Applied Statistical Methods II

Chapter #14

Logistic Regression, Poisson Regression, and Generalized Linear Models

Part II

Remember Our Setting

- We have $i = 1, \ldots, n$ subjects/trials.
- The outcome for each subject/trial is a binary random variable Y_i.
- Think that $\pi_i = \Pr(Y_i = 1)$ depends on a set of covariates $X_{i1}, \ldots, X_{i(p-1)}$.
- Assume the logistic regression model:
  - Y_i are independent for $i = 1, \ldots, n$.
  - $Y_i \sim \mathit{Bernoulli}(\pi_i)$
  - $\text{logit}(\pi_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i(p-1)} = X_i^T \beta$

Model Fitting

- Get the MLE's $\hat{\beta}_0, \ldots, \hat{\beta}_{p-1}$ of $\beta_0, \ldots, \beta_{p-1}$. Must use a numerical optimization routine such as Newton-Raphson or Fisher Scoring.
- In R: function "glm". The default is to use Fisher Scoring, which is usually efficient. You will get practice using this function on homework.
- In SAS: There are a lot of procedures for fitting logistic regression models. GENMOD, CATMOD, LOGISTIC, GLIMMIX

Testing

- So far, we have points estimates. What about inference?
- Assume that we fit the simple logistic regression model with $\text{logit}(\pi_i) = \beta_0 + \beta_1 X_i$.
- A popular goal: determine if X_i is associated with Y_i.
  - $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$
  - H_0 : OR between $X_i + \delta_x$ and X_i is 1.
  - $H_0 : \pi_i$ does not depend on X_i.

Testing, cont.

- Another popular goal: test if a subject if $X_i = x$ has a $50/50$ chance of $Y_i = 1$
  - $H_0 : \pi = 1/2$
  - $H_0 : \text{odds} = 1$
  - $H_0 : \text{logit}(\pi) = 0$
  - $H_0 : \beta_0 + \beta_1 x = 0$

Multiple logistic regression

- The additive model: $\text{logit}(\pi_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$.
- Assumes that the effect of X_{i1} are the same for all values of X_{i2}.
- For a given value of $X_{i2} = x_2$, then one unit increase in X_{i1} leads to β_1 increase in log odds of having $Y = 1$.
- For a given value $X_{i2} = x_2$:
  - $\log\left(\frac{odds(X_{i1}+1, x_2)}{odds(X_{i1}, x_2)}\right) = \beta_1$.
  - Odds ratio for one unit increase in X_1 is $\exp(\beta_1)$
- If X_1 is a categorical variable, then the odds ratio has a very natural interpretation. Will see examples.

Interactions

- The model: $\text{logit}(\pi_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$ assumes that the odds ratio for X_{i1} are the same for all values of X_{i2}.
- What if the odds ratio of X_{i1} depends on the values of X_{i2}?
  - Introduce an interaction term.
- $\text{logit}(\pi_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_{12} X_{i1} X_{i2}$
- For a given value $X_{i2} = x_2$:
  - $\log\left(\frac{odds(X_{i1}+1, x_2)}{odds(X_{i1}, x_2)}\right) = \beta_1 + \beta_{12} x_2$.
  - Odds ratio for one unit increase in X_{i1} is $e^{\beta_1} e^{\beta_{12} x_2}$

Inference

- In general, tests will be based on the parameters β.
- Should convert hypotheses about probabilities to equations of parameters: in general, we can test linear combinations of parameters.

Three Types of Inference

- There are 3 types of tests for logistic regression based on MLE:
  - Wald tests
  - Likelihood ratio tests
  - Score tests
- Wald tests:
  - These are the most popular.
  - They are easy to compute.
  - Are not reliable when there are small sample sizes.
- Likelihood ratio tests
  - Most popular in testing if a group of parameters are simultaneously zero.
  - More robust to small sample sizes than Wald tests.
- Score tests
  - Not used as much in practice.
  - These are usually computationally inexpensive, although power is sometimes small.

Preliminaries

- Y has density function $f(Y;\theta)$, where θ is the parameter. Let $\ell(Y;\theta) = \log f(Y;\theta)$
- $U(\theta) = \nabla_\theta \ell(Y;\theta)$ is the score function.
- $E_\theta(U) = 0$ under regularity conditions. Example on board.
- Fisher-information:
  $\mathcal{I}(\theta) = E_\theta\{\nabla_\theta \ell(Y;\theta) \nabla_\theta \ell(Y;\theta)^T\} = Var(U)$
- With regularity conditions, $\mathcal{I}(\theta)$ can also be written as $-E_\theta\left(\nabla_\theta^2 \ell(Y;\theta)\right)$.

