Chapter 6

Simple Logistic Regression

6.1 General Regression Setting

- Types of variables
  - Binary data: e.g. sex, death
  - Nominal (unordered categorical) data: e.g. race, martial status
  - Ordinal (ordered categorical data): e.g. cancer stage, asthma severity
  - Quantitative data: e.g. age, blood pressure
  - Right censored data: e.g. time to death

- The measures used to summarize and compare distributions vary according to the type of variable
  - Means: Binary, quantitative
  - Medians: Ordered, quantitative, censored
– Proportions: Binary, nominal, ordinal

– Odds: Binary, nominal, ordinal

– Hazards: Censored
  * Hazard is the instantaneous risk of failure

  * The hazard is an unusual idea– would it even have been invented if not for censored data? Probably not.

· Which regression model you choose to use is based on the parameter being compared across groups
  
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Regression Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Means</td>
<td>Linear regression</td>
</tr>
<tr>
<td>Geometric means</td>
<td>Linear regression on log scale</td>
</tr>
<tr>
<td>Odds</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>Rates</td>
<td>Poisson regression</td>
</tr>
<tr>
<td>Hazards</td>
<td>Proportional Hazards (Cox) regression</td>
</tr>
</tbody>
</table>

· General notation for variables and parameters

  \[
  Y_i \quad \text{Response measured on the } i\text{th subject} \\
  X_i \quad \text{Value of the predictor measured on the } i\text{th subject} \\
  \theta_i \quad \text{Parameter summarizing distribution of } Y_i | X_i
  \]

  – The parameter (\(\theta_i\)) might be the mean, geometric mean, odds, rate, instantaneous risk of an event (hazard), etc.

  – In linear regression on means, \(\theta_i = E[Y_i | X_i]\)

  – Choice of correct \(\theta_i\) should be based on scientific understanding of problem

· General notation for simple regression model
\[ g(\theta_i) = \beta_0 + \beta_1 \times X_i \]

- The link function is often either the identity function (for modeling means) or log (for modeling geometric means, odds, hazards)
  - Identity function: \( f(x) = x \)

### 6.1.1 Uses of General Regression

- **Borrowing information**
  - Use other groups to make estimates in groups with sparse data
    - Intuitively, 67 and 69 year olds would provide some relevant information about 68 year olds
    - Assuming a straight line relationship tells us about other, even more distant, individuals
  - If we do not want to assume a straight line, we may only want to borrow information from nearby groups

- **Defining “Contrasts”**
  - Define a comparison across groups to use when answering scientific questions
  - If the straight line relationship holds, the slope is the difference in parameter between groups differing by 1 unit in \( X \)
  - If a non-linear relationship in parameter, the slope is still the average difference in parameter between groups differing by 1 unit in \( X \)
    - Slope is a (first order or linear) test for trend in the parameter
CHAPTER 6. SIMPLE LOGISTIC REGRESSION

· Statistical jargon: “a contrast” across groups

· The major difference between different regression models is the interpretation of the parameters
  – How do I want to summarize the outcome?
    · Mean, geometric mean, odds, hazard
  – How do I want to compare groups?
    · Difference, ratio
  – Answering these two simple questions provides a starting road-map as to which regression model to choose

· Issues related to the inclusion of covariates remains the same
  – Address the scientific question: Predictor of interest, effect modification
  – Address confounding
  – Increase precision

6.2 Simple Logistic Regression

6.2.1 Uses of logistic regression

· Use logistic regression when you want to make inference about the odds

· Allows continuous (or multiple) grouping variables
  – Is OK with binary grouping variables too

· Compares odds of responses across groups using ratios
CHAPTER 6. SIMPLE LOGISTIC REGRESSION

- “Odds ratio”

• Binary response variable
  - When using regression with binary response variables, we typically model the (log) odds using logistic regression
    * Conceptually there should be no problem modeling the proportion (which is the mean of the distribution)

    * However, there are several technical reasons why we do not use linear regression very often with binary responses

• Why not use linear regression for binary responses?
  - Many misconceptions about the advantages and disadvantages of analyzing the odds

