Regression Analysis

Poisson Regression

Nicoleta Serban, Ph.D.
Professor
Stewart School of Industrial and Systems Engineering

Goodness of Fit Assessment

1

About This Lesson

2

Poisson Regression Model

Data: $\{(x_{11},..,x_{1p}),Y_1\},....,\{(x_{n1},..,x_{np}),Y_n\}$ where $Y_1,..,Y_n$ are event count data per observation unit with a Poisson distribution

Assumptions:

- *Linearity Assumption*: $\log(E(Y|x_1, \ldots, x_p)) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$
- *Independence Assumption*: $Y_1,..,Y_n$ are independent random variables
- *Variance Assumption*: $E(Y|x_1, \ldots, x_p) = V(Y|x_1, \ldots, x_p)$

There is no error term! How to check the assumptions?

3

Residuals in Poisson Regression

Poisson Regression:

$Y_i|(x_{i1},..,x_{ip}) \sim \text{Poisson}(l(x_{i1},\ldots,x_{ip}))$

- Estimated rates are:
  $\hat{\lambda}_i = \hat{\lambda}(x_{i1,..,}x_{ip}) = e^{\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\ldots+\hat{\beta}_p x_{ip}}$
- Pearson Residuals:

$$r_i = \frac{Y_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}}$$

- Deviance Residuals:

$$d_i = 2\sum_{i=1}^{n}\left\{Y_i \log\left(\frac{Y_i}{\hat{\lambda}_i}\right) - (Y_i - \hat{\lambda}_i)\right\}$$

- Pearson's residuals follow directly a normal approximation to a binomial. Hence approximately N(0,1)
- The deviance residuals are the signed square root of the log-likelihood evaluated at the saturated model vs. the fitted model. Thus approximately N(0,1) if the model is a good fit.
- Deviances play the role of sum of squares in a linear model.

4

Goodness of Fit

GOF Visual Analytics:
- Normal Probability plot & Histogram of the Residuals
- Log of the event rate vs predictors

Hypothesis Testing Procedure:
$\mathrm{H_0}$: *the Poisson model fits the* data
$\mathrm{H_A}$: *the Poisson model does not fit the data*
Deviance test statistic: $\mathrm{D} = \sum_{i=1}^{n} d_i^2$
Under null hypothesis, $\mathrm{D} \sim \chi^2_{df}$ with $df = \text{n-p-1}$
Reject the null that the model is correct if $\text{p-value} = \mathrm{P}(\chi^2_{df} > \mathrm{D})$ small.
Note that for this test, we want large p-values!!!!

5

What if No Goodness of Fit?

- Add predicting variables, consider interaction terms, or/and transform predicting variables to improve linearity;
- Identify unusual observations (outliers, leverage points);
- The Poisson distribution isn't appropriate:
  - Overdispersion: the variability of the estimated rates is larger than would be implied by a Poisson model
    - Correlation in the observed responses
    - Heterogeneity in the rates that hasn't been modeled

6

Overdispersion

Overdispersion: the variability of the response variable is larger than would be implied by the model

Binomial regression model:
- $V(Y_i|x_1, \ldots, x_p) = n_i p(x_{i1}, \ldots, x_{ip})(1\text{-}p(x_{i1}, \ldots, x_{ip}))$
- Overdispersed Binomial: $V(Y_i|x_1, \ldots, x_p) = \phi\, n_i p(x_{i1}, \ldots, x_{ip})(1\text{-}p(x_{i1}, \ldots, x_{ip}))$

Poisson regression model:
- $V(Y_i|x_1, \ldots, x_p) = \lambda(x_{i1}, \ldots, x_{ip})$
- Overdispersed Poisson: $V(Y_i|x_1, \ldots, x_p) = \phi\lambda(x_{i1}, \ldots, x_{ip})$

Overdispertion Parameter: ϕ
- Estimate: $\hat{\phi} = \frac{D}{n-p-1}$ where D is the sum of the squared deviances
- If $\hat{\phi} > 2$ then overdispersed model

Georgia Tech

7

Summary

Georgia Tech

8