Score equation

Let $f(\theta; Y)$ be the density with respect to some measure μ and $\ell(\theta; Y)$ its log-likelihood.

$$E_\theta\{\nabla_\theta \ell(\theta; Y)\} = \int \nabla_\theta \ell(\theta; y) f(\theta; y)\, d\mu(y) = \int \nabla_\theta e^{\ell(\theta;y)} d\mu(y)$$

$$\underbrace{=}_{\text{Reg. conditions}} \nabla_\theta \int e^{\ell(\theta;y)} d\mu(y) = \nabla_\theta \int f(\theta; y)\, d\mu(y)$$

$$= \nabla_\theta 1 = 0$$

- The regularity conditions guarantee we can exchange integration and differentiation (i.e. when can we apply the Dominated Convergence Theorem).
- Typically satisfied whenever the support of $\ell(\theta; Y)$ does not depend on θ.
  - Example: check that we cannot exchange integration and differentiation for a uniform, i.e. $\ell(\theta; Y) = \frac{1}{\theta} 1\{Y \in (-\theta, \theta)\}$.

Fisher Information

$$0 = \nabla^2_\theta 1 = \nabla^2_\theta \int e^{\ell(\theta;y)} d\mu(y) \underbrace{=}_{\text{Reg.}\mathit{conditions}} \int \nabla^2_\theta e^{\ell(\theta;y)} d\mu(y)$$

$$= \int \nabla^2_\theta \ell(\theta; y)\, e^{\ell(\theta;y)} d\mu(y) + \int \nabla_\theta \ell(\theta; y) \{\nabla_\theta \ell(\theta; y)\}^T e^{\ell(\theta;Y)} d\mu(y)$$

$$= E\left\{\nabla^2_\theta \ell(\theta; Y)\right\} + E\left[\nabla_\theta \ell(\theta; Y) \{\nabla_\theta \ell(\theta; Y)\}^T\right]$$

Observed values

In Newton-Raphson approach, we have the observed score, the observed information I_{obs}, and the expected observed information I

- the observed score are for n independent (but not identically distributed) observations, so that $U_{obs} = \sum_{i=1}^{n} U_i = \sum_{i=1}^{n} \nabla_\beta \log f(Y_i; X_i, \beta)$, where the distribution of Y_i depends on X_i and β.
- $I_{obs} = -\sum_{i=1}^{n} \nabla_\beta^2 \{\log f(Y_i; X_i, \beta)\}$
- $E(I_{obs}) = \sum_{i=1}^{n} \mathcal{I}_i$, sometimes, we use the notation I_n.

Wald Tests

- Wald tests are based on the asymptotic normality of MLE $\hat{\theta}$ (from i.i.d data), that $\sqrt{n}(\hat{\theta} - \theta) \to N(0, \mathcal{I}^{-1}(\theta))$.
- In the GLM regression setting, $I_n(\beta) = E(I_{obs}(\beta))$, $\sqrt{n}(\hat{\beta} - \beta) \to N(0, (I(\beta)_{lim})^{-1})$, where $I(\beta)_{lim} = \lim_{n\to\infty} I_n(\beta)/n$.
  - See for example Appendix A of McCullagh and Nelder for details.
- $\hat{\beta} \overset{n/p \text{ large}}{\approx} N(\beta, I_n^{-1}(\beta))$.

Application to logistic regression

- $$\begin{aligned}\ell(\beta_0,\ldots,\beta_{p-1}) &= \sum_{i=1}^{n}\{Y_i\log(\pi_i)+(1-Y_i)\log(1-\pi_i)\}\\ &= \sum_{i=1}^{n}\left\{Y_i\sum_j x_{ij}\beta_j-\log\left[1+\exp\sum_j x_{ij}\beta_j\right]\right\}\end{aligned}$$
- $U_r = \frac{\partial l}{\partial \beta_r} = \sum_{i=1}^{n}(y_i-\pi_i)x_{ir}$
- $h_{rs} = \frac{\partial U_r}{\partial \beta_s} = -\sum_{i=1}^{n} x_{ir}x_{is}(1-\pi_i)\pi_i$.
- $I_n(\beta) = E(I_{obs}) = X'WX$.
- X is the $n\times p$ matrix with ij^{th} element X_{ij}.
- W is the diagonal matrix of $\pi(X_i,\beta)[1-\pi(X_i,\beta)]$.
- $\Rightarrow I_n(\beta) = X'\mathrm{Var}\,(Y)\,X$

