Chapter 14: Generalized Linear Models (GLMs)

• GLMs are a useful general family of models having three characteristics:

(1) The response values $Y_1, \ldots, Y_n$ are independent and follow a distribution that is in the exponential family; i.e., the density may be written in the form:

Note: Using this form,

(2) The model has a linear predictor (based on the predictor variables $X_1, \ldots, X_k$) denoted:

(3) There is a monotone link function $g(\cdot)$ that relates the mean response $E(Y_i) = \mu_i$ to the linear predictor:

Note: Our classical regression model for normal data,

is a GLM:

Why?

(1) Normal distribution is in the exponential family:

(2) A linear predictor is clearly used.

(3) It uses the "identity" link function:

• We now study GLMs for two other common types of data.

Logistic Regression

• First we consider situations in which the response variable is binary (has two possible outcomes).

Example 1: Study of the effect of various predictors (age, weight, cholesterol, smoking level) on the incidence of heart disease. For each individual, the response $Y = 1$ if the person developed heart disease, and $Y = 0$ if no heart disease.

Example 2: We examine the effect of study habits on passing the state driver's test. For each examinee, the response is $Y = 1$ if the examinee passed the test, and $Y = 0$ if the examinee failed the test.

• We assume each Y_i is a Bernoulli r.v. with

Therefore

• **If we were to use a standard regression model, say,**
$E(Y_i) = \beta_0 + \beta_1 X_i$, then

Problems with using the standard model:

(1) Errors are clearly non-normal since Y_i can only be 0 or 1.

(2) Error variance is not constant.
• **A Bernoulli r.v. has variance**

• **If $E(Y) = \pi = \beta_0 + \beta_1 X$, then this variance is**

→

→

(3) Most importantly, since $E(Y)$ is a probability here, it should always be between 0 and 1.
• **For the model $E(Y) = \beta_0 + \beta_1 X$,**

• **A better model for binary data is the <u>Logistic Mean Response Model</u>:**

• **This function is constrained to fall between 0 and 1.**
• **It has a sigmoidal ("S") shape.**
• **It approaches 0 or 1 at the left/right limits.**
• **It is monotone.**
• **The value of β_1 determines whether the function is increasing or decreasing:**

<u>Note</u>:

So the odds that $Y_i = 1$, defined as are:

under this model.

• So the log-odds that $Y_i = 1$ (also called the logit of π_i) is:

<u>Note</u>: This logistic regression model is a GLM.

(1) Y_i has a distribution in the exponential family:

(2) Linear predictor is present.
(3) The link function is the <u>logit</u>:

• We could use other link functions for binary data.

• Letting $g(\pi_i) = \Phi^{-1}(\pi_i)$, the inverse of a standard normal cdf), yields a <u>probit</u> model.
• Letting $g(\pi_i) = \ln[-\ln(1 - \pi_i)]$ yields a <u>complementary log-log</u> model.

• Logistic and probit models have a <u>symmetric property</u>: If the coding of 0's and 1's in the data is reversed, the signs of all coefficients are reversed. (c-log-log does not have this)

<u>Estimating a Simple Logistic Regression Model</u>

• The parameters β_0 and β_1 are generally estimated via maximum likelihood (we do not use ordinary least squares because of the nonconstant error variance problem).

• Estimates b_0 and b_1 may be found using SAS or R.

Fitted logistic model:

<u>**Example**</u> **(Programming Task data, Table 14.1):**

Y = completion of task:

X = amount of programming experience (in months)

From SAS's PROC LOGISTIC:

Example:

Interpreting b_1:

Example (Programming task):

<u>Note</u>:

Multiple Logistic Regression

• This simply extends the <u>linear predictor</u> to include several predictor variables:

• Again, maximum likelihood is used to find estimates b_0, b_1, …, b_k.

<u>Example</u> (Disease outbreak, modified from Table 14.3):

Y = disease status (1 = yes, 0 = no)
X_1 = age (quantitative)
X_2 = city sector of residence (qualitative, 0 or 1)

SAS example:

Note: When all predictors are qualitative, the logistic regression model is often called a **log-linear model** (very common in categorical data analysis).

Inferences About Regression Parameters

• To determine the significance of individual predictors on the binary response variable, we may use tests or CIs about the β_j's.

Testing whether all β_j's are zero (Likelihood Ratio Test)

• Use Full Model vs. Reduced Model approach.

