic regression, we use the *logistic function*

$$p(X) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}$$

Odds

- We can rewrite that last eqn to find

$$\frac{p(X)}{1-p(X)} = \exp(\beta_0 + \beta_1 X)$$

  - The quantity $p(X)/[1-p(X)]$ is called the *odds*
- Odds can take on any values on $[0, \inf)$.
- Ex: With an odds of $1/4$, on avg 1 in 5 people will default.
  - $p(X) = 0.2 \rightarrow 0.2/(1-0.2) = 1/4$
- Ex: With an odds of 9, on average 9 in 10 people will default.
  - $p(X) = 0.9 \rightarrow 0.9/(1-0.9) = 9$

The Logistic Model

$$p(X) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}$$

can be written as

$$\frac{p(X)}{1 - p(X)} = \exp(\beta_0 + \beta_1 X)$$

and finally as

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X$$

where the LHS is called the *log odds* or *logit* function.

The Logistic Model

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

- When we run a logistic regression, predicted values are *log odds*.
- This makes interpretation a little more involved!
- If we increase X by one unit, the log odds change by β_1.
  - Or we can say it multiplies the odds by e^{β_1}.

Estimating the Coefficients

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

Here, we use a method called *maximum likelihood* to estimate the coefficients.

Idea: want to find estimates for β_0 and β_1 and such that the predicted probability $\hat{p}(x_i)$ for each observation corresponds as closely as possible to the individual's observed default status.

We are not going to discuss the mathematical details of MLE, but will note that least squares is a special case of it.

Example: Logistic Regression in R

```
glm(default ~ balance + income, family=binomial, data=Default)
```

```
##
## Call:  glm(formula = default ~ balance + income, family = binomial,
##     data = Default)
##
## Coefficients:
## (Intercept)      balance       income
##  -1.154e+01    5.647e-03    2.081e-05
##
## Degrees of Freedom: 9999 Total (i.e. Null);  9997 Residual
## Null Deviance:       2921
## Residual Deviance: 1579  AIC: 1585
```

```
##
## Call:
## glm(formula = default ~ balance + income, family = binomial,
##     data = Default)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.154e+01  4.348e-01 -26.545  < 2e-16 ***
## balance      5.647e-03  2.274e-04  24.836  < 2e-16 ***
## income       2.081e-05  4.985e-06   4.174 2.99e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1579.0  on 9997  degrees of freedom
## AIC: 1585
##
```

Making Predictions

$$\hat{p}(X) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 X)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 X)}$$

and we plug in values for X.

- We can also plug in values for X into the odds or log odds formulations to predict those values.

Multiple Logistic Regression

This extension is very similar to the one from simple to multiple linear regression.

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

and

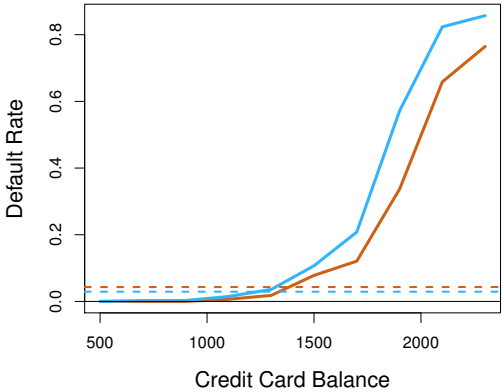
$$\hat{p}(X) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}$$

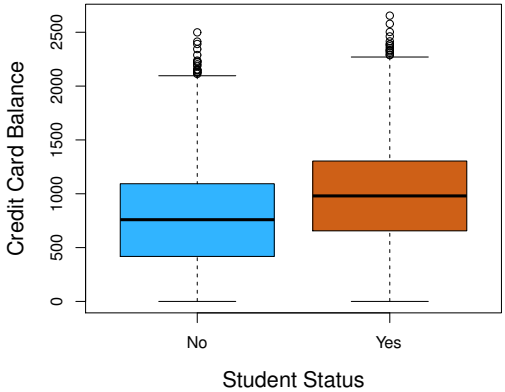
Example

```
##
## Call:
## glm(formula = default ~ ., family = binomial, data = Default)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
## studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
## balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
## income       3.033e-06  8.203e-06   0.370  0.71152
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1571.5  on 9996  degrees of freedom
```

```
Coefficients: Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
income       3.033e-06  8.203e-06   0.370  0.71152
```

- Negative coefficient suggests students less likely to default than non students.
- However, a model with *only* `student` has a positive coefficient. Why?





- Overall average default rates are higher for students.
- As a function of credit card balance, default rates are *lower* for students.
  - So students are riskier than non-students *on average*, but a student is less risky than a non-student with the same credit card balance.

Multinomial Logistic Regression

What if the response variable has more than two outcomes, $K > 2$?

Select one class to serve as the *baseline*. WLOG, we choose class K. Then the model becomes

$$\mathrm{P}(Y = k|X = x) = \frac{\exp(\beta_{k0} + \beta_{k1}x_1 + \beta_{k2}x_2 + \cdots + \beta_{kp}x_p)}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l0} + \beta_{l1}x_1 + \beta_{l2}x_2 + \cdots + \beta_{lp}x_p)}$$

for p predictor variables and $k = 1, \ldots, K-1$ and

$$\mathrm{P}(Y = K|X = x) = \frac{1}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l0} + \beta_{l1}x_1 + \beta_{l2}x_2 + \cdots + \beta_{lp}x_p)}.$$

Multinomial Logistic Regression

Then for $k = 1, \ldots, K-1$,

$$\log\left(\frac{\mathrm{P}(Y = k|X = x)}{\mathrm{P}(Y = K|X = x)}\right) = \beta_{k0} + \beta_{k1}x_1 + \beta_{k2}x_2 + \cdots + \beta_{kp}x_p$$

- Baseline is unimportant from a modeling perspective, but it does serve as the point of comparison for other classes.
- Say `seizure` is our baseline for some medical classification problem and let k represent `stroke`.
  - Then the model gives the log odds of `stroke` versus `seizure`.

Softmax Coding

Another way to do the estimation for a multinomal logistic regression is, for $k = 1, 2, \ldots, K$

$$\mathrm{P}(Y = k|X = x) = \frac{\exp(\beta_{k0} + \beta_{k1}x_1 + \beta_{k2}x_2 + \cdots + \beta_{kp}x_p)}{\sum_{l=1}^{K} \exp(\beta_{l0} + \beta_{l1}x_1 + \beta_{l2}x_2 + \cdots + \beta_{lp}x_p)}$$

which requires the estimation of coefficients for all K classes, instead of $K - 1$ classes.

Then the log odds ratio between *any two classes* k and k' is

$$\log\left(\frac{\mathrm{P}(Y = k|X = x)}{\mathrm{P}(Y = k'|X = x)}\right) = (\beta_{k0} - \beta_{k'0}) + (\beta_{k1} - \beta_{k'1})x_1 + (\beta_{k2} - \beta_{k'2})x_2 + \cdots + (\beta_{kp} - \beta_{k'p})x_p$$