- Wald test of $H_0 : \beta_j = C$ uses the test statistics
  - $Z = \frac{\hat{\beta}_j - C}{\hat{s}(\hat{\beta}_j)}$
  - $\hat{s}^2(\hat{\beta}_j) = \left[I_n(\hat{\beta})^{-1} \right]_{j+1,j+1}$
  - Under the null, $Z \approx N(0, 1)$
  - Can be used for one- and two- sided tests.
- Wald test of $H_0 : A\beta = C$, for $A \in R^{q \times p}$.
  - $A\hat{\beta} \sim N(A\beta, AI_n(\beta)^{-1}A^T)$ approximately.
  - $(A\hat{\beta} - A\beta)^T (AI_n(\hat{\beta})^{-1}A^T)^{-1}(A\hat{\beta} - A\beta) \sim \chi^2_q$ asymptotically.
  - In logistic regression, $I_n(\hat{\beta}) = X'WX$, with $W_{ii} = \hat{\pi}_i(1 - \hat{\pi}_i)$ and $\hat{\pi}_i = \frac{\exp(x_i^T \hat{\beta})}{1+\exp(x_i^T \hat{\beta})}$.

Likelihood Ratio Tests

- Assume you want to test
  $H_0 : \beta_j = 0$ vs. $H_a : \beta_j \neq 0$.
- The likelihood ratio test first fits
  - $\text{logit}(\pi_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i(p-1)}$
  - Gets the MLE $\hat{\beta}$.
  - This is the "full model".
- then fit the "reduced" model:
  - $\text{logit}(\pi_i) =$
    $\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{j-1} X_{i(j-1)} + \beta_{j+1} X_{i(j+1)} + \cdots + \beta_{p-1} X_{i(p-1)}$
  - Gets the MLE $\hat{\beta}^{(-j)}$.
- $\Lambda = -2\left[\ell(\hat{\beta}^{(-j)}) - \ell(\hat{\beta})\right] \sim \chi^2_1$ under H_0
- Reject the null at level $1 - \alpha$ when Λ is larger than the $(1 - \alpha)$ percentile of χ^2_1

- Can also test if a group of β's are simultaneously equal to zero.
- Test if $H_0 : \beta_{p-q} = \cdots = \beta_{p-1} = 0$.
- Or testing if the last q coefficients are zero.
- Let $\hat{\beta}^{-(p-q,\ldots,p-1)}$ be the MLE's for the model $\text{logit}(\pi_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-q-1} X_{i(p-q-1)}$
- $\Lambda = -2\left[\ell(\hat{\beta}^{-(p-q,\ldots,p-1)}) - \ell(\hat{\beta})\right] \sim \chi^2_q$ under H_0
- Reject the null at level $1-\alpha$ when Λ is larger than the $(1-\alpha)$ percentile of χ^2_q

Score Tests

- Score tests for $H_0 : \beta = C$ are based on the score statistics

$$\mathcal{I}^{-1/2}(C)U(C).$$

- Uses the idea that the score function should be close to zero for values near the true value.
- Under the null, the score statistic is approximately a standard normal.
- This is convenient, because we only have to estimate parameters from the full model, so it's fast.
- It's used frequently in genetics when computation is the bottleneck.

Confidence Intervals for Parameters

- Three type of confidence intervals:
  - Wald
  - Likelihood Ratio
  - Score
- The Wald confidence intervals are easiest to compute and used most.
  - Especially for linear combination of parameters.
- Likelihood Ratio CIs are used for small sample sizes.
  - Require an iterative procedure to compute.
  - SAS only computes these for parameters and not linear combinations of parameters.
  - Can be used for joint confidence regions of a few parameters.

Wald CI for Parameters

- Let z_p be the p^{th} percentile of $N(0, 1)$.
- A $(1 - \alpha) \times 100\%$ confidence interval for β_j is: $\hat{\beta}_j \pm z_{1-\alpha/2}\hat{s}(\hat{\beta}_j)$.
- A $(1 - \alpha)\%$ confidence interval for $\sum_{j=0}^{p-1} a_j\beta_j$ is: $\sum_{j=0}^{p-1} a_j\hat{\beta}_j \pm z_{1-\alpha/2}\hat{s}(\sum_{j=0}^{p-1} a_j\hat{\beta}_j)$.

