Applied Statistical Methods II

Logistic Regression, Poisson Regression, and Generalized Linear Models

Part V

Looking forward:

We will look at Poisson regression

- GLM for count data.
- Also used for modeling rates.

Truancy Example

- You want to understand some factors which are associated with high school students missing school and might want to do prediction.
- You collect data about last year's junior classes from two schools (157 students from one school and 159 from the other).
- Your dependent variable is the number of absent school days per student.
- Your explanatory variables are sex and scores on standardized exams for math and language skills.

Count Data vs. Continuous

- Count data takes on discrete values.
- The variance is a function of the mean
  - As the mean increases, variance tends to increase
- The distribution of the values is skewed to the right.
  - What does that say about the frequency of large counts?
- We can often assume that the data follow a Poisson distribution.
  - $P(Y = y) = \frac{\mu^y \exp(-\mu)}{y!}$ for $y = 0, 1, 2, \ldots$.
- Parameterization of the Poisson distribution
  - $EY = \mu$
  - $\text{Var}(Y) = \mu$
  - $Y \sim \text{Poisson}(\mu)$
- Distribution resembles $N(\mu, \mu)$ as $\mu \to \infty$.

Poisson Regression

Count data can be modeled through Poisson Regression.

- Conditional on the covariates $X_{i1}, \ldots, X_{i(p-1)}$.
- Assume $Y_i \sim \text{Poisson}(\mu_i)$.
- Model $\log(\mu_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{ip-1} = X_i^T \beta$.
  - This is different from the log-normal regression.
  - Note that $\mu \in (0, +\infty)$.
  - $\log : (0, +\infty) \to (-\infty, +\infty)$ is one-to-one.
- Again the model $Y_i = EY_i + \epsilon_i$ on page 619 of KKLN is not intuitive.
- Why do you think we most often use the link $\log(\mu_i)$?

Poisson Regression through MLE

- Will fit the model through MLE.
  - Newton-Raphson is used to maximize the likelihood.
  - Newton-Raphson and Fisher's Scoring are the same when using canonical link log.
  - $\beta_l = \beta_{l-1} + I_n^{-1}(\beta_{l-1})U(\beta_{l-1})$.
  - $I_n = X^T WX$, where W is an diagonal matrix with $W_{ii} = \mu_i$.

Three Types of Inference

Just like logistic regression, three approaches:

- Wald
  - These are the most popular.
  - They are easy to compute.
  - Are not reliable when there are small sample sizes.
- Likelihood ratio
  - Most popular in testing if a group of parameters are simultaneously zero.
  - Requires computing two different models then comparing them.
  - More robust to small sample sizes than Wald tests.
- Score test

Wald Tests

- Wald test of $H_0 : \beta_j = C$ uses the test statistic
  - $Z = \frac{\hat{\beta}_j - C}{\hat{se}(b_j)}$
  - $\hat{se}(\hat{\beta}_j) = \sqrt{[I_n^{-1}]_{j+1,j+1}}$
  - Under the null, $Z \sim N(0,1)$
- Wald test of $H_0 : a_0\beta_0 + \cdots + a_{p-1}\beta_{p-1} = C$:
  - Let $a = [a_0, \ldots, a_{p-1}]^T$.
  - $Z = \frac{a^T\hat{\beta} - C}{\hat{se}(a^T\hat{\beta})}$
  - $\hat{se}(a^T\hat{\beta}) = \sqrt{a^T I_n^{-1} a}$
  - Under the null, $Z \sim N(0,1)$
- Can easily invert to get confidence intervals.

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
  - $\log(\mu_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i(p-1)}$
  - Gets the MLE's $\hat{\beta}$, and the log likelihood $\ell(\hat{\beta})$
  - This is the "full model".
- then fit the "reduced" model:
  - $\log(\mu_i) =$
    $\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{j-1} X_{i(j-1)} + \beta_{j+1} X_{i(j+1)} + \cdots + \beta_{p-1} X_{i(p-1)}$
  - Gets the MLE's $\hat{\beta}^{-j}$, and the log likelihood $\ell(\hat{\beta}^{-j})$
- $\Lambda = -2\left[\ell(\hat{\beta}^{-j}) - \ell(\hat{\beta})\right] \sim \chi^2_1$
- Reject the null at level $1-\alpha$ when Λ is greater than the $1-\alpha$ percentile of χ^2_1
- Confidence intervals require an iterative procedure.