Test statistic is:

L_R = maximized likelihood function under reduced model
L_F = maximized likelihood function under full model

For large samples, under H_0,

• Reject H_0 when full model is

• A similar full/reduced test can be used to test whether some (not all) predictor variables are needed.

SAS example (disease outbreak):

Test About a Single Parameter

- **To test whether a single predictor is useful, we could use a form of the LR test.**
- **Another approach is the Wald test.**

Note: For large samples, maximum likelihood estimates are approximately normal.

Hence, for any predictor X_j,

Hence to test

we may use:

- **Often computer packages will report the Wald chi-square statistic $(z^*)^2$ and use the χ^2_1 distribution to obtain the P-value.**

- **This is completely equivalent to the (two-sided) z-test.**

• An approximate (large-sample) 100(1 – α)% CI for β_j is:

and thus an approximate 100(1 – α)% CI for the odds ratio for predictor X_j is:

SAS example:

Model Selection

• This is done similarly as in linear regression.
• The SELECTION=STEPWISE option can be used in the MODEL statement.

• SAS gives values of

for each fitted model, where L = maximized likelihood function for that model.
• Again, models with small AIC and small BIC are preferred.

Tests for Goodness of Fit

• We typically wish to formally test whether the logistic model provides a good fit to the data.

• The Hosmer-Lemeshow test breaks the data into *c* classes (usually between 5 and 10) and compares the observed number of successes (Y = 1 values) in each class to the expected number under the logistic model.

• The Hosmer-Lemeshow test statistic has an approximate ___ distribution under

• A small p-value indicates the logistic model does not fit well.

• SAS and R will give P-values of the H-L test (see examples).

Residuals:

• In logistic regression, the ordinary residuals

are not too meaningful.

• The Pearson residuals are obtained by dividing by the estimated standard deviation of Y_i:

• The INFLUENCE option gives Pearson residuals and other diagnostic measures.

• A r_{p_i} value with large magnitude indicates a possible outlier.

CI for the "Mean Response" π_h

• For a particular *x*-value X_h (or set of values

we may wish to estimate

• A point estimate is obtained simply by

• If $s\{\hat{\pi}_h\}$ is the estimated standard error of $\hat{\pi}_h$, by maximum likelihood theory, for large samples:

→ A large-sample approximate 100(1 – α)% CI for π_h is:

• In practice, SAS or R will find these.

Example: Find a 90% CI for the probability that programmers with 10 months experience are successful at the task.

Predicting a New Observation

• A simple rule for predicting Y_h for a new observation having predictor values $\underline{X}_h$ is:

• This assumes outcomes 0 and 1 are equally likely in the population.

• Another option is to use a different cutoff than 0.5; use the cutoff for which the <u>fewest observations in the sample</u> are "misclassified".

Poisson Regression (Count Regression)

• This is used when the response variable *Y* represents a count (the number of occurrences of an event).

<u>Example 1</u>: number of trips to a grocery store per month by a household
<u>Example 2</u>: number of cars passing an intersection per minute

• When the counts in a data set are very large, we may view Y as an approximately normal r.v. and use standard linear regression.

• When counts are typically small to moderate, we should use specialized <u>count regression</u> methods.

• The Poisson regression model is a GLM appropriate for modeling counts:

• If Y ~ Poisson(μ), then

• The most common link function for Poisson regression is the

So

• Fitting the model (estimating β_0, β_1, …, β_k) is again done via maximum likelihood.

<u>Example</u> (Miller lumber): A store surveyed its customers from 110 census tracts.
• The response Y_i = the number of customers from each census tract, i = 1, …, 110.

• We model Y_i using a Poisson distribution.
• They also measured other variables for the 110 tracts.

Poisson regression of Y against X_1 = # of housing units:

• Inference about several parameters is again done with the Likelihood Ratio test.

• For large samples, approximate CIs and tests about individual parameters can be done with the Wald statistic.

<u>Miller lumber example</u>:

• Goodness of fit may be checked with the "residual deviance":

or Pearson's χ^2 statistic =

• These each have an approximate distribution when the Poisson is the correct model.

• Values of *Dev* or χ^2 much larger than $n - k - 1$ indicate a poor fit.

• The contributions of each observation to *Dev* or χ^2 are the "deviance residuals" or "Pearson residuals" and these are examined to detect outliers.

• Model selection is often based on AIC, as with logistic regression (see multiple Poisson regression example).

<u>Prediction</u>: SAS or R gives predicted mean response values, and CIs for $\hat{\mu}_i$.

<u>Example</u>: