

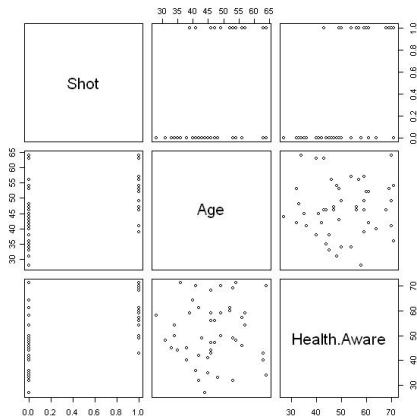
Lecture 14: Weighted Least Squares

Nancy R. Zhang

Statistics 191, Stanford University

March 5, 2008

Review - Binary responses data example



A clinic sent fliers to its clients to encourage everyone, but especially older persons, to get a flu shot for protection against an expected flu epidemic.

- 1 50 clients randomly sampled
- 2 Y: did they get flu shot?
- 3 Predictor variables: Age, health awareness.

Review - Binary responses model

Model: $Y \in \{0, 1\}$,

$$P(Y = 1 | X_1, \dots, X_p) = g^{-1}(\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p).$$

Where

$$g(\pi) = \log\left(\frac{\pi}{1 - \pi}\right).$$

The inverse g^{-1} is

$$g^{-1}(z) = \frac{e^z}{1 + e^z}.$$

We have no choice but to accept non-constant variance,

$$\text{Var}(Y) = \pi(X)[1 - \pi(X)].$$

Review - Model interpretation

An intuitive quantity to assess probabilities:

$$odds = \frac{P(Y = 1|X)}{P(Y = 0|X)}.$$

In the logistic regression model,

$$\log(odds) = \beta X.$$

The parameter β is the contribution of unit increase in X to the increase (decrease) in odds. For example, if X were binary as well,

$$\log\left(\frac{odds(X = 1)}{odds(X = 0)}\right) = \beta.$$

Review - Model fitting

Fitting can be done by Newton-Raphson:

- 1 Let $u' = \left(\frac{\delta l(\beta)}{\delta \beta_i}\right)_{i=1,\dots,p}$ be the gradient vector.
- 2 Let H be the Hessian matrix $h_{i,j} = \frac{\delta^2 l(\beta)}{\delta \beta_i \delta \beta_j}$.
- 3 Start with an initial $\beta^{(0)}$, then iterate $\beta^{(t+1)} = \beta^{(t)} - (H^{(t)})^{-1} u^{(t)}$.

The idea is, for each iteration t , to approximate $l(\beta)$ locally by a quadratic:

$$l(\beta) \approx l(\beta^{(t)}) + u^{(t)'}(\beta - \beta^{(t)}) + \frac{1}{2}(\beta - \beta^{(t)})' H^{(t)}(\beta - \beta^{(t)}),$$

and solve for $\delta l(\beta)/\delta \beta \approx u^{(t)} + H^{(t)}(\beta - \beta^{(t)}) = 0$.

For logistic regression model,

$$\beta^{(t+1)} = \beta^{(t)} + \{X' \text{diag}[\pi_i^{(t)}(1 - \pi_i^{(t)})]X\}^{-1} X'(y - \pi^{(t)}).$$

This is equivalent to doing a weighted linear regression at each step.

Inference for β

In Gaussian case:

$$\hat{\beta} = (X'X)^{-1}X'Y, \quad Y \sim N(X\beta, \sigma^2 I).$$

Since Gaussian vectors remain Gaussian under linear transforms,

$$\hat{\beta} \sim N(\beta, (X'X)^{-1}\sigma^2).$$

For logistic regression, $\hat{\beta}$ is no longer linear in Y . However, *asymptotically* (i.e. n large), it is Gaussian. It's covariance can be estimated by

$$\widehat{\text{cov}}(\hat{\beta}) = (X' \text{diag}[\hat{\pi}_i(1 - \hat{\pi}_i^{(t)})]X)^{-1}.$$

From the square root of the diagonal elements of the above matrix you can get $\widehat{\text{s.e.}}(\hat{\beta})$.