  - Reasons I consider valid
    * Scientific basis
      · Uses of odds ratios in case control studies

      · Plausibility of linear trends and no effect modifiers

    * Statistical basis
      · There is a mean variance relationship (if not using robust SE)

6.2.2 Reasons to use logistic regression

• First (scientific) reason: Case-Control Studies
  - Studying a rare disease, so we do study in reverse
    * e.g. find subjects with cancer (and suitable controls) and then ascertain exposure of interest
- Estimate distribution of the “effect” across groups defined by “cause”  
  * e.g. proportion (or odds) of smokers among people with or without lung cancer

  * In contrast, a cohort study samples by exposure (smoking) and then estimates the distribution of the effect in exposure groups

- In a case-control study, we cannot estimate prevalence (without knowing selection probabilities)  
  * e.g. if doing a 1:1 case-control study, it would look like 50% of the subjects have cancer

- Odds ratios are estimable in either case-control or cohort sampling scheme

  * Cohort study: Odds of cancer among smoker compared to odds of cancer among nonsmokers

  * Case-control study: Odds of smoking among cancer compared to odds of smoking among non-cancer

  * Mathematically, these two odds ratios are the same

- Odds ratios are easy to interpret when investigating rare events  
  * Odds = prob / (1 - prob)

  * For rare events, (1 - prob) is approximately 1  
    · Odds is approximately the probability

    · Odds ratios are approximately risk ratios

  * Case-control studies usually used when events are rare

  · Second (scientific) reason: Linearity
– Proportions are bounded by 0 and 1

– It is thus unlikely that a straight line relationship would exists between a proportion and a predictor
  * Unless the predictor itself is bounded

  * Otherwise, there eventually must be a threshold above which the probability does not increase (or only increases a little)

• Third (scientific) reason: Effect modification

– The restriction on ranges for probabilities makes it likely that effect modification *must* be present with proportions

– Example: Is the association between 2-year relapse rates and having a positive scan modified by gender?
  * Women relapse 40% of the time when the scan is negative, and 95% of the time when the scan is positive (an increase of 55%)

  * If men relapse 75% of the time when the scan is negative, then a positive scan can increase the relapse rate to at most 100%, which is only a 25% increase

<table>
<thead>
<tr>
<th>Proportion</th>
<th>Women</th>
<th>Men</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative Scan</td>
<td>0.40</td>
<td>0.75</td>
</tr>
<tr>
<td>Positive Scan</td>
<td>0.95</td>
<td>(up to 1.00)</td>
</tr>
<tr>
<td>Diff</td>
<td>0.55</td>
<td>(≤ 0.25)</td>
</tr>
<tr>
<td>Ratio</td>
<td>1.64</td>
<td>(≤ 1.33)</td>
</tr>
</tbody>
</table>

– With the odds, the association can hold without effect modification
<table>
<thead>
<tr>
<th>Odds</th>
<th>Women</th>
<th>Men</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative Scan</td>
<td>0.67</td>
<td>3</td>
</tr>
<tr>
<td>Positive Scan</td>
<td>19 (up to $\infty$)</td>
<td></td>
</tr>
<tr>
<td>Ratio</td>
<td>28.5 ($&lt; \infty$)</td>
<td></td>
</tr>
</tbody>
</table>

- Fourth (statistics) reason:
  - Classical linear regression requires equal variances across each predictor group
    * But, with binary data, the variance within a group depends on the mean
  
  * For binary $Y$, $E(Y) = p$ and $Var(Y) = p(1 - p)$

- With robust standard errors, the mean-variance relationship is not a major problem. However, a logistic model that correctly models the mean-variance relationship will be more efficient.