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

Fitting and Model Checking

- In R: use `glm` with `family = poisson(link = "log")`.
- In SAS: Poisson regression models can be easily fit in Proc GENMOD.
- Both deviance and standardized Pearson's residuals can be defined similar to those for logistic regression.
  - $r_{p_i} = \frac{Y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$
  - $r_{d_i} = sign(Y_i - \hat{\mu}_i)\sqrt{2\left[Y_i \log(Y_i/\hat{\mu}_i) - (Y_i - \hat{\mu}_i)\right]}$
  - Use the convention that $0 \log(0) = 0$.
- Deviance can be used for comparing models:
  - Is just the likelihood ratio test for nested models.
  - Can be used to test overall lack-of-fit.

Interpretation of Parameters

- We fit a Poisson regression model to

  $$\log\left[E(Y_i|X_{i1},\ldots,X_{i(p-1)})\right] = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i(p-1)}.$$

- β_0 is the log of EY when all covariates are 0.
- $\exp(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i(p-1)})$ is the expected value of Y at covariates X_i.
- β_j is the increase in the log of EY, conditional on other covariates, with an increase in X_{ij} by one unit.
- $\exp(\beta_j)$ is the ratio of the EY, conditional on other covariates, between $X_{ij} + 1$ and X_{ij}.

Fitting the Poisson Regression

- Let's analyze the truancy example:
  - The rate ratio of days missed between males and females controlling for test scores.
  - The expected number days missed by an average male (test scores = 50).
  - The expected number days missed by an average female (test scores = 50).

Inference

- The rate ratio of days missed between males and females controlling for test scores is estimated as   with an estimated 95% Wald CI of [,   ].
- The expected number of days missed by a male with math and verbal scores of 50 is estimated as   with an estimated 95% Wald CI of [,   ].
- The expected number of days missed by a female with math and verbal scores of 50 is estimated as   with an estimated 95% Wald CI of [,   ].

Poisson Regression for Rates

- In the truancy example, each subject was observed for the same number of days or time (the entire school year).
- What if each subject is observed over a different time period?
- We would want to estimate a rate.
  - The expected number of events given a time interval.
  - Assume that the number of events that will occur in an interval of time of length T is $\mu * T$.

Poisson Regression for Rates, Cont.

- Let Y_i be the number of events observed from subject i over time T_i.
- We will model:

$$
\begin{aligned}
\log(E(Y_i/T_i)) &= \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{(p-1)} X_{i(p-1)} \\
\log(E(Y_i)) &= \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{i(p-1)} + \log(T_i)
\end{aligned}
$$

- $\log(T_i)$ is called the offset.
- $\hat{\mu}_i$ is the estimated rate per base unit of time for the ith subject.
- $\hat{\mu}_i T_i$ is the fitted count for the ith subject.

The Estimates

- $\hat{\mu}_i = \exp(\hat{\beta}_0 + \cdots + \hat{\beta}_{p-1} X_{i(p-1)})$ is the estimate for the expected number of events per unit time.
- How would you interpret $\exp(\hat{\beta}_j)$?
- We choose the units of the time interval.
- In the truancy example, it was school days.
- Should report a rate that is in appropriate units:
  - Do you want to know that the murder rate in Philadelphia was .00016 per resident in 2013?
  - Or that there were 16 homicides per 100,000 residents?
- You can make this conversion either through the choice of units of the offset or by transforming estimates.

The Estimates

- $\hat{\mu}_i = \exp(\hat{\beta}_0 + \cdots + \hat{\beta}_{p-1} X_{i(p-1)})$ is the estimate for the expected number of events per unit time.
- How would you interpret $\exp(\hat{\beta}_j)$?
- We choose the units of the time interval.
- In the truancy example, it was school days.
- Should report a rate that is in appropriate units:
  - Do you want to know that the murder rate in Philadelphia was .00016 per resident in 2013?
  - Or that there were 16 homicides per 100,000 residents?
- You can make this conversion either through the choice of units of the offset or by transforming estimates.

Horse Example

- Consider a study of the infection rate among hospitalized horses.
- The outcome is the number of hospital acquired infections acquired by a horse.
- We know the time that each horse spent in the hospital in days.
- We also know the age of the horse at the time of admission in months.
- Some questions to ask:
  - What is the rate of infection for a 2 year old horse?
  - How does age affect the rate of infection?