Wald tests for β

- 1 Confidence intervals for β :

$$[\hat{\beta} - z_{\alpha/2} \widehat{\text{s.e.}}(\hat{\beta}), \hat{\beta} + z_{\alpha/2} \widehat{\text{s.e.}}(\hat{\beta})]$$

- 2 Two sided test $H_0 : \beta = 0$, reject if

$$\left| \frac{\hat{\beta}}{\widehat{\text{s.e.}}(\hat{\beta})} \right| > z_{\alpha/2}$$

- 3 Test of constraint $H_0 : C_{j \times p} \beta_{p \times 1} = h_{j \times 1}$, reject if

$$(C\hat{\beta} - h)'(C\widehat{\text{cov}}(\hat{\beta})C')^{-1}(C\hat{\beta} - h)$$

is larger than $\chi_{j,1-\alpha}^2$.

Assessment of model fit

In linear regression, we used the F -test:

$$F = \frac{[SSE(RM) - SSE(FM)]/[\Delta df]}{SSE(FM)/[n - df(FM)]}.$$

$$F \sim F_{\Delta df, n - df(FM)}.$$

The analogous quantity of SSE for non-linear models is *deviance*:

$$\text{Deviance}(\hat{\beta}) = -2[l(\tilde{\beta}, Y) - l(\hat{\beta}, Y)],$$

where

$l(\cdot, Y)$ is log-likelihood,

$\tilde{\beta}$ is fit of data using saturated model (n predictors).

Assessment of model fit

For Gaussian model,

$$\text{Deviance}(\hat{\beta}) = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

For Logistic model,

$$\begin{aligned} \text{Deviance}(\hat{\beta}) &= 2 \sum_{i=1}^n \left[Y_i \log \frac{Y_i}{\hat{\pi}_i} + (1 - Y_i) \log \frac{1 - Y_i}{1 - \hat{\pi}_i} \right] \\ &= 2 \sum \text{observed} \times \log \frac{\text{observed}}{\text{fitted}}. \end{aligned}$$

Convention: $0 \times \log 0 = 0$.

Nested Chi-squared tests of model fit

As for SSE, the greater the deviance, the poorer the fit. If reduced model (RM) were true, then

$$\text{Deviance}(RM) - \text{Deviance}(FM) \rightarrow \chi^2_{df(FM) - df(RM)}.$$

Thus, reject RM at asymptotic level α if

$$\text{Deviance}(RM) - \text{Deviance}(FM) > \chi^2_{df(FM) - df(RM), 1 - \alpha}.$$

Model diagnosis

In linear regression, the standardized residuals were used to diagnose model fit.

$$r_i = y_i - \hat{y}_i, \quad r_i^* = \frac{r_i}{\hat{\sigma}_{r_i}} = \frac{r_i}{\sqrt{1 - \rho_{ii}}}.$$

The analogous quantity here is the Pearson residual,

$$r_i = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)}},$$

$$r_i^* = \frac{r_i}{\sqrt{1 - \hat{h}_{ii}}}.$$

where \hat{h}_{ij} are diagonals of

$$\text{Hat} = W^{1/2} X (X' W X)^{-1} X' W^{1/2}.$$

$$W = \text{diag}[\pi_i(1 - \pi_i)].$$

Model diagnosis

Another quantity you can use is the deviance residuals:

$$\begin{aligned} \text{Deviance}(\hat{\beta}) &= 2 \sum_{i=1}^n \left[Y_i \log \frac{Y_i}{\hat{\pi}_i} + (1 - Y_i) \log \frac{1 - Y_i}{1 - \hat{\pi}_i} \right] \\ &= 2 \sum \text{observed} \times \log \frac{\text{observed}}{\text{fitted}}. \end{aligned}$$

So let d_i be the contribution of data point i to the above measure of mis-fit:

$$d_i = Y_i \log \frac{Y_i}{\hat{\pi}_i} + (1 - Y_i) \log \frac{1 - Y_i}{1 - \hat{\pi}_i}$$

$$\text{Deviance residual: } \sqrt{|d_i|} \times \text{sign}(y_i - \hat{\pi}_i).$$

Model Selection

- 1 Nested models: Deviance χ^2 . Thus, reject RM at asymptotic level α if

$$\text{Deviance}(RM) - \text{Deviance}(FM) > \chi_{df(FM) - df(RM), 1 - \alpha}^2$$

$$\text{Deviance} = -2[l(\tilde{\beta}, Y) - l(\hat{\beta}, Y)]$$

- 2 Non-nested models:

- 1 AIC $-2\loglik + 2p$
- 2 C_p , but you would be doing a local asymptotic approximation using Pearson's residuals.

$$r_i = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)}}, \quad C_p = \sum_i r_i^2 + 2p\hat{\sigma}^2,$$

- 3 BIC $-2\loglik + p \log n$

Multinomial Data

If the response Y belongs to K categories.

① Designate one category as the “base” category.

②

$$P(Y = k|X) = \frac{e^{X\beta_k}}{1 + \sum_{l=1}^{K-1} e^{X\beta_l}}$$

Here, $\beta_k = (\beta_{k1}, \dots, \beta_{kp})'$.

$$P(Y = K|X) = \frac{1}{1 + \sum_{l=1}^{K-1} e^{X\beta_l}}$$

③ $p \times (K - 1)$ parameters.

④ β_{ki} for k -th category and i -th predictor interpreted as increase in log-odds from base category.

Count data

- 1 Men and women were asked whether they believed in the after life (1991 General Social Survey).

- 2 Results:

	Y	N or U	
M	435	147	582
F	375	134	509
Total	810	281	1091

- 3 Question: is belief in afterlife independent of gender?

Contingency Tables

	Y	N or U	
M	435	147	582
F	375	134	509
Total	810	281	1091

- 1 Model: $Y_{ij} \sim \text{Poisson}(\lambda_{ij})$.
- 2 H_0 : Independence. i.e. $\lambda_{ij} = \lambda\alpha_i\beta_j$.
- 3 H_A : λ_{ij} arbitrary.
- 4 Pearson's χ^2 Test:

$$\chi^2 = \sum_{i,j} \frac{(Y_{ij} - E_{ij})^2}{E_{ij}} \sim \chi_1^2 \quad (\text{under } H_0)$$

- 5 Why 1 df? Independence model has 5 (λ , 2 α 's, 2 β 's) parameters, 2 constraints \Rightarrow 3 df. Unrestricted model has 4 parameters.

Under independence:

$$\log E(Y_{ij}) = \log \lambda_{ij} = \log \lambda + \log \alpha_i + \log \beta_j.$$

What about variance? Because the data is Poisson,

$$\text{Var}(Y_{ij}) = E(Y_{ij}) = \lambda_{ij}.$$

Thus, the variance scales with the mean.

- Log stabilizes variance.
- But unlike before, we are explicitly modeling data as Poisson rather than Gaussian – added power if the data is indeed Poisson.

Why Poisson?

- 1 Count data is always > 0 .
- 2 Poisson distribution:

$$Poisson(k) = \sum_{i=1}^k Poisson(1)$$

By central limit theorem,

$$\frac{Poisson(k) - k}{\sqrt{k}} \rightarrow N(0, 1)$$

Thus “large Poissons are like Gaussians”. But small Poissons are quite different.

Similarities and differences with Gaussian, Logistic

① Mean = $g(X\beta)$.

① Gaussian: g is identity.

② Binomial: g is logit.

③ Poisson: g is log.

g is called the “link” function.

