Logistic Regression

November 25, 2019

Example

	Estimate	Std. Error	z value	Pr(> $\lvert z \rvert$)
(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.

Example

Use the logistic regression model to estimate the probability of receiving a callback for a job in Chicago where the candidate lists 14 years experience, no honors, no military experience, includes an email address, and has a first name that implies they are a White male.

...then calculate the probability of receiving a callback for a candidate who is the same on all characteristics but race.

On average, how many jobs does each candidate need to apply to in order to receive a callback?

Resume Data

- This is a simplified version of the actual data used in the 2003 article.
  - The full data produced the same basic conclusions.
- Because regression is about trends or averages, it is impossible (using this data) to point fingers at any particular employer or hiring managers.
- All we can say for sure is that the data shows a clear racial bias in job callbacks.

Logistic Regression Diagnostics

There are two key conditions for the logistic regression model:

1. Each outcome y_i is independent of the other outcomes.
2. Each predictor x_i is linearly related to logit(p_i) if all other predictors are held constant.

Note: the linear regression assumptions of constant variance and normally distributed residuals come from the normal model.

In a logistic regression, we assume a binomial model.

Diagnostics for the Callback Rate Model: Independence

Independence is satisfied because the resume characteristics were randomly assigned.

Diagnostics for the Callback Rate Model: Linearity

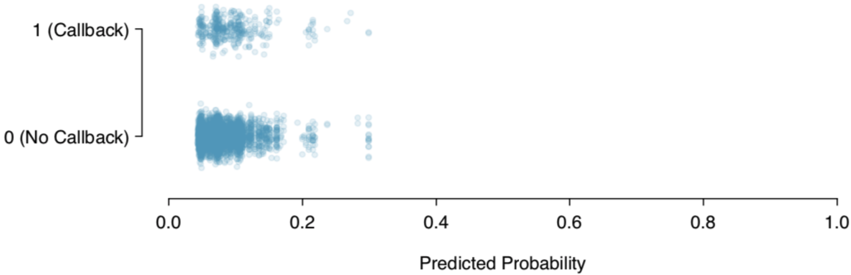
- It is difficult to check linearity without a fairly large dataset.
- Fortunately, the resume data has 4870 entries.

Diagnostics for the Callback Rate Model

We want to assess the quality of the model.

Ex: if we look at resumes that we modeled as having a 10% chance of getting a callback, do we find about 10% of them actually receive a callback?

Diagnostics for the Callback Rate Model



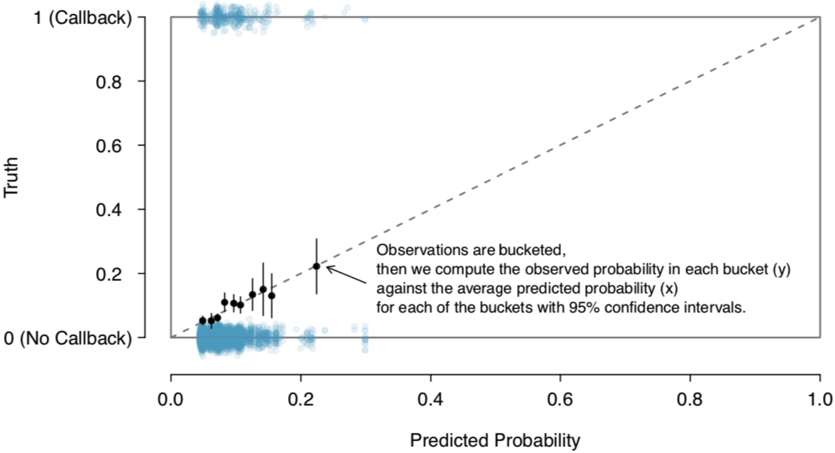
We will start to think about this using a plot.

Diagnostics for the Callback Rate Model

We will add to this plot as follows:

1. Bucket the data into groups based on predicted probabilities.
2. Compute the average predicted probability for each group.
3. Compute the observed probability for each group along with 95% confidence intervals.
4. Plot the observed probabilities (with the confidence intervals) against the average predicted probabilities for each group.

Diagnostics for the Callback Rate Model



Observations are bucketed,
then we compute the observed probability in each bucket (y)
against the average predicted probability (x)
for each of the buckets with 95% confidence intervals.

Groups of Different Sizes

Section 9.5 closes with an interesting probability problem.

- In the resume data, we found that individuals with names perceived to be from a Black person would have to send $\approx 50\%$ more resumes in order to receive a callback.
- We want to consider other dimensions to this.

Example

Consider a hypothetical company made up of 20% women and 80% men.

- Suppose the company has $20,000$ employees.
- Whenever someone goes up for promotion, 5 of their colleagues are randomly chosen to provide feedback on their work.
- Now suppose that in this company, 10% of the people are prejudiced against the other sex.

How often will people experience sex-based discrimination?

Groups of Different Sizes

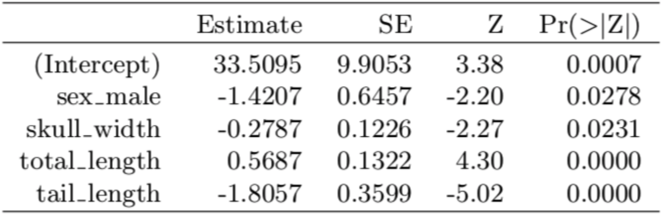
This is a very simplified example, but we've used probability to highlight something:

- Even when both groups are equally discriminatory, the smaller group will experience more discrimination.
- Increasing the imbalance in population size increases this discrepancy.

There is a lot of nuance being left out here, but nonetheless this is an important probability property to be aware of.

Example: 9.17 Possum classification, Part II.

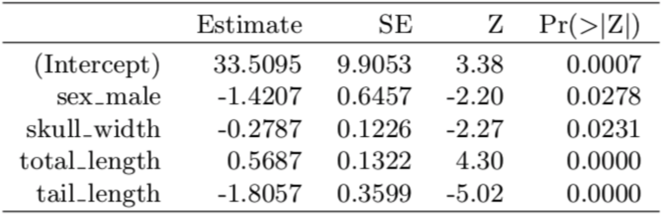
A logistic regression model was proposed for classifying common brushtail possums into their two regions. The outcome variable took value 1 if the possum was from Victoria and 0 otherwise.



	Estimate	SE	Z	Pr(>|Z|)
(Intercept)	33.5095	9.9053	3.38	0.0007
sex_male	-1.4207	0.6457	-2.20	0.0278
skull_width	-0.2787	0.1226	-2.27	0.0231
total_length	0.5687	0.1322	4.30	0.0000
tail_length	-1.8057	0.3599	-5.02	0.0000

- Write out the form of the model. Also identify which of the variables are positively associated when controlling for other variables.

Example: 9.17 Possum classification, Part II.



	Estimate	SE	Z	Pr(>|Z|)
(Intercept)	33.5095	9.9053	3.38	0.0007
sex_male	-1.4207	0.6457	-2.20	0.0278
skull_width	-0.2787	0.1226	-2.27	0.0231
total_length	0.5687	0.1322	4.30	0.0000
tail_length	-1.8057	0.3599	-5.02	0.0000

- Suppose we see a brushtail possum at a zoo in the US, and a sign says the possum had been captured in the wild in Australia, but it doesn't say which part of Australia. However, the sign does indicate that the possum is male, its skull is about 63 mm wide, its tail is 37 cm long, and its total length is 83 cm. What is the model's computed probability that this possum is from Victoria? How confident are you in the model's accuracy of this probability calculation?