Interpretation

- The estimated infection rate for 2 years old horses is infections per 6 months of hospitalization with a 95% confidence interval of [,].
- An increase in age by 1 year increases the expected rate of infection by % with an 95% confidence interval of [%, %].
  - Why don't we have to worry about the units of our rate when discussing the affect of a covariate?

Binomial and Poisson

- Consider $Y \sim \text{Bin}(n, \pi)$. If n is large compared to π, then the $Y \approx \text{Poi}(n\pi)$.
  - Proof: for $\lambda = n\pi$, moment generating function $M(t)$ for $\text{Bin}(n, \pi)$ is $\{1 + \lambda/n(e^t - 1)\}^n$. Take $n \to \infty$.
  - Poisson data is often easier to model than binomial data.
- Poisson regression to approximate a binomial is used a lot when you do not have subject specific data but group level data.
- Note: sum of independent Poisson is still Possion. But sum of Binomial with different π is not Binomial.

Example

- Determine how having a single parent household is associated with rate of crime conviction in 2005.
  - This problem can be thought of as a Bernoulli problem at individual level.
  - You get your data from a national data base on the county level.
  - Response: Y_i = number of people convicted of a crime (also know the total population), percentage single parent households are included in the data set.
  - Don't know subject level information.
  - Is Y_i Binomial?
  - Poisson regression with offset can estimate the crime rate.

Other examples of Poisson for rates

- Example: In social sciences often data are not in subject level, but for a group of people. The denominator in rate is number of people.
- Example: In health sciences often the observation for a subject is during some amount of time. The denominator in rate is time.
  - Example: study to look at number of days with suicidal thoughts in bipolar adolescents.
  - Number of days each patient is observed is different.

Looking forward:

Continue Poisson regression

- Over-dispersion
- Some model building and examples.

Over-dispersion: an example

- Remember that the mean and variance of a Poisson regression are the same.
- $Y_i \mid x_i \sim Poi(\mu_i)$ are independent, where $\log(\mu_i) = \beta_0 + \beta x_i$, but we do not observe x_i.
- Suppose $x_i \sim F$ are i.i.d with cumulant generating function $\log\left\{E\left(e^{tx_i}\right)\right\} = K(t)$. Are $Y_1, \ldots, Y_n$ independent unconditional on x_i?

$$E(Y_i) = E\{E(Y_i \mid x_i)\} = \exp\{\beta_0 + K(\beta)\}$$
$$\begin{aligned} \mathrm{Var}(Y_i) &= E\{E\mathrm{Var}(Y_i \mid x_i)\} + \mathrm{Var}\{E(Y_i \mid x_i)\} \\ &= \exp\{\beta_0 + K(\beta)\} + \exp(2\beta_0)\mathrm{Var}\{\exp(\beta x_i)\} \end{aligned}$$

- $\mathrm{Var}(Y_i) > E(Y_i)$ if x_i is not observed!
- In general, over-dispersion: the variance is larger than predicted by the model, i.e. $Var(Y_i) > v(\mu_i)$.

Over-dispersion: an example

- Remember that the mean and variance of a Poisson regression are the same.
- $Y_i \mid x_i \sim Poi(\mu_i)$ are independent, where $\log(\mu_i) = \beta_0 + \beta x_i$, but we do not observe x_i.
- Suppose $x_i \sim F$ are i.i.d with cumulant generating function $\log\left\{E\left(e^{tx_i}\right)\right\} = K(t)$. Are $Y_1, \ldots, Y_n$ independent unconditional on x_i?

$$
\begin{aligned}
E(Y_i) &= E\{E(Y_i \mid x_i)\} = \exp\{\beta_0 + K(\beta)\} \\
\mathrm{Var}(Y_i) &= E\{E\mathrm{Var}(Y_i \mid x_i)\} + \mathrm{Var}\{E(Y_i \mid x_i)\} \\
&= \exp\{\beta_0 + K(\beta)\} + \exp(2\beta_0)\mathrm{Var}\{\exp(\beta x_i)\}
\end{aligned}
$$

- $\mathrm{Var}(Y_i) > E(Y_i)$ if x_i is not observed!
- In general, over-dispersion: the variance is larger than predicted by the model, i.e. $Var(Y_i) > v(\mu_i)$.

