

Lecture 28: Logistic Regression

Chapter 8.4

1 / 19

Binary Outcome Variables

Instead of numerical outcomes, we have observations Y_i for $i = 1, \dots, n$ where

- ▶ $Y_i = 1$ with probability p_i
- ▶ $Y_i = 0$ with probability $1 - p_i$

Logistic regression: we are modeling p_i 's with a linear model.

2 / 19

Outcome Variable

Let

$$x_{1,i}, \dots, x_{k,i}$$

be the k predictor variables associated with the i^{th} observation

One's first thought might be to model the p_i 's using linear regression:

$$p_i = \beta_0 + \beta_1 x_{1,i} + \dots + \beta_k x_{k,i}$$

However, you may end up fitting p_i 's that are either

- ▶ less than 0
- ▶ greater than 1

3 / 19

Outcome Variable

Rather, what is modeled is the **logit transformation** or **log-odds** of p_i

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \dots + \beta_k x_{k,i}$$

Why this transformation? It maps the $[0, 1]$ interval to a $(-\infty, \infty)$ interval.

4 / 19

Outcome Variable

First, convert p_i into odds:

“Two to one odds for event X” \equiv “There is a 66% chance of event X occurring.”

Then we take the natural log of it. So

- ▶ for $p_i = 0 \Rightarrow \log\left(\frac{p_i}{1-p_i}\right) = \log\left(\frac{0}{1}\right) = -\infty$
- ▶ for $p_i = 0.5 \Rightarrow \log\left(\frac{p_i}{1-p_i}\right) = \log\left(\frac{0.5}{0.5}\right) = 0$
- ▶ for $p_i = 1 \Rightarrow \log\left(\frac{p_i}{1-p_i}\right) = \log\left(\frac{1}{0}\right) = \log(\infty) = \infty$

Outcome Variable

Figure 8.14 from page 369

Simple Logistic Regression Example p.370

So say we fit a logistic regression with ($n = 3921$):

- ▶ Y_i is spam: binary variable of whether message was classified as spam (1 if spam)
- ▶ x_i is to_multiple: binary variable indicating if more than one recipient listed

| | Estimate | Std. Error | z value | Pr(> z) |
|-------------|----------|------------|---------|----------|
| (Intercept) | -2.1161 | 0.0562 | -37.67 | 0.0000 |
| to_multiple | -1.8092 | 0.2969 | -6.09 | 0.0000 |

The regression equation is

$$\log\left(\frac{p_i}{1-p_i}\right) = -2.12 - 1.81 \times \text{to_multiple}$$

7 / 19

Inverse Logit Transformation

How do we convert back into p_i 's?

$$\text{Say } x = \log\left(\frac{p_i}{1-p_i}\right) \text{ then } p_i = \frac{\exp(x)}{1 + \exp(x)}$$

is the **inverse logit transformation**.

So to convert the regression equation to probabilities, we compute

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_{1,i} + \dots + \beta_k x_{k,i})}{1 + \exp(\beta_0 + \beta_1 x_{1,i} + \dots + \beta_k x_{k,i})}$$

8 / 19

Fitted Probabilities

To compute the fitted probabilities \hat{p}_i :

- ▶ `to_multiple= 0` (only one recipient):

$$\hat{p}_i = \frac{\exp(-2.12 - 1.81 \times 0)}{1 + \exp(-2.12 - 1.81 \times 0)} = 0.11$$

- ▶ `to_multiple= 1` (many recipients):

$$\hat{p}_i = \frac{\exp(-2.12 - 1.81 \times 1)}{1 + \exp(-2.12 - 1.81 \times 1)} = 0.02$$

Note: 11% and 2% are not dramatically different. In an ideal world of binary predictors, we'd have fitted probabilities of 100% and 0%.

