Introducing Logistic Regression

November 22, 2019

Types of Outcome Variable

- So far, all of our outcome variables have been numeric.
- Values of $\hat{y}$ are continuous numeric.
- What happens when we have categorical outcomes?
- Enter **logistic regression**.

Generalized Linear Models

(Multiple) linear regression and logistic regression are both a type of **generalized linear model (GLM)**.

- Logistic regression will allow us to model binary response variables.
- That is, we will be able to model categorical variables with two levels.

Generalized Linear Models

We can think of GLMs as a two-stage approach:

1. Model the response variable using some probability distribution.
2. Model the distribution's parameter(s) using a collection of predictors (as in multiple regression).

We've already been doing this!

For a continuous outcome,

1. The response variable is assumed to follow a normal distribution.
2. The mean of this normal distribution is $\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$.

Example: Resume Data

Consider data from a study to examine the effect of race and sex on job application callback rates.

- Fake resumes were sent to job ads in Boston and Chicago.
- Researchers wanted to see which would elicit callbacks.
- Experience and education were randomly generated.
- Finally, names were randomly generated and added to the resumes.
  - Names were generated such that hiring managers would be likely to assume both race and gender.

Example: Resume Data

The response variable of interest is

$$\texttt{callback} = \begin{cases} 1 & \text{if} \quad \text{received callback} \\ 0 & \text{otherwise} \end{cases}$$

Example: Resume Data

The variables in this dataset are

`callback`	yes or no
`job_city`	Boston or Chicago
`college_degree`	yes or no
`years_experience`	Numeric, number of years experience
`honors`	Resume lists some type of honors, yes or no
`military`	yes or no
`email_address`	Listed, yes or no
`race`	Black or white (implied by name)
`sex`	implied by name

Example

Race and sex are protected classes in the US, meaning that employers are not legally allowed to make hiring decisions based on these factors.

This study...

- has random assignment.
- is a true experiment.

Therefore we may infer causation between (statistically significant) variables and the callback rate.

Modeling the Probability of an Event

With logistic regression,

- The outcome Y_i takes values 1 or 0 with some probability.
  - $P(Y_i = 1) = p_i$
  - $P(Y_i = 0) = 1 - p_i$
- The subscript i refers to the ith observation (in this case the ith resume).
- We will model the probability p, which takes values $p_1, \dots, p_n$.

Logistic Regression

We want to relate the probability of a callback for each resume, p, to the predictors $x_1, \ldots, x_k$.

This will look a lot like multiple regression!

$$\text{transformation}(p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$

Transforming p

Why do we transform p?

- We want the range of possibilities for the outcome to match the range of p
  - $p = P(Y = 1)$ is between 0 and 1!
- Without a transformation, $\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ could take values outside of 0 to 1.

Transforming p

A common transformation for p is the **logit transformation**:

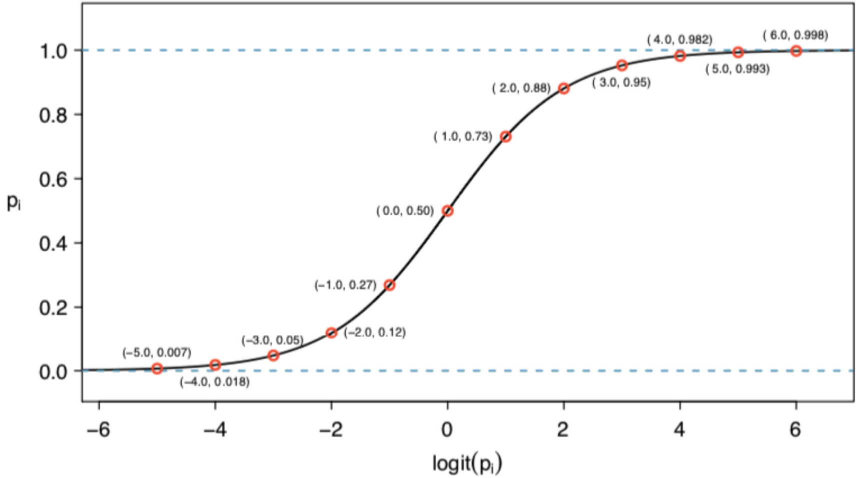
$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

The Logistic Model

Then the model looks like

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$

Example: Transforming the Resume Data



Example: Fitting the Model

We start with the model that includes only `honors`.

$$\log\left(\frac{p}{1-p}\right) = -2.4998 + 0.8668 \times \texttt{honors}$$

For a resume with no honors listed, what is the probability of a callback?

A Note

As with multiple regression, we'll fit all of these models using a computer (the computer will do the logit transformation for you, too!), but we do need to know how to interpret the results.

Converting Back to probabilities

To make probability predictions using a logistic regression, use

$$p = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}{1 \ + \ e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}$$

Example: Resume Data

The summary for the full model is

	Estimate	Std. Error	z value	$Pr(>\|z\|)$
(Intercept)	-2.6632	0.1820	-14.64	<0.0001
job_city:Chicago	-0.4403	0.1142	-3.85	0.0001
college_degree	-0.0666	0.1211	-0.55	0.5821
years_experience	0.0200	0.0102	1.96	0.0503
honors	0.7694	0.1858	4.14	<0.0001
military	-0.3422	0.2157	-1.59	0.1127
email_address	0.2183	0.1133	1.93	0.0541
race:white	0.4424	0.1080	4.10	<0.0001
sex:male	-0.1818	0.1376	-1.32	0.1863

Variable Selection for Logistic Regression

- The approach is similar to using R^2_{adj} in multiple regression.
- Use a statistic called **Akaike information criterion (AIC)**.
  - This is similar to R^2_{adj} in that it balances model fit and number of parameters.
- We will prefer models with a *lower* AIC value.

Running all possible seven-variable models for the resume data, the model with the lowest AIC has `college_degree` removed.

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-2.7162	0.1551	-17.61	<0.0001
job_city:Chicago	-0.4364	0.1141	-3.83	0.0001
years_experience	0.0206	0.0102	2.02	0.0430
honors	0.7634	0.1852	4.12	<0.0001
military	-0.3443	0.2157	-1.60	0.1105
email_address	0.2221	0.1130	1.97	0.0494
race:white	0.4429	0.1080	4.10	<0.0001
sex:male	-0.1959	0.1352	-1.45	0.1473

Notice that the coefficients barely changed!

The Logistic Regression Model

- Sex is not statistically significant.
- However, race is associated with a near-zero p-value.
  - The coefficient corresponds to `white`.
  - To interpret this coefficient, we would say that the *probability of callback* is higher for `white`.
  - These data provide very strong evidence for racial bias in job application callbacks.

Example

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-2.7162	0.1551	-17.61	<0.0001
job_city:Chicago	-0.4364	0.1141	-3.83	0.0001
years_experience	0.0206	0.0102	2.02	0.0430
honors	0.7634	0.1852	4.12	<0.0001
military	-0.3443	0.2157	-1.60	0.1105
email_address	0.2221	0.1130	1.97	0.0494
race:white	0.4429	0.1080	4.10	<0.0001
sex:male	-0.1959	0.1352	-1.45	0.1473

Write the logistic regression model for these data.