Over-dispersion: an example

- Remember that the mean and variance of a Poisson regression are the same.
- $Y_i \mid x_i \sim Poi(\mu_i)$ are independent, where $\log(\mu_i) = \beta_0 + \beta x_i$, but we do not observe x_i.
- Suppose $x_i \sim F$ are i.i.d with cumulant generating function $\log\left\{E\left(e^{tx_i}\right)\right\} = K(t)$. Are $Y_1, \ldots, Y_n$ independent unconditional on x_i?

$$\begin{aligned} E(Y_i) &= E\{E(Y_i \mid x_i)\} = \exp\{\beta_0 + K(\beta)\} \\ \mathrm{Var}(Y_i) &= E\{E\mathrm{Var}(Y_i \mid x_i)\} + \mathrm{Var}\{E(Y_i \mid x_i)\} \\ &= \exp\{\beta_0 + K(\beta)\} + \exp(2\beta_0)\mathrm{Var}\{\exp(\beta x_i)\} \end{aligned}$$

- $\mathrm{Var}(Y_i) > E(Y_i)$ if x_i is not observed!
- In general, over-dispersion: the variance is larger than predicted by the model, i.e. $Var(Y_i) > v(\mu_i)$.

Assessing Over-Dispersion

- Over-Dispersion is a problem with lack-of-fit.
  - Assessment can be done similarly to assessment of lack-of-fit.
- Pearson's χ^2. Assume that Y_i are counts at time T_i with covariates $\mathbf{X}_i$.
  - $\mu_i = E(Y_i/T_i)$ and $\log \mu_i = \mathbf{X}_i^T \beta$
  - $X^2 = \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i T_i)^2}{v(\hat{\mu}_i T_i)} = \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i T_i)^2}{\hat{\mu}_i T_i}$.
- Ad-hoc: $\hat{\phi} = X^2/(n-p)$ should be about 1.
- The scaled (by what?) Deviance should also be about 1 but can have poor asymptotic properties.
- Deviance and Pearson's residuals can be plotted as well.
- Can also use Carmeron and Trivedi (1990) to test for over-dispersion in the Poisson model (see course documents).

What Next?

- You have over-dispersion. What do you do?
  1. Check to make sure that you have the proper regression function.
     - Check if you need more terms.
  2. If you can't fix the regression function, adjust the Poisson distribution. Two popular ways:
     - Use a quasi-likelihood by setting $\text{Var}(Y_i) = \phi\mu_i$. This is easy, and appropriate when we are missing covariates.
     - Use the 2 parameter Negative-Binomial distribution rather than the Poisson

What Next?

- You have over-dispersion. What do you do?
  1. Check to make sure that you have the proper regression function.
     - Check if you need more terms.
  2. If you can't fix the regression function, adjust the Poisson distribution. Two popular ways:
  - Use a quasi-likelihood by setting $\text{Var}(Y_i) = \phi\mu_i$. This is easy, and appropriate when we are missing covariates.
  - Use the 2 parameter Negative-Binomial distribution rather than the Poisson

Using the Over-Dispersion Parameter

- $\sqrt{\phi}$ is known as the scale parameter.
- If you force $\text{Var}(Y_i) = \phi\mu_i \neq \mu_i$, you are not using a Poisson distribution. Sometimes called over-dispersed Possion.
- Use a quasi-likelihood:
  - A quasi-likelihood is based just on the first two moments.
  - Obtain point estimates as usual.
  - When doing inference, inflate the inverse Hessian by $\hat{\phi}$.

One way to motivate the Negative Binomial

- Suppose $\mu_i = x_i^T \beta$ for x_i observed.

$$
\begin{aligned}
Y_i \mid x_i, \gamma_i &\sim Poi\left(\gamma_i \mu_i\right) \\
\gamma_i \mid \delta &\sim \text{Gamma}\left(\delta^{-1}, \delta^{-1}\right) \\
E\left(Y_i \mid x_i\right) &= \mu_i \\
\text{Var}\left(Y_i \mid x_i\right) &= E\left\{\text{Var}\left(Y_i \mid x_i, \gamma_i\right) \mid x_i\right\} + \text{Var}\left[\left\{E\left(Y_i \mid x_i, \gamma_i\right)\right\} \mid x_i\right] \\
&= \mu_i + \delta \mu_i^2
\end{aligned}
$$

δ is precision parameter. Then $Y_i \sim NB\left(\frac{\mu_i}{\delta^{-1}+\mu_i}, \delta^{-1}\right)$

$$
\begin{aligned}
E(Y_i \mid x_i) &= \mu_i \\
\mathrm{Var}(Y_i \mid x_i) &= E\{\mathrm{Var}(Y_i \mid x_i, \gamma_i) \mid x_i\} + \mathrm{Var}[\{E(Y_i \mid x_i, \gamma_i)\} \mid x_i] \\
&= \mu_i + \delta\mu_i^2
\end{aligned}
$$