② Distribution of $Y \Rightarrow$ dependence of variance on mean.

① Gaussian: Variance constant in mean.

② Binomial: $Var(\pi) = \pi(1 - \pi)$.

③ Poisson: $Var(\lambda) = \lambda$.

There are many other models of this type, collectively called “generalized linear models.”

Contingency table - regression model

Suppose that we have a k by m table. After life example: $k = m = 2$. We call this a $k \times m$ contingency table.

1 Model:

$$Y_{ij} \sim \text{Poisson}(\lambda_{ij})$$

2 Mean function:

$$\log \lambda_{ij} = \log \lambda + \log \alpha_i + \log \beta_j$$

3 Pearson test for independence:

$$\chi^2 = \sum_{ij} \frac{(Y_{ij} - E_{ij})^2}{E_{ij}} \sim \chi_{k-1, m-1}^2 \quad (\text{under } H_0)$$

Poisson Regression

- 1 Model fitting: Newton Raphson.
- 2 Confidence intervals: same as for Binomial, use local Gaussianity.
- 3 Assessment of model fit: Deviance residuals.

Log-linear versus Logit models

- 1 Loglinear models are of use primarily when at least two variables are response variables. With a single categorical response, it is simpler and more natural to use logit models.
- 2 When you have two variables (e.g. Gender versus after-life belief), then logit might treat one as explanatory and the other as response, while there is an equivalent loglinear model.
- 3 Loglinear models view data as N independent cell counts rather than individual classifications of n subjects, $n = \sum_{i=1}^N Y_i$, and do not treat the row sums as fixed.

2 × 2 tables

	Y	N or U	
M	435	147	582
F	375	134	509
Total	810	281	1091

Model: $Y_{ij} \sim \text{Poisson}(\lambda_{ij})$.

If you have two Poissons, $\text{Poiss}(\lambda_{i1})$ and $\text{Poiss}(\lambda_{i2})$, then conditioned on their sum, each count is a binomial.

$$Y_{i,1} | Y_{i,1} + Y_{i,2} \sim \text{Binomial} \left(Y_{i,1} + Y_{i,2}, \frac{\lambda_{i1}}{\lambda_{i1} + \lambda_{i2}} \right)$$

Then,

$$\begin{aligned} \text{logit} P(1 | \text{row} = i, \text{row sum} = n_i) &= \log \frac{P(1 | \text{row} = i, \text{row sum} = n_i)}{P(2 | \text{row} = i, \text{row sum} = n_i)} \\ &= \log \frac{\lambda_{i1}}{\lambda_{i2}} \\ &= \log \lambda_{i1} - \log \lambda_{i2}. \end{aligned}$$

2 × 2 tables

$$\text{logit}P(1 | \text{row} = i, \text{row sum} = n_i) = \log \lambda_{i1} - \log \lambda_{i2}.$$

Under the null hypothesis:

$$H_0 : \lambda_{ij} = \lambda * \alpha_i * \beta_j,$$

$$\log \lambda_{ij} = \log \lambda + \log \alpha_i + \log \beta_j.$$

$$\text{logit}P(1 | \text{row} = i, \text{row sum} = n_i) = \log \beta_1 - \log \beta_2 \equiv \delta$$

The key is that the above logit does not depend on i . In binomial regression, we are modeling

$$\text{logit}P(1 | X) = \beta_0 + \beta_1 X.$$

So testing H_0 is equivalent to testing $\beta_1 = 0$ in logistic regression.

2 × 2 tables

Thus...

- 1 Testing the hypothesis $H_0 : \lambda_{ij} = \lambda * \alpha_i * \beta_j$ in the Poisson model is the same as testing independence in the logistic model.
- 2 To test this hypothesis, you fit the model with λ_{ij} arbitrary, and then use Chi-square test on the difference of deviances.
- 3 The difference of deviances will be the same as the logit model, but the absolute deviances will be different.