Logistic Regression

Aims

- When and Why do we Use Logistic Regression?
  - Binary
  - Multinomial
- Theory Behind Logistic Regression
  - Assessing the Model
  - Assessing predictors
- Interpreting Logistic Regression

When to use Logistic Regression

Select A Statistical Test

- Hypothesis tests to find relationships between project Y and potential X's

		Y: Continuous	Y: Discrete
X	Continuous	Simple Linear Regression	Logistic Regression
X	Discrete	2 Sample t-Test *(Compare Means of two samples)* ANOVA *(Compare means of multiple samples)* Homgeneity of Variance *(Compare variances)*	Chi-Square Test

When And Why

- To predict an outcome variable that is categorical from one or more categorical or continuous predictor variables.
- Used because having a categorical outcome variable violates the assumption of linearity in normal regression.

With One Predictor

$$P(Y) = \frac{1}{1+e^{-(b_0+b_1X_1+\varepsilon_i)}}$$

- Outcome
  - We predict the *probability* of the outcome occurring
- b_0 *and* b_1
  - Can be thought of in much the same way as multiple regression
  - Note the normal regression equation forms part of the logistic regression equation

With Several Predictor

$$P(Y) = \frac{1}{1 + e^{-(b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_n X_n + \varepsilon_i)}}$$

- Outcome
  - We still predict the *probability* of the outcome occurring
- Differences
  - Note the multiple regression equation forms part of the logistic regression equation
  - This part of the equation expands to accommodate additional predictors

Assumptions

- Logistic regression does not make any assumptions of normality, linearity and homogeneity of variance for the independent variables.
- Because it does not impose these requirements, it is preferred to Discriminant analysis when the data does not satisfy these assumptions.
- The only “real” limitation on logistic regression is that the outcome must be discrete.

Sample size requirements

- The minimum number of cases per independent variable is 10, using a guideline provided by Hosmer and Lemeshow, authors of *Applied Logistic Regression*, one of the main resources for Logistic Regression.

- For preferred case-to-variable ratios, we will use 20 to 1 for simultaneous and hierarchical logistic regression and 50 to 1 for stepwise logistic regression.

The logistic function

- Advantages of the logit
  - Simple transformation of P(y|x)
  - Linear relationship with x
  - Can be continuous (Logit between $-\infty$ to $+\infty$)
  - Known binomial distribution (P between 0 and 1)

$$\ln\left(\frac{P}{1-P}\right) = \alpha + \beta x \qquad \frac{P}{1-P} = e^{\alpha+\beta x}$$

Interpretation of b

	Exposure (x)	
Disease (y)	**Yes**	**No**
Yes	$P(y\|x=1)$	$P(y\|x=0)$
No	$1-P(y\|x=1)$	$1-P(y\|x=0)$

$$\frac{P}{1-P}=e^{\alpha+\beta x}$$

$$Odds_{d|e} = e^{\alpha+\beta}$$

$$Odds_{d|\bar{e}} = e^{\alpha}$$

$$OR = \frac{e^{\alpha+\beta}}{e^{\alpha}} = e^{\beta}$$

$$\ln(OR) = \beta$$

Methods for including variables

- There are three methods available for including variables in the regression equation:
  - The simultaneous method in which all independents are included at the same time
  - The hierarchical method in which control variables are entered in the analysis before the predictors whose effects we are primarily concerned with.
  - The stepwise method in which variables are selected in the order in which they maximize the statistically significant contribution to the model.

- For all methods, the contribution to the model is measures by model chi-square is a statistical measure of the fit between the dependent and independent variables, like R^2.

Computational method

- Multiple regression uses the least-squares method to find the coefficients for the independent variables in the regression equation, i.e. it computed coefficients that minimized the residuals for all cases.

- Logistic regression uses maximum-likelihood estimation to compute the coefficients for the logistic regression equation. This method finds attempts to find coefficients that match the breakdown of cases on the dependent variable.

Computational method…

- The overall measure of how will the model fits is given by the likelihood value, which is similar to the residual or error sum of squares value for multiple regression.

- Maximum-likelihood estimation is an iterative procedure that successively tries works to get closer and closer to the correct answer.

Maximum Likelihood Estimation

- Sample $\mathcal{X} = \{ x^t \}_t$ where x^t is drawn from a known probability density function $p(x \mid \theta)$, defined upto parameters θ.
- Parametric estimation: We want to find θ that makes sampling x^t as likely as possible.
- Likelihood of θ given the sample $\mathcal{X}$

$$l(\theta \mid \mathcal{X}) = p(\mathcal{X} \mid \theta) = \prod_t p(x^t \mid \theta)$$

Examples: Bernoulli distribution

- Bernoulli:
- Two states, failure/success, x in {0,1}
- $P(x) = p^x (1 - p)^{(1-x)}$
- $\mathcal{L}(p|\mathcal{X}) = \log \prod_t p^{x^t} (1 - p)^{(1-x^t)}$
- MLE is maximizes this $\mathcal{L}$ i.e. $\partial \mathcal{L} / \partial p = 0$
- Hence MLE= $p_o = \sum_t x^t / N$

