Logistic Regression

Objectives

- Describe the motivation for logistic regression
- Understand how to fit a logistic model and interpret its coefficients
- Explain common classification metrics, and how they tie into the ROC curve

Agenda

Afternoon: **Logistic Regression**

- Logistic regression details
  - Sigmoid (logistic) function
- Interpreting the results
  - Fitted values (probabilities)
  - Coefficients (log odds ratios)
- Pair Programming: Parts 2, 3, 4, and 5

Logistic Regression - Motivation II

Breakout: Pair Exercise, 5 mins

- Why logistic and not just plain old linear?
  1. What shape does the logistic function take?
  2. Why is the logistic function a good, logical fit for binary classification?
  3. Discuss the problems with using standard linear regression for modeling binary response

Linear Review - Underlying Assumptions

- With linear regression, we assume that our **response is normally distributed**:

$$y_i|X \sim N(X\beta, \sigma^2)$$

- **With classification, that isn't the case**

Classification - Observed Distribution

- With binary classification setting, **response is binary**:

$$Y_i = \begin{cases} 1, \text{if an event occurs} \\ 0, \text{if it doesn't} \end{cases}$$

- We are interested in the probability that an event occurs given a subject's profile: $p_i = P(y_i = 1|X)$
  - Each observation is drawn from a **Bernoulli distribution**: $y_i|X \sim Bernoulli(p)$
- Our standard linear model won't work

A Model for Classification

- We need a model that:
  - Takes continuous input (e.g., from $-\infty$ to ∞)
  - Produces output between 0 and 1
  - Transitions between 0 and 1 "without wasting much time"
  - Has interpretable coefficients (like our standard linear regression model)
    - Takes the mean response of our observations and links it to a linear combination of our inputs (e.g., $X\beta$)
- Just as with linear regression, the linear predictor is $X\beta = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$ where X is your design matrix: x_0 is a column vector of 1's and $x_1, ..., x_p$ are the feature column vectors
  - The explanatory variables may be quantitative, categorical, or mixed

Logistic Regression for Classification

- Enter **logistic regression**...

$$p(y) = \frac{1}{1 + e^{-X\beta}}$$

Note: $p(y)$ denotes the probability of success for y. We can think of this as the mean of the response

An Aside: Odds and Probabilities

- Given the probability an event occurring, the odds of that event are

$$odds = \frac{p}{1-p}$$

- Similarly, given the odds, you can calculate the probability

$$p = \frac{odds}{1+odds}$$

An Aside: Odds and Probabilities

- Odds are commonly used in gambling, especially horse-racing
  - Even odds (1:1): $p = \frac{1}{1+1} = 0.5$
  - Odds are 3:1 for an event: $p = \frac{3}{1+3} = 0.75$
  - Long shot: 20:1 against: $1 - p = 1 - \frac{20}{1+20} = 0.0476$

Logistic Regression - The Details

Logistic Regression

- Logistic regression fits a **logistic function** that we use to obtain the probability that the response of an individual observation (y) is a success (typically denoted as a 1, where a failure is denoted by a 0)

$$p(y) = \frac{1}{1 + e^{-X\beta}}$$

- How do we get this function, though?

Logistic Regression - Link Function

- The **link function** provides the relationship between a linear combination of our inputs ($X\beta$) and the mean of our response (p)
- For logistic regression, we use the following link function

$$log(\frac{p}{1-p}) = X\beta$$

- The logistic model can be rewritten in terms of odds via the logit function

$$log(\frac{p}{1-p}) = logodds = logit(p)$$

  - Giving us a nice framework that seems familiar

$$logit(p) = X\beta$$

- See the appendix for a derivation of how to move from this to the logistic function that we use to predict the mean of our response

Estimating the Parameters

- The parameters of our logistic regression are estimated via maximum likelihood. We know that each individual observation follows a Bernoulli distribution:

$$y_i|X \sim Bernoulli(p)$$

- Given this, we can construct the likelihood of our β matrix as:

$$\mathcal{L}(\beta) = \prod_{i=1}^{N} p(y_i)^{y_i}(1 - p(y_i))^{1-y_i}$$

And from there, our log likelihood:

$$\ell(\beta) = \sum_{i=1}^{N} y_i log\, p(y_i) + (1 - y_i) log(1 - p(y_i))$$

Estimating the Parameters

- The regression coefficients can be estimated using maximum likelihood estimation (MLE)
- Unlike linear regression, no closed form solution exists, therefore an iterative method such as Newton-Rhapson or Gradient Descent is needed
- Reasons that the model may not reach convergence
  - A large number of features relative to subjects
    - rule of thumb is at least 10 cases for each explanatory variable
  - Multicollinearity
  - Sparseness, specifically low cell counts for categorical predictors

An Aside: Odds Ratio

- Given the definition of odds above, the odds ratio is

$$OR = \frac{Odds_1}{Odds_2} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

- For example, say the probability of a disease in individuals with a certain genetic trait is $p_1 = 0.05$ while in the general population its $p_2 = 0.001$ the resulting odds ratio would be

$$OR = \frac{0.05/0.95}{0.001/0.999} \approx 53$$

- This represents a measure of relative risk such that an individual with the genetic trait is 53 time more likely to develop the disease than a randomly chosen person