- This arises naturally in gene expression and other data types.
  - Suppose each individual i has N_i total genes in a cell (known), and we are interested in gene A.
  - Suppose gene A comprises $f_i = \gamma_i \mu_i / N_i$ of all genes in individual i's cell.
  - If we know γ_i, we sample genes from individual i according to $Poi(f_i N_i)$. The variability in sampling (i.e. $E\{\mathrm{Var}(Y_i \mid x_i, \gamma_i) \mid x_i\}$) is the technical variability.
  - Since individuals are heterogeneous, model $\gamma_i \sim \mathrm{Gamma}(1/\delta, 1/\delta)$. The variability in the mean (i.e. $\mathrm{Var}[\{E(Y_i \mid x_i, \gamma_i)\} \mid x_i]$) is the biological variability.

Examples

- We will consider some examples.
- The Horse infection data. Example of a good fit.
- The crab data: how are weight and color associated with number of satellites.
  - Example of an over-dispersed model.
  - First look at its mean.
  - Deal with a scaled over-dispersion parameter.
  - Try a negative binomial model.
    - Will have to deal with model building and outlier detection.

Horse Example

- Consider a study of the infection rate among hospitalized horses.
- The outcome is the number of hospital acquired infections acquired by a horse.
- We know the time that each horse spent in the hospital in days.
- We also know the age of the horse at the time of admission in months.

Crab Data Set Revisited

- In the past we used the outcome if there are any satellites.
- Let's use the outcome number of satellites.
- We want to know how weight and color are associated with the number of satellites.
- Assume these are the only covariates we know.

- Tie linear regression, logistic regression, and Poisson regression into examples of generalized linear models (GLM).

Overview of GLM

- Why are we grouping the analysis of normal, binomial, and Poisson data into one category?
- What else is included in this group?
- We will now take a general overview of GLM.

The Exponential Family of Distributions

- Assume we observe independent univariate outcomes Y_i.
  - Can be extended to multivariate outcomes but we will not consider that here.
- Assume the pdf or pf and log-likelihood of Y_i are:

$$
\begin{aligned}
f(Y_i|\theta_i,\phi) &= \exp\left\{\frac{Y_i\theta_i - b(\theta_i)}{\phi} + c(Y_i,\phi)\right\} \\
\ell(\theta_i,\phi|Y_i) &= \frac{Y_i\theta_i - b(\theta_i)}{\phi} + c(Y_i,\phi)
\end{aligned}
$$

  - b is twice differentiable.
- We call f a member of the family of exponential distributions.

Normal Distribution

- $Y_i \sim N(\mu_i, \sigma^2)$

$$
\begin{aligned}
\ell(\mu, \sigma^2 | Y_i) &= \frac{-1}{2}\frac{(Y_i - \mu_i)^2}{\sigma^2} + \frac{1}{2}\log\left(2\pi\sigma^2\right) \\
&= \frac{Y_i\mu_i - \mu_i^2/2}{\sigma^2} - \frac{1}{2}\left[\frac{Y_i^2}{\sigma^2} + \log\left(2\pi\sigma^2\right)\right]
\end{aligned}
$$

$\theta_i = \mu_i$ $\qquad \phi = \sigma^2$

$b(\theta_i) = \mu_i^2/2$ $\qquad$ c is the last term

Bernoulli Distribution

- Assume that $Y_i \sim Ber(\pi_i)$.

$$\begin{aligned}\ell(\pi_i | Y_i) &= Y_i \log(\pi_i) + (1 - Y_i)\log(1 - \pi_i) \\ &= Y_i \log\left(\frac{\pi_i}{1 - \pi_i}\right) + \log(1 - \pi_i)\end{aligned}$$

$\theta_i = logit(\pi_i)$ $\qquad \phi = 1$
$b(\theta) = -\log(1 + e^{\theta_i})$

Poisson Distribution

- Assume that $Y_i \sim$ Poisson (μ_i).

$$\ell(\mu_i | Y_i) = Y_i \log(\mu_i) - \mu_i - \log(Y_i!)$$

$$\begin{array}{ll} \theta_i = \log(\mu_i) & \phi = 1 \\ b(\theta_i) = \exp(\theta_i) & c = -\log(Y_i!) \end{array}$$