Examples: Multinomial distribution

- Generalization of Bernoulli where instead of two states, he outcome of a random event is one of K mutually exclusive and exhaustive states
- K>2 states, x_i in {0,1}
- $P\,(x_1,x_2,\ldots,x_K) = \prod_i p_i^{x_i}$

Examples: Multinomial distribution

- Let us say we do N such independent experiments with outcomes $\mathcal{X} = \{ x^t \}_{t=1}^{N}$ where $x_i^t = 1$ if experiment chooses state I,

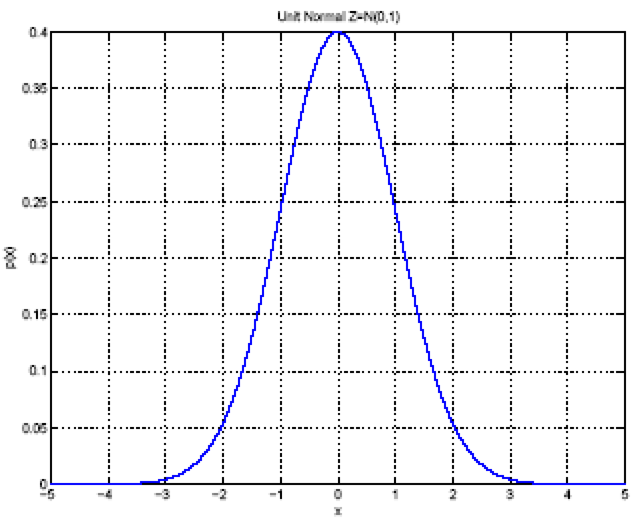
  = 0 otherwise
- $\mathcal{L}(p_1,p_2,\ldots,p_K \mid \mathcal{X}) = \log \prod_t \prod_i p_i^{x_i^t}$
- MLE: $p_i = \sum_t x_i^t / N$

Examples: Gaussian (Normal) distribution

- $p(x) = \mathcal{N}(\mu, \sigma^2)$

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

μ



- MLE for μ and σ^2:

$$m = \frac{\sum_t x^t}{N}$$

$$s^2 = \frac{\sum_t (x^t - m)^2}{N}$$

Assessing the Model

$$\log-\text{likelihood}=\sum_{i=1}^{N}\left[Y_i\ln(P(Y_i))+(1-Y_i)\ln(1-P(Y_i))\right]$$

- The Log-likelihood statistic
  - Analogous to the residual sum of squares in multiple regression
  - It is an indicator of how much unexplained information there is after the model has been fitted.
  - Large values indicate poorly fitting statistical models.

Assessing Changes in Models

- It's possible to calculate a log-likelihood for different models and to compare these models by looking at the difference between their log-likelihoods.

$$\chi^2 = 2\left[LL(New) - LL(Baseline)\right]$$

$$\left(df = k_{new} - k_{baseline}\right)$$

Assessing Predictors: The Wald Statistic

$$\boxed{Wald\_stat = \frac{b}{SE_b}}$$

- Similar to *t*-statistic in Regression.
- Tests the null hypothesis that $b = 0$.
- Is biased when b is large.
- Better to look at Likelihood-ratio statistics.

Assessing Predictors: The Odds Ratio or Exp(*b*)

$$\text{Exp(b)} = \frac{\text{Odds after a unit change in the predictor}}{\text{Odds before a unit change in the predictor}}$$

- Indicates the change in odds resulting from a unit change in the predictor.
  - OR > 1: Predictor ↑, Probability of outcome occurring ↑.
  - OR < 1: Predictor ↑, Probability of outcome occurring ↓.

Methods of Regression

- Forced Entry: All variables entered simultaneously.
- Hierarchical: Variables entered in blocks.
  - Blocks should be based on past research, or theory being tested. Good Method.
- Stepwise: Variables entered on the basis of statistical criteria (i.e. relative contribution to predicting outcome).
  - Should be used only for exploratory analysis.

Things That Can go Wrong

- Assumptions from Linear Regression:
  - Linearity
  - Independence of Errors
  - Multicollinearity
- Unique Problems
  - Incomplete Information
  - Complete Separation
  - Overdispersion

Incomplete Information From the Predictors

- Categorical Predictors:
  - Predicting cancer from smoking and eating tomatoes.
  - We don't know what happens when non-smokers eat tomatoes because we have no data in this cell of the design.
- Continuous variables
  - Will your sample contain a to include an 80 year old, highly anxious, Buddhist left-handed cricket player?

Do you smoke?	*Do you eat tomatoes?*	*Do you have cancer?*
Yes	No	Yes
Yes	Yes	Yes
No	No	Yes
No	Yes	??????

Complete Separation

- When the outcome variable can be perfectly predicted.
  - E.g. predicting whether someone is a burglar or your teenage son or your cat based on weight.
  - Weight is a perfect predictor of cat/burglar unless you have a very fat cat indeed!

Overdispersion

- Overdispersion is where the variance is larger than expected from the model.
- This can be caused by violating the assumption of independence.
- This problem makes the standard errors too small!