Likelihood ratio CI for Parameters

- Assume we have $\text{logit}(\pi_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i(p-1)}$.
- We want a CI around β_j.
- Let ℓ_0 be the maximum log-likelihood.
- Let $\ell_j(\tilde{\beta}_j)$ be the maximum log-likelihood when β_j is set to be the value $\tilde{\beta}_j$.
- The two-sided $100(1-\alpha)$ % likelihood ratio CI for β_j is the set that satisfies:

$$\left\{ \tilde{\beta}_j : \ell_0 - \ell_j(\tilde{\beta}_j) < 0.5 * \chi^2_{1-\alpha,1} \right\}.$$

Why does this work?

- SAS starts with the mle of β_j and progressively search until reaches the endpoint.
- Could take some time for large data sets.

CI for Probabilities

- What if we want to do inference on the probability $\pi_i = \Pr(Y_i = 1|X_i)$?
- $\pi_i = \text{expit}\left(X_i^T \beta\right)$.
- Inference for $\pi_i = \Pr(Y_i = 1|X_{i1}, \ldots, X_{i(p-1)})$ are computed by
  1. Compute confidence interval for $X_i^T \beta$, usually Wald type confidence interval.
  2. CI for π_i is
     $$\text{expit}\Big[\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_{p-1} X_{i(p-1)} \pm z_{1-\alpha/2}\hat{s}(\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_{p-1} X_{i(p-1)})\Big]$$
  3. Note: CI for π_i is not centered around $\hat{\pi}_i$.

Likelihood ratio statistic and deviance

Let $\ell(Y;\theta)$ be the log-likelihood for $Y \in \mathbb{R}^n$ and $\theta \in \Theta_F \subseteq \mathbb{R}^p$. Let $\Theta_0 \subset \Theta_F$. The likelihood ratio statistic is defined as:

$$LR = -2\left[\sup_{\theta\in\Theta_0}\{\ell(Y;\theta)\} - \sup_{\theta\in\Theta_F}\{\ell(Y;\theta)\}\right].$$

- One will almost always assume that Θ_0 is “locally linear”, i.e. $\Theta_0 = \{\theta \in \Theta_F : h(\theta) = 0_{p-q}\}$ for some differentiable function $h : \mathbb{R}^p \to \mathbb{R}^{p-q}$ (i.e. Θ_0 is is q-dimensional).
- If $H_0 : \theta \in \Theta_0$ is true, then $LR \to \chi^2_{p-q}$ as $n \to \infty$ (under the proper regularity conditions).
- LR is equivalent to numerator of F-test $SSE(\text{null model}) - SSE(\text{full model})$.
- We therefore define the **deviance** to be $D_\Theta = 2\left[\ell_{\max} - \sup_{\theta\in\Theta}\{\ell(Y;\theta)\}\right]$. $LR = D_{\Theta_0} - D_{\Theta_F}$.
- Deviance is analogous to the SSE in linear regression.

Likelihood ratio statistic and deviance

Let $\ell(Y;\theta)$ be the log-likelihood for $Y \in \mathbb{R}^n$ and $\theta \in \Theta_F \subseteq \mathbb{R}^p$. Let $\Theta_0 \subset \Theta_F$. The likelihood ratio statistic is defined as:

$$LR = -2\left[\sup_{\theta\in\Theta_0}\{\ell(Y;\theta)\} - \sup_{\theta\in\Theta_F}\{\ell(Y;\theta)\}\right].$$

- One will almost always assume that Θ_0 is "locally linear", i.e. $\Theta_0 = \{\theta \in \Theta_F : h(\theta) = 0_{p-q}\}$ for some differentiable function $h : \mathbb{R}^p \to \mathbb{R}^{p-q}$ (i.e. Θ_0 is is q-dimensional).
- If $H_0 : \theta \in \Theta_0$ is true, then $LR \to \chi^2_{p-q}$ as $n \to \infty$ (under the proper regularity conditions).
- LR is equivalent to numerator of F-test $SSE(\text{null model}) - SSE(\text{full model})$.
- We therefore define the **deviance** to be $D_\Theta = 2\left[\ell_{\max} - \sup_{\theta\in\Theta}\{\ell(Y;\theta)\}\right]$. $LR = D_{\Theta_0} - D_{\Theta_F}$.
- Deviance is analogous to the SSE in linear regression.

