Lecture 28: Logistic Regression

Chapter 8.4

Binary Outcome Variables

Outcome Variable

Logit Transformation

Odds

Outcome Variable

Figure 8.13 from page 388

Simple Logistic Regression Example p.388

So say we fit a logistic regression with:

- Y_i is `spam`: binary variable of whether message was classified as spam (1 if spam)

Simple Logistic Regression Example p.388

So say we fit a logistic regression with:

- 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

Simple Logistic Regression Example p.388

So say we fit a logistic regression with:

- 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

Inverse Logit Transformation

Fitted Probabilities

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
to_multiple?	-2.7514	0.3074	-8.95	0.0000
word winner used?	1.7251	0.3245	5.32	0.0000
special formatting?	-1.5857	0.1201	-13.20	0.0000
'RE:' in subject?	-3.0977	0.3651	-8.48	0.0000
attachment?	0.2127	0.0572	3.72	0.0002
word password used?	-0.7478	0.2956	-2.53	0.0114

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

Fitted Model Using Backwards Regression

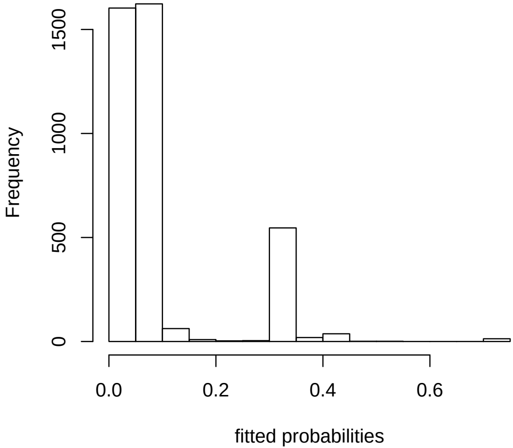
The following variables decrease the probability that the email is spam, since $b < 0$

	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

Assumptions for Logistic Regression

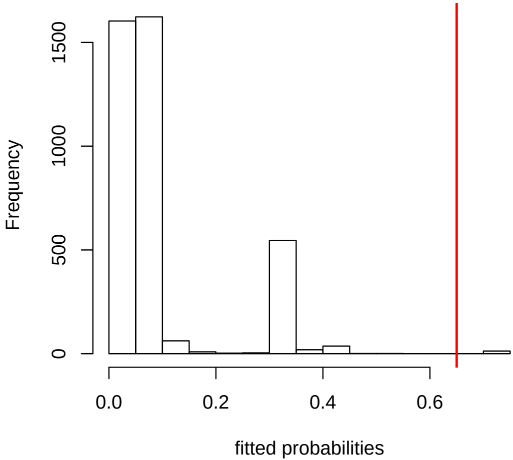
Fitted Probabilities

These are all 3921 fitted probabilities:



Using Cutoffs to Classify Emails as Spam

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



Using Cutoffs to Classify Emails as Spam

Using a cutoff of 65%:

		Classification: Not Spam	Classification: Spam
Truth	Not Spam	3351	3
	Spam	357	10

Using Cutoffs to Classify Emails as Spam

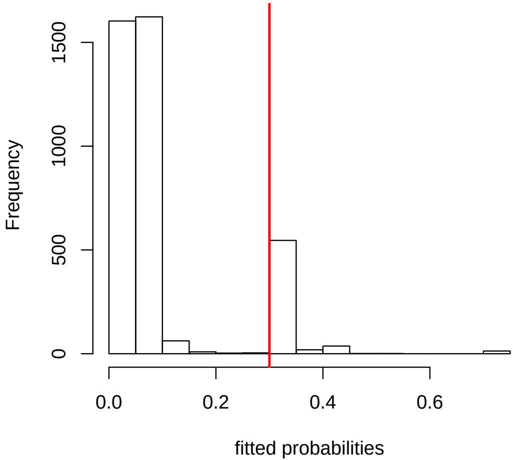
Using a cutoff of 65%:

		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

Using Cutoffs to Classify Emails as Spam

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



Using Cutoffs to Classify Emails as Spam

Using a cutoff of 30%:

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

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

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