STAT3015/7030:
Generalised Linear Modelling
Logistic regression for binomial counts

Bronwyn Loong

Semester 2 2014

References

Ch 21 - Ramsey and Schafer, Statistical Sleuth

Ch 6.3 - Gelman and Hill

Ch 2 - Faraway

Outline

Extend logistic regression model to the case where the responses are proportions from counts. For example, grouped binary responses (such as the proportion of subjects of a given age who develop lung cancer) or a proportion based on a count for each subject (such as the proportion of subjects for which the student received a HD). The response variable is more general, (that is, the number of "successes" out of $n > 1$ trials), but we still model the probability of success through the logit link, as a linear function of predictors.

Binomial Logistic Regression Model

$$Y \sim Bin(n_i, p_i)$$

$$Pr(Y_i = y_i) = \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{(n_i - y_i)}$$

The GLM model:

$$\eta_i = g(p_i) = X_i^T \beta$$

Choices for $g(p_i)$:

- Logit: $g(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$
- Probit: $g(p_i) = \Phi^{-1}(p_i)$
- Comp. log-log: $g(p_i) = \log(-\log(1 - p_i))$

Use maximum likelihood to obtain parameter estimates β, identify factors signficant in predicting the probability of a success.

Binomial Logistic Regression Model

What information about the response values do we need to fit the model in R? $\rightarrow$ y and n.

In R we can do this by forming a two-column matrix, with the first column for the number of successes y, and the second column for the number of failures $n - y$.

Example: Trout data

An experiment was conducted where boxes of trout eggs were buried at five different stream locations and retrieved at four different times, specified by the number of weeks after the original placement. The number of surviving eggs was recorded.
(see R code)

Model Inference

As before....

- Test for significance of single coefficient estimates, interpretation
- Goodness - of -fit test adequacy of overall model
- Drop in deviance test for subsets of predictors

See R code for case study example.

Model Diagnostics

- Scatterplots of empirical logits $\hat{\pi}_i = y_i/n_i$ versus predictors
- Residual plots
- Overdispersion test
  See R code for case study example.

Binomial GLM - OVERDISPERSION

Sometimes the data may exhibit more variation in the response than is predicted by the Binomial GLM model (eg clustering of events).

The dispersion parameter estimate $\hat{\phi}$ represents a multiplicative departure from binomial variation.

$E[Y_i|X_i] = n_i p_i$; $logit(p_i) = X_i^T \beta$

$$Var[Y_i|X_i] = \hat{\phi} n_i p_i (1 - p_i)$$

We want to account for this extra variation somehow in how our model - introduce a dispersion parameter to account for overdispersion (or sometimes underdispersion).

Binomial GLM - OVERDISPERSION

How to test for over/underdispersion

- Are the binary responses included in the count independent?
- Are observations with identical values of the explanatory variables likely to have different p_i's?
- Too many outliers
- Rejection of goodness of fit test

Binomial GLM - OVERDISPERSION

How to test for over/underdispersion - a numerical check

Define standardized residual

$$z_i = \frac{y_i - \hat{y}_i}{sd(\hat{y}_i)}$$

If the Binomial model is true, the z_i's should be approximately independent each with mean 0 and standard deviation 1.

If there is overdispersion, then we expect the z_i's to be larger in absolute value, reflecting the extra variation beyond what is predicted under the Binomial model.

estimated overdispersion $\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} z_i^2$.

Compare $\sum_{i=1}^{n} z_i^2$ to the quantiles of a χ^2_{n-p} distribution.

Binomial GLM - adjusting inferences for overdispersion

Multiply all standard errors by $\sqrt{\hat{\phi}}$.

OR

Fit the model using the `quasibinomial` family. Why quasi?
Because we do not fix the variance function $V(\mu) = \mu$.
Notes: because we have estimated an extra parameter ϕ

- The drop in deviance test statistics is divided by $\hat{\phi}$ to form an F-statistic. This is analagous to the extra-sum-of-squares F-test for ordinary regression, except that the deviance residuals (rather than ordinary residuals) are used.
- Use the t-distribution as the reference distribution for tests of significance of individual coefficient estimates.

(see R code)

Binary-data model as a special case of the binomial-count data model

Logistic regression for binary data ($Y_i = 0$ or 1) is a special case of the binomial form with $n_i \equiv 1$ for all i.

However, the concept of overdispersion does not exist for $n_i = 1$ because then y_i only takes on the values 0 or 1, and it must follow that $Y_i \sim Bern(p_i)$ and that the variance must be $Var(Y_i) = p_i(1 - p_i)$. There is no other distribution with support $\{0, 1\}$.

That is, if the original data are ungrouped binary data, then overdispersion cannot occur.

Binomial count data as a special case of the binary-data model

The binomial model for

$$Y \sim Bin(n_i, p_i) \text{ and } p_i = g^{-1}(X_i^T \beta)$$

can be expressed in binary-data form

$$Pr(Y_i = 1) = g^{-1}(X_i^T \beta)$$

Binomial count data as a special case of the binary-data model

By considering each of the n_i cases as a separate data point. What are the implications??

The sample size of the expanded regression is $\sum_i n_i$, and the data points are 0's and 1's

The X matrix is expanded to have $\sum_i n_i$ rows, where the i^{th} row of the original X matrix becomes n_i identical rows in the expanded matrix.

Binomial count data as a special case of the binary-data model

By grouping the binary data into binomial counts for each combination of explanatory variables, the data set is reduced to a smaller size, so it is often more convenient.

Grouping the binary data and modelling as binomial counts does not change the inferential results (that is tests and confidence intervals for coefficients are identical).

However, the residual deviance will change - why? - because the saturated model is different (there are different number of data points, hence parameters in the saturated model).

Also note, the goodness-of-fit test is not appropriate for binary data. - why? - because the chi-square approximation requires expected frequencies for each observation to exceed one (not possible by definition of binary data).