Likelihood ratio statistic and deviance

Let $\ell(Y;\theta)$ be the log-likelihood for $Y \in \mathbb{R}^n$ and $\theta \in \Theta_F \subseteq \mathbb{R}^p$. Let $\Theta_0 \subset \Theta_F$. The likelihood ratio statistic is defined as:

$$LR = -2\left[\sup_{\theta\in\Theta_0}\{\ell(Y;\theta)\} - \sup_{\theta\in\Theta_F}\{\ell(Y;\theta)\}\right].$$

- One will almost always assume that Θ_0 is “locally linear”, i.e. $\Theta_0 = \{\theta \in \Theta_F : h(\theta) = 0_{p-q}\}$ for some differentiable function $h : \mathbb{R}^p \to \mathbb{R}^{p-q}$ (i.e. Θ_0 is is q-dimensional).
- If $H_0 : \theta \in \Theta_0$ is true, then $LR \to \chi^2_{p-q}$ as $n \to \infty$ (under the proper regularity conditions).
- LR is equivalent to numerator of F-test $SSE(\text{null model}) - SSE(\text{full model})$.
- We therefore define the **deviance** to be $D_\Theta = 2\left[\ell_{\max} - \sup_{\theta\in\Theta}\{\ell(Y;\theta)\}\right]$. $LR = D_{\Theta_0} - D_{\Theta_F}$.
- Deviance is analogous to the SSE in linear regression.

Horseshoe Crab Example

- Study looks at female horseshoe crabs.
- The outcome of interest is if there are any male crabs (called satellites) living near by.
- There are other variables in the data set, but we will focus on the width of the crab and color of the crab for now.
- There are 173 female crabs in the data set.
- $Y_i = 1$ if a female has satellites.
- Width is measured in centimeters.
- Let's look at some basic commands in R using "glm".

We find

- There is a significant association between width and the presence of satellites (LRT p-value < 0.001).
- The estimated increase in log-odds associated with an increase in width by 1cm is with an 95% Wald CI of .
- For one unit increase in width, the odds of having satellite is estimated to be times of the original odds.
- For width = 21 cm, the estimated odds of having male satellite is , the estimated probability of having male satellite is .

Multiple Logistic Regression

- Let's now look at the variable color as well.
- There are 4 different colors.
- The variable color is coded as 1,2,3,4.
- Want to know how having any satellites is associated jointly with color and width.
    - Assume that there is no interaction. What does this mean?
- Fit the model
    - $\text{logit}(\pi_i) \sim WIDTH_i + COLOR_i$, here $COLOR$ has four levels.
    - β_{colj} corresponds to crabs of color "j"
    - Proc genmod with param = GLM.

Questions We Are Asked to Answer.

Let's assume that our collaborator wants to know:

- If width is associated with the presence of satellites when controlling for color.
  - $H_0 : \beta_w = 0$
- If color is associated with the presence of satellites when controlling for width.
  - $H_0 : \beta_{col1} = \beta_{col2} = \beta_{col3} = \beta_{col4}$.
- The estimated odds ratio, while controlling for width, between crabs of color=1 and crabs of color=2.
  - $\exp(\widehat{\beta_{col1} - \beta_{col2}})$
- The estimated odds ratio, while controlling for width, between crabs of color=2 and crabs of color=4.
  - $\exp(\widehat{\beta_{col2} - \beta_{col4}})$

We find

- There is a significant association between width and the presence of satellites conditional on color (LRT p-value < 0.001).
  - The estimated increase in log-odds associated with an increase in width by 1cm while controlling for color is with an 95% Wald CI of .
  - For one unit increase in width, the odds of having satellite is estimated to be times of the original odds.

We find: (cont.)

- There is a moderately/marginally significant association between color and the presence of satellites conditional on width (LRT p-value = 0.07).
  - Wald test for the difference in log odds between color=1 and color=2 while controlling for width has a p-value of .
  - Wald test for the difference in log odds between color=2 and color=4 while controlling for width has a p-value of .
  - The estimated odds of having any satellites for color=1 is times the odds for color=2 with a 95% confidence interval that ranges from to .
  - The estimated odds of having any satellites for color=2 is times odds for color=4 with a 95% confidence intervals of [,].
  - These are some wide intervals. Probably do not have enough data to address these questions with proper power.

Example With Interactions

- Our collaborator needs to know if the odds ratio for width is different in different colors.
- Fit $\text{logit}(\pi_i) \sim WIDTH_i + COLOR_i + WIDTH_i\, COLOR_i$
- Wants to test the interaction term $\beta_{w,col1} = \beta_{w,col2} = \beta_{w,col3} = \beta_{w,col4}$.

We find:

- There is a no significant interaction between width and color in the odds of having any satellites with a LRT p-value of   .
- Usually we do not stepdown when the overall interaction not significant. But for illustration purpose, we also look at
    1. The estimated odds ratio of width in color=3 is   .
    2. The estimated odds ratio of width in color=4 is   .