Nice Properties

- There are some nice properties of the exponential family.
- Recall that $E\frac{\partial \ell}{\partial \theta_i} = 0$ at the true parameter θ_i.
- Some calculations:

$$
\begin{aligned}
\frac{\partial \ell}{\partial \theta_i} &= \frac{Y_i - b'(\theta_i)}{\phi} \\
E\frac{\partial \ell}{\partial \theta_i} &= \frac{\mu_i - b'(\theta_i)}{\phi} \\
E(Y_i) = \mu_i &= b'(\theta_i)
\end{aligned}
$$

The Variance

- Fisher's Information Equality Tells us that:

$$E\frac{\partial^2 \ell}{\partial \theta_i^2} + E\left[\frac{\partial \ell}{\partial \theta_i}\right]^2 = 0$$

$$\begin{aligned}
\frac{\partial \ell}{\partial \theta_i} &= \frac{Y_i - b'(\theta_i)}{\phi} \\
\frac{\partial^2 \ell}{\partial \theta_i^2} &= \frac{-b''(\theta_i)}{\phi} \\
0 &= \frac{-b''(\theta_i)}{\phi} + \frac{E\left(Y_i - EY_i\right)^2}{\phi^2}
\end{aligned}$$

- $\mathrm{Var}(Y_i) = b''(\theta_i)\phi$.

Two Parts to GLM

- The GLM has two parts.
- Random or Stochastic: choose the distribution for Y_i.
- Systematic: Model EY_i as a function of covariates.
  - $g(EY_i) = \mathbf{X}_i^T \beta$
  - g is the link function.
- $g = b'^{-1}$ is called the canonical link, i.e., $\theta = b'^{-1}(\mu) = X\beta$.
- Canonical link has nice properties in terms of computation and asymptotics.
  - The observed and expected information matrices (-Hessians) are the same.
  - $I_n = X^T WX$, where W is diagonal with W_{ii} being $[\{b''(\mu_i)\}^{-1}\phi]^{-1}$
- Other than these, choose a link function for interpretability and the fit of the data.

Fitting The Models

- Newton-Ralphson and Fisher's Scoring can be use to find the MLE's.
- Inference will be based on large sample properties.
    - MLE are asymptotically normal.
    - Their variance/covariance matrix is asymptotically the inverse Fisher Information I_n^{-1}.

Deviance

- The Deviance is a type of log-likelihood statistic.
- Let $\ell(\mu|\mathbf{Y}) = \sum \ell(\mu_i|Y_i)$ be the joint likelihood.
- Fit a model to obtain **B** and corresponding $\hat{\mu}_i$.
- The Deviance for a particular model is $D = -2\left[\ell(\hat{\mu}|\mathbf{Y}) - \ell(\mathbf{Y}_i|\mathbf{Y}_i)\right]$.
- We are more interested in the difference of DEVs between two models: likelihood ratio test.

Pearson's χ^2

- $X^2 = \sum \frac{(Y_i - \hat{\mu}_i)^2}{V(\mu_i)}$
- When used for Poisson regression, known as Cochran's lack-of-fit. Under the null that the model fits well, $X^2 \sim \chi^2_{n-p}$.
- Note that for binary data (logistic regression), Pearson's χ^2 test can only be used for grouped data (Binomial with n_i sufficiently large).
- Which one do I use?
  - If they don't give you close to the same answer, you have bigger problems.
  - Deviance provides a nice nested set of tests for choosing groups of parameters.
  - X^2 often has nicer asymptotic and is easier to interpret.

Residuals

- We consider two types of residuals: Pearson's and Deviance.
- Pearson's are the individual contributions of each observation to X^2.
  - we often look at their standardized version.
- Deviance residuals are a function of the contribution of each observation to D_i.
- Test vs. Residuals for diagnostics:
  - Tests give you a definitive value while residual plots are rather subjective.
  - Residual plots give more insight into what might be going wrong.

Dispersion

- If you have poor fit check to make sure you have the correct mean modeled.
  - Add and subtract terms.
  - Possibly need a different link function.
- For one-parameter families, could have over-dispersion.
  - Extra variation breaks the relationship between the mean and variance.
  - Over-dispersed binomial: $Var(Y) > n\mu(1-\mu)$.
  - If Y is binary and we model $Y \sim \text{Ber}(\pi)$, do we have to worry about over-dispersion?

Dispersion

- Can either fit an appropriate two-parameter model or an over-dispersed model.
- Change ϕ to adjust the variance.
  - Recall $\text{Var}(Y_i) = b''(\theta_i)\phi$ and $EY_i = b'(\theta_i)$
  - Called the scale parameter.
- Do not have a full likelihood but a quasi-likelihood.
  - SAS fixes the parameter ϕ to make $X^2/(n-p) = \phi$ then finds β.
  - In R, can specify dispersion=$X^2/(n-p)$.
  - Should only change the variance.

Overview

- GLM is a very general framework to answering many questions.
- We only looked at a small section of its scope.
  - Small but most popular.
  - Ideas are easy to generalize to other situations.

GLM beyond canonical link

For $\theta_i = x_i^T \beta$, consider the log-likelihood

$$\ell(\theta_i; y_i) = \frac{y_i g(\theta_i) - K\{g(\theta_i)\}}{\phi} + c(y_i; \phi) \qquad (1)$$

- Things we know:
  - $Ey_i = \mu_i(\theta_i) = K'\{g(\theta_i)\}$
  - $\mathrm{Var}(y_i) = \phi K''\{g(\theta_i)\}$
  - How can we interpret g in terms of a link function h?
- Consider the model $h(\mu_i) = x_i^T \beta = \theta_i$.
- Then $\mu_i(\theta_i) = h^{-1}(\theta_i)$.
- We also have $\mu_i(\theta_i) = K'\{g(\theta_i)\}$.
- $\Rightarrow g(\theta_i) = K'^{-1}\{h^{-1}(\theta_i)\}$.
- Given a parametric family of distributions (i.e. Poisson, binomial, etc.), for every link function h, there is a function g such that the likelihood can be written as (1).

GLM beyond canonical link

For $\theta_i = x_i^T \beta$, consider the log-likelihood

$$\ell(\theta_i; y_i) = \frac{y_i g(\theta_i) - K\{g(\theta_i)\}}{\phi} + c(y_i; \phi) \qquad (1)$$

- Things we know:
  - $Ey_i = \mu_i(\theta_i) = K'\{g(\theta_i)\}$
  - $\text{Var}(y_i) = \phi K''\{g(\theta_i)\}$
  - How can we interpret g in terms of a link function h?
- Consider the model $h(\mu_i) = x_i^T \beta = \theta_i$.
- Then $\mu_i(\theta_i) = h^{-1}(\theta_i)$.
- We also have $\mu_i(\theta_i) = K'\{g(\theta_i)\}$.
- $\Rightarrow g(\theta_i) = K'^{-1}\{h^{-1}(\theta_i)\}$.
- Given a parametric family of distributions (i.e. Poisson, binomial, etc.), for every link function h, there is a function g such that the likelihood can be written as (1).

GLM beyond canonical link

For $\theta_i = x_i^T \beta$, consider the log-likelihood

$$\ell(\theta_i; y_i) = \frac{y_i g(\theta_i) - K\{g(\theta_i)\}}{\phi} + c(y_i; \phi) \tag{1}$$

- Things we know:
  - $Ey_i = \mu_i(\theta_i) = K'\{g(\theta_i)\}$
  - $\text{Var}(y_i) = \phi K''\{g(\theta_i)\}$
  - How can we interpret g in terms of a link function h?
- Consider the model $h(\mu_i) = x_i^T \beta = \theta_i$.
- Then $\mu_i(\theta_i) = h^{-1}(\theta_i)$.
- We also have $\mu_i(\theta_i) = K'\{g(\theta_i)\}$.
- $\Rightarrow g(\theta_i) = K'^{-1}\left\{h^{-1}(\theta_i)\right\}$.
- Given a parametric family of distributions (i.e. Poisson, binomial, etc.), for every link function h, there is a function g such that the likelihood can be written as (1).

Estimation with arbitrary link

$$\theta_i = x_i^T \beta$$

$$\ell\left(\theta_i; y_i\right) = \frac{y_i g(\theta_i) - K\left\{g(\theta_i)\right\}}{\phi} + c\left(y_i; \phi\right)$$

- Score function:

$$\nabla_\beta \ell\left(\beta; y_i\right) = \phi^{-1} x_i g'\left(\theta_i\right)\left[y_i - \underbrace{K'\left\{g\left(\theta_i\right)\right\}}_{=\mu_i}\right]$$

- Fisher information:

$$-E\left\{\nabla_\beta^2 \ell\left(\beta; y_i\right)\right\} = \phi^{-1} g'(\theta_i)\underbrace{K''\left\{g\left(\theta_i\right)\right\}}_{v(\mu_i)} x_i x_i^T$$

Quasi-likelihood and non-linear models

Setup: Let $Y \in \mathbb{R}^n$ be such that $EY = \mu(\beta)$ and $\text{Var}(Y) = \phi V\{\mu(\beta)\}$, where $\mu : \mathbb{R}^p \to \mathbb{R}^n$ and $V : \mathbb{R}^n \to \mathbb{R}^{n \times n}$.

Examples when data $Y_1, \ldots, Y_n$ are independent:

- Binomial with logit link
  - $Y_i =$ #of success in sample i out of m_i trials.
  - $logit\{\mu_i(\beta)/m_i\} = x_i^T\beta$
  - $V(\mu) = \text{diag}\left\{\frac{\mu_1(m_1-\mu_1)}{m_1}, \ldots, \frac{\mu_p(m_p-\mu_p)}{m_p}\right\}$, $\phi = 1$.
- Normal with identity link and unknown variance
  - $Y_i \in \mathbb{R}$
  - $\mu_i(\beta) = x_i^T\beta$
  - $V(\mu) = I_n$, $\phi > 0$.
- Overdispersed Poisson with log link
  - $Y_i =$ #of counts ≥ 0.
  - $log\{\mu_i(\beta)\} = x_i^T\beta$
  - $V(\mu) = \text{diag}(\mu_1, \ldots, \mu_p)$, $\phi > 0$.
- Can we do inference on β when we ONLY specify mean and variance of Y (and NOT the distribution)?

Quasi-likelihood and non-linear models

Setup: Let $Y \in \mathbb{R}^n$ be such that $EY = \mu(\beta)$ and $\text{Var}(Y) = \phi V\{\mu(\beta)\}$, where $\mu : \mathbb{R}^p \to \mathbb{R}^n$ and $V : \mathbb{R}^n \to \mathbb{R}^{n\times n}$.
Examples when data $Y_1, \ldots, Y_n$ are independent:

- Binomial with logit link
  - Y_i = #of success in sample i out of m_i trials.
  - $logit\{\mu_i(\beta)/m_i\} = x_i^T\beta$
  - $V(\mu) = \text{diag}\left\{\frac{\mu_1(m_1-\mu_1)}{m_1}, \ldots, \frac{\mu_p(m_p-\mu_p)}{m_p}\right\}$, $\phi = 1$.
- Normal with identity link and unknown variance
  - $Y_i \in \mathbb{R}$
  - $\mu_i(\beta) = x_i^T\beta$
  - $V(\mu) = I_n$, $\phi > 0$.
- Overdispersed Poisson with log link
  - Y_i = #of counts ≥ 0.
  - $log\{\mu_i(\beta)\} = x_i^T\beta$
  - $V(\mu) = \text{diag}(\mu_1, \ldots, \mu_p)$, $\phi > 0$.
- Can we do inference on β when we ONLY specify mean and variance of Y (and NOT the distribution)?

Estimation

$EY = \mu(\beta)$, $\mathrm{Var}(Y) = \phi V\{\mu(\beta)\}$

- Construct an estimating equation that is analogous to the score function: $E_{\text{true }\beta}(\text{score}) = 0$.
- Consider the function $f(\beta) = H^T\{Y - \mu(\beta)\}$, where $H \in \mathbb{R}^{n \times p}$ can be a function of β but not Y.
    - $E\{f(\beta)\} = 0$.
    - $\mathrm{Var}\{f(\beta)\} = \phi H^T V\{\mu(\beta)\} H$
- Estimator: $\hat{\beta}$ satisfies $f\left(\hat{\beta}\right) = 0$. How do we choose H?
- "Best" H is the one with the smallest asymptotic variance.
- Let $D(\beta) = \nabla_\beta \mu(\beta) \in \mathbb{R}^{n \times p}$.

$$
\begin{aligned}
0 = & n^{-1} f\left(\hat{\beta}\right) \approx n^{-1} f(\beta) \\
& + \left[-n^{-1} H^T D(\beta) + \begin{pmatrix} n^{-1}\{Y - \mu(\beta)\}^T (\nabla_\beta H_1) \\ \vdots \\ n^{-1}\{Y - \mu(\beta)\}^T (\nabla_\beta H_p) \end{pmatrix}\right]\left(\hat{\beta} - \beta\right)
\end{aligned}
$$

Estimation

$EY = \mu