### 6.2.3 The simple logistic regression model

- Modeling the odds of binary response variable $Y$ on predictor $X$

  Distribution $\Pr(Y_i = 1) = p_i$

  Model $\text{logit}(p_i) = \log \left( \frac{p_i}{1-p_i} \right) = \beta_0 + \beta_1 \times X_i$

  $X_i = 0$ \hspace{1cm} log odds = $\beta_0$

  $X_i = x$ \hspace{1cm} log odds = $\beta_0 + \beta_1 \times x$

  $X_i = x + 1$ \hspace{1cm} log odds = $\beta_0 + \beta_1 \times x + \beta_1$

- To interpret as odds, exponentiate the regression parameters
CHAPTER 6. SIMPLE LOGISTIC REGRESSION

Distribution
\[ \Pr(Y_i = 1) = p_i \]

Model
\[ \frac{p_i}{1-p_i} = \exp(\beta_0 + \beta_1 \times X_i) = e^{\beta_0} \times e^{\beta_1 \times X_i} \]

- To interpret as proportions (remember proportion = odds / (1 + odds))

Distribution
\[ \Pr(Y_i = 1) = p_i \]

Model
\[ p_i = \frac{e^{\beta_0} e^{\beta_1 \times X_i}}{1 + e^{\beta_0} e^{\beta_1 \times X_i}} \]

- Most common interpretations found by exponentiating the coefficients
  - Odds when predictor is 0 found by exponentiating the intercept: \( \exp(\beta_0) \)
  - Odds ratio between groups differing in the values of the predictor by 1 unit found by exponentiating the slope: \( \exp(\beta_1) \)

- Stata commands
  - “logit respvar predvar, [robust]”
    * Provides regression parameter estimates an inference on the log odds scale (both coefficients with CIs, SEs, p-values)
  - “logistic respvar predvar, [robust]”
    * Provides regression parameter estimates and inference on the odds ratio scale (only slope with CIs, SEs, p-values)
• R Commands
  – With rms package, \texttt{lrm(respvar \sim predvar, ...)}

  – In general, \texttt{glm(respvar \sim predvar, family=\textquoteleft \textquoteleft binomial\textquoteleft \textquoteleft)}

6.3 Example: Survival on the Titanic and Age

• Dataset at http://biostat.mc.vanderbilt.edu/DataSets

• Describe the survival status of individual passengers on the Titanic

• Data on age available for many, but not all, subjects (data continually being updated)

• Response variables is Survival
  – Binary variable: 1=Survived, 0=Died

• Predictor variable is Age
  – Continuous grouping variable

• Effect modification by gender
6.3.1 Descriptive Plots
CHAPTER 6. SIMPLE LOGISTIC REGRESSION

Survival with lowess line

Survived (1=Yes, 0=No)

Age (years)
Survival by gender with lowess lines

Survived (1=Yes, 0=No)

Age (years)
Figure obtained using `Hmisc::plsmo()` in R
Comments on the plots
- Age is missing for many subjects, which we will not worry about in the following analysis
- The simple scatterplot, even with superimposes lowess smooth, is clearly worthless
- More advanced plotting available in R (in this case, the plsmo() function) can help to visualize the data

6.3.2 Regression Model

- Regression model for survival on age (ignoring possible effect modification for now)
  - Answer question by assessing linear trends in log odds of survival by age
  - Estimate the best fitting line to log odds of survival within age groups
    * logodds(Survival|Age) = \( \beta_0 + \beta_1 \times \text{Age} \)
  - An association will exist if the slope \( \beta_1 \) is nonzero
    * In that case, the odds (and probability) of survival will be different across different age groups
CHAPTER 6. SIMPLE LOGISTIC REGRESSION

. logit survived age

Iteration 0: log likelihood = -707.31022
Iteration 1: log likelihood = -705.69166
Iteration 2: log likelihood = -705.69139

Logistic regression                   Number of obs = 1046
LR chi2(1) = 3.24                       Prob > chi2 = 0.0720
Log likelihood = -705.69139            Pseudo R2 = 0.0023

------------------------------------------------------------------------------
survived | Coef. Std. Err. z  P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
age | -.0078985  .0044065 -1.79  0.073   -.0165351   .0007381
_cons | -.136534   .1447153 -0.94  0.345   -.4201708   .1471028
------------------------------------------------------------------------------

· logodds(Survival|Age) = -0.1365 - 0.007899 × Age

· General interpretation
  – Intercept is labeled “_cons”
  – Slope for age is labeled “age”