9 / 19

Fitted Model Using Backwards Regression

The following model was selected in the text using backwards selection using $\alpha = 0.05$.

| | Estimate | Std. Error | z value | Pr(> z) |
|----------------------------------|----------|------------|---------|----------|
| (Intercept) | -0.8057 | 0.0880 | -9.15 | 0.0000 |
| <code>to_multiple?</code> | -2.7514 | 0.3074 | -8.95 | 0.0000 |
| <code>word winner used?</code> | 1.7251 | 0.3245 | 5.32 | 0.0000 |
| <code>special formatting?</code> | -1.5857 | 0.1201 | -13.20 | 0.0000 |
| <code>'RE:' in subject?</code> | -3.0977 | 0.3651 | -8.48 | 0.0000 |
| <code>attachment?</code> | 0.2127 | 0.0572 | 3.72 | 0.0002 |
| <code>word password used?</code> | -0.7478 | 0.2956 | -2.53 | 0.0114 |

10 / 19

Fitted Model Using Backwards Regression

The following variables increase the probability that the email is spam, since $b > 0$

| | Estimate | Std. Error | z value | Pr(> z) |
|-------------------|----------|------------|---------|----------|
| (Intercept) | -0.8057 | 0.0880 | -9.15 | 0.0000 |
| word winner used? | 1.7251 | 0.3245 | 5.32 | 0.0000 |
| attachment? | 0.2127 | 0.0572 | 3.72 | 0.0002 |

11 / 19

Fitted Model Using Backwards Regression

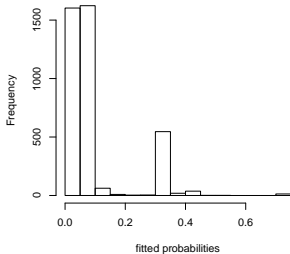
The following variables decrease the probability that the email is spam, since $b < 0$

| | Estimate | Std. Error | z value | Pr(> z) |
|---------------------|----------|------------|---------|----------|
| (Intercept) | -0.8057 | 0.0880 | -9.15 | 0.0000 |
| to_multiple? | -2.7514 | 0.3074 | -8.95 | 0.0000 |
| special formatting? | -1.5857 | 0.1201 | -13.20 | 0.0000 |
| 'RE:' in subject? | -3.0977 | 0.3651 | -8.48 | 0.0000 |
| word password used? | -0.7478 | 0.2956 | -2.53 | 0.0114 |

12 / 19

Fitted Probabilities

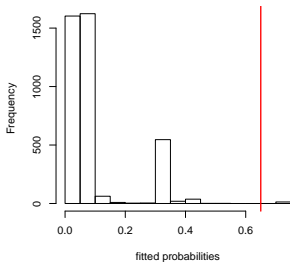
These are all 3921 fitted probabilities:



13 / 19

Using Cutoffs to Classify Emails as Spam

Say we use a cutoff of 65% to **classify** an email spam or not:



14 / 19

Using Cutoffs to Classify Emails as Spam

Using a cutoff of 65%:

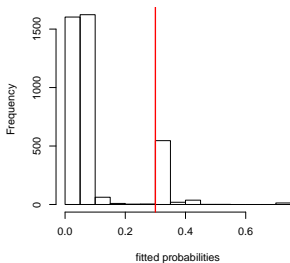
| | | Classification | |
|-------|----------|----------------|------|
| | | Not Spam | Spam |
| Truth | Not Spam | 3351 | 3 |
| | Spam | 357 | 10 |

- ▶ Of the emails classified as spam: $\frac{10}{10+3} = 76\%$ correct
- ▶ Of the emails classified not as spam: $\frac{3351}{3351+357} = 90.3\%$ correct

15 / 19

Using Cutoffs to Classify Emails as Spam

Now say we use a cutoff of 30% to **classify** an email spam or not:



16 / 19

Using Cutoffs to Classify Emails as Spam

Using a cutoff of 30%:

| | | Classification | |
|-------|----------|----------------|------|
| | | Not Spam | Spam |
| Truth | Not Spam | 3138 | 416 |
| | Spam | 166 | 201 |

- ▶ Of the emails classified as spam: $\frac{201}{201+416} = 32.6\%$ correct
- ▶ Of the emails classified not as spam: $\frac{3138}{3138+166} = 95.0\%$ correct

17 / 19

Using Cutoffs to Classify Emails as Spam

Moral of the Story: most classifiers are never perfect (like hypothesis tests). There will almost always be a trade-off between:

- ▶ Type I errors: labeling an email spam when it is not
- ▶ Type II errors: failing to label an email as spam when it is

18 / 19

Assumptions for Logistic Regression

- ▶ There is a roughly linear relationship between each of the predictors and $\log\left(\frac{p}{1-p}\right)$.
- ▶ Each outcome Y_i is independent of the other outcomes. This can be verified using the residuals $e_i = Y_i - \hat{p}_i$

Please read pages 375 and 376 from the text.