Model Interpretation

- In linear regression, the $\hat{\beta}$ coefficientscients can be interpreted directly as the change in y for a 1-unit increase in the explanatory variable
- In logistic regression, however, this would represent the change in logit value for a 1-unit increase in the explanatory variable, which is not interpretable
- We can however convert the $\hat{\beta}$ coefficient to an estimate of Odds Ratio for a 1-unit increase in the explanatory variable

$$\widehat{OR} = e^{\hat{\beta}}$$

Model Interpretation - Example 1

- Say we fit a logistic regression model with the outcome/response as whether or not a person works (yes/no, which is denoted with a 1/0) and only one predictor, income:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_{income} X_{income})}}$$

- Let's say that β_{income} is 0.00001. This means that a one-unit increase in income (\$1) causes an $e^{0.00001}$ increase in the odds of somebody working
  - $e^{0.00001} = 1.00001$

Model Interpretation - Example 2

- This ultimately means that for each additional dollar that a person makes, we expect a 0.001% increase in the odds that they are working
  - For an additional $1000 dollars that a person makes, we expect a 1% increase in the odds that they are working

Uses of Logistic Regression

- To **predict probabilities** that subjects fall into one of 2 categories on a dichotomous response variable
- To **classify** subjects into one of 2 categories
  - one of our main focuses
- Lots of other possibilities

Predict Probabilities

Once the $\hat{\beta}$ coefficients have been calculated, we can estimate the probability p of the event occurring with

$$p = \frac{1}{1 + e^{-X\hat{\beta}}}$$

Classification

General Method

- For each unclassified subject, you would calculate the probability the subject falls in a specific class using the fitted model
- You would compare that probability to a predetermined decision rule boundary
  - default is 0.5
- If the predicted probability is > 0.5, classify as 1
  - Otherwise, classify as 0

Breakout: Pair, 5 mins

Understanding your chances

- State what each of the following terms are:
  - Probability
  - Odds
  - Log-Odds
  - Odds Ratio
- Give an example to demonstrate what each of the 4 terms are

Breakout: Individual, 5 mins; then pair, 5 mins

Interpret the results from this logistic regression model

- What are my current chances of getting into grad school?
- How would my chances change if I increased my GPA by 100 points?
- What score would I need on the GRE's to increase my chances to 95%?

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                  admit   No. Observations:                  400
Model:                          Logit   Df Residuals:                      397
Method:                           MLE   Df Model:                            2
Date:                Fri, 02 Dec 2016   Pseudo R-squ.:                 0.03927
Time:                        16:43:29   Log-Likelihood:                -240.17
converged:                       True   LL-Null:                       -249.99
                                        LLR p-value:                 5.456e-05
==============================================================================
                 coef    std err          z      P>|z|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const         -4.9494      1.075     -4.604      0.000        -7.057    -2.842
gre            0.0027      0.001      2.544      0.011         0.001     0.005
gpa            0.7547      0.320      2.361      0.018         0.128     1.381
==============================================================================
```

Breakout: Individual, 5 mins; then pair, 5 mins

Models 1 and 2 are from the same dataset. Explain what you see

Dep. Variable:	Survived	**No. Observations:**	712
Model:	Logit	**Df Residuals:**	709
Method:	MLE	**Df Model:**	2
Date:	Tue, 22 Nov 2016	**Pseudo R-squ.:**	0.2528
Time:	15:27:35	**Log-Likelihood:**	-359.02
converged:	True	**LL-Null:**	-480.45
		LLR p-value:	1.825e-53

	coef	std err	z	P>|z|	[95.0% Conf. Int.]
Intercept	0.6590	0.167	3.935	0.000	0.331 0.987
Sex[T.male]	-2.3711	0.189	-12.524	0.000	-2.742 -2.000
Fare	0.0121	0.003	4.595	0.000	0.007 0.017

Dep. Variable:	Survived	**No. Observations:**	712
Model:	Logit	**Df Residuals:**	708
Method:	MLE	**Df Model:**	3
Date:	Tue, 06 Dec 2016	**Pseudo R-squ.:**	0.3013
Time:	08:33:07	**Log-Likelihood:**	-335.70
converged:	True	**LL-Null:**	-480.45
		LLR p-value:	1.852e-62

	coef	std err	z	P>|z|	[95.0% Conf. Int.]
Intercept	3.1335	0.399	7.863	0.000	2.352 3.915
Sex[T.male]	-2.5536	0.204	-12.528	0.000	-2.953 -2.154
Fare	0.0019	0.002	0.850	0.395	-0.002 0.006
Pclass	-0.9283	0.137	-6.788	0.000	-1.196 -0.660

Appendix

Logistic Regression - From link to probability

1. $log(\frac{p}{1-p}) = X\beta$
2. $\frac{p(y)}{1-p} = e^{X\beta}$
3. $p = (1-p)e^{X\beta}$
4. $p = e^{X\beta} - pe^{X\beta}$
5. $p + p(y)e^{X\beta} = e^{X\beta}$
6. $p(1 + e^{X\beta}) = e^{X\beta}$
7. $p = \frac{e^{X\beta}}{1+e^{X\beta}}$
8. $p = \frac{\frac{e^{X\beta}}{X\beta}}{\frac{1+e^{X\beta}}{X\beta}}$
9. $p = \frac{1}{1+e^{-X\beta}}$