· Interpretation of intercept
  – Estimated log odds for newborns (age=0) is -0.136534
    * Odds of survival for newborns is $e^{-0.136534} = 0.8724$

    * Probability of survival
      · Prob = odds / (1 + odds)

        · $0.8724/(1 + 0.8724) = 0.4659$

· Interpretation of slope
  – Estimate difference in the log odds of survival for two groups differing by one year in age is -0.0078985
– This estimate averages over males and females

– Older groups tend to have lower log odds
  * Odds Ratio: $e^{-0.0078985} = 0.9921$
  * For five year difference in age: $e^{-0.0078985 \times 5} = 0.9612$
  * In Stata use “lincom age, or” or “lincom 5*age, or”

– Note that if the straight line relationship does not hold true, we interpret the slope as an average difference in the log odds of survival per one year difference in age

· There are two ways to get the odds ratio in Stata
  – Use the logit command to fit the model, followed by lincom post-estimation commands
  
  – Note that standard errors (and CIs) for odds ratios are calculated using the delta method
  
  – Alternatively, skip the log odds scales and use the logistic command to just get the odds ratio so we don’t have to bother exponentiating the slope coefficients
. lincom age, or
( 1) age = 0

survived | Odds Ratio Std. Err. z P>|z| [95% Conf. Interval]
---------+-------------------------------------------------------------
(1) | .9921326 .0043718 -1.79 0.073 .9836009 1.000738

. lincom 5*age, or
( 1) 5 age = 0

survived | Odds Ratio Std. Err. z P>|z| [95% Conf. Interval]
---------+-------------------------------------------------------------
(1) | .9612771 .0211794 -1.79 0.073 .9206499 1.003697

. logistic survived age

Logistic regression
Number of obs = 1046
LR chi2(1) = 3.24
Prob > chi2 = 0.0720
Log likelihood = -705.69139 Pseudo R2 = 0.0023

survived | Odds Ratio Std. Err. z P>|z| [95% Conf. Interval]
---------+-------------------------------------------------------------
age | .9921326 .0043718 -1.79 0.073 .9836009 1.000738

6.3.3 Comments on Interpretation

· The slope for age is expressed as a difference in group means, not the difference due to aging. We did not do a longitudinal study in which repeated measurements were taken on the same subject.

· If the group log odds are truly linear, then the slope has an exact interpretation as the change in survival due to a one year change in (any) age
  – Otherwise, the slope estimates the first order trend of the sample data and we should not treat the estimates of group odds or probabilities as
It is difficult to see in the above example, but the CIs around the odds ratios are not symmetric.

- (Symmetric) CIs are calculated on the log odds scale, and then transformed to the odds scale by exponentiating the lower and upper limits of the CI.

- “From logistic regression analysis, we estimate that for each 5 year difference in age, the odds of survival on the Titanic decreased by 3.9%, though this estimate is not statistically significant ($p = 0.07$). A 95% CI suggests that this observation is not unusual if a group that is five years older might have an odds of survival that was anywhere between 7.9% lower and 0.4% higher than the younger group.

- The confidence interval and statistical test given in the Stata output is called a Wald test. Other tests (Score, Likelihood Ratio) are also possible.

  - All tests are asymptotically equivalent

  - The Wald test is easiest to obtain, but generally performs the poorest in small sample sizes

  - The Likelihood Ratio test performs the best in small samples. We will discuss it later, including how to obtain the test using post-estimation commands.

  - The Score test is not bad in small samples, but is often hard to obtain from software. It is exactly equal to the Chi-squared test for binary outcomes and categorical predictors.
6.4 Inference with Logistic Regression

- The ideas of Signal and Noise found in simple linear regression do not translate well to logistic regression
  - We do not tend to quantify an error distribution with logistic regression

- Valid statistical inference (CIs, p-values) about associations requires three general assumptions
  - Assumption 1: Approximately Normal distributions for the parameter estimates
    * Large N
      - Need for either robust standard errors or classical logistic regression
    - Definition of large depends on the underlying probabilities (odds)
    - Recall the rule of thumb for chi-squared tests based on the expected number of events
  
  - Assumption 2: Assumptions about the independence of observations
    * Classical regression: Independence of all observations
    * Robust standard errors: Correlated observations within identified clusters

  - Assumption 3: Assumptions about variance of observations within groups
    * Classical regression: Mean-variance relationship for binary data
      - Classical logistic regression estimates SE using model based estimates
    - Hence in order to satisfy this requirement, linearity of log odds across groups must hold
CHAPTER 6. SIMPLE LOGISTIC REGRESSION

- Robust standard errors
  - Allows unequal variance across groups
  - Hence, do not need linearity of log odds across groups to hold

- Valid statistical inference (CIs, p-values) about odds of response in specific groups requires a further assumption
  - Assumption 4: Adequacy of the linear model
    - If we are trying to borrow information about the log odds from neighboring groups, and we are assuming a straight line relationship, the straight line needs to be true

- Needed for either classical or robust standard errors

- Note that we can model transformations of the measured predictor

- Inference about individual observations (prediction intervals, P-values) in specific groups requires no further assumptions because we have binary data
  - For binary data, if we know the mean (proportion), we know everything about the distribution including the variance

6.4.1 Implications for Inference

- The Moral: Regression based inference about associations is far more trustworthy than estimation of group odds of responses. Now, the story...

- We will now consider a hierarchy of null hypotheses
  - Strong (and intermediate) null: Total independence of $Y$ and $X$
    - A binary distribution only depends on the mean, which is why these two nulls are combined
– Weak null: No linear trend in log odds of $Y$ across $X$ groups

· Under Strong Null (response and predictor are totally independent)
  – Probability of response, and hence the odds and log odds, would be the same in all groups
    * A flat line would describe the log odds response across groups (a linear model is correct, and the slope is zero)
    * Within group variance is correctly estimated by the model
    * In large sample sizes, the regression parameters are Normally distributed

· Under Weak Null
  – Linear trend in log odds across predictor groups would lie on a flat line
    * Slope of best fitting line is zero
    * Within group variance could vary across groups as predicted by the model
    * In large sample sizes, the regression parameters are Normally distributed

· Classical Logistic Regression
  – Inference about the slope tests the strong null
    * All tests make inference by assuming the strong null is true
    * If the data appear non-linear on the log odds scale, merely evidence the strong null is not true

– Limitations
* We cannot be confident that there is a trend in log odds across groups (valid inference about the trends demands a correct model)

* We cannot be confident in estimate of group probabilities/odds (valid estimates of group probabilities require a correct model)

· Robust Standard Errors
  – Inference about the slope tests the weak null
    * All tests make inference by assuming the weak null is true

* Data can appear non-linear in log odds
  · Robust SE estimate true variability
    · Nonlinearity decreases precision, but inference about first-order trends still valid

· So, which inference is correct?
  – Classical logistic regression and logistic regression with robust standard errors differ in the strength of necessary assumptions

  – As a rule, if all of the assumptions of classical logistic regression hold, it will be more precise
    * Hence, we will have the greatest precision to detect associations if the linear model is correct

  – The robust standard error methods are valid for detection of associations without relying on the classical assumptions

· Back to the Moral: Regression based inference about associations is far more trustworthy than estimation of group means or individual predictions
  – A non-zero slope suggests an association between the response and predictor
– If use robust SE, can also make inference about linear trends in the log odds

• Interpreting “Positive” Results
– Slope is statistically different from 0 using robust standard errors
  * Observed data is atypical of a setting with no linear trend in odds of response across groups

  * Data suggests evidence of a trend toward larger (or smaller) odds in groups having larger values of the predictor

  * (To the extent the data appears linear, estimates of the group odds or probabilities will be reliable)

• Interpreting “Negative” Results
– Many possible reasons why the slope is not statistically different from 0 using robust standard errors
  * There may be no association between the response and predictor

  * There may be an association, but not in the parameter considered (the log odds of response)

  * There may be an association in the parameter considered, but the best fitting line has zero slope

  * There may be a first order trend in the log odds, but we lacked the precision to be confident that it truly exists (a type II error)

6.5 Example analysis revisited: Effect Modification

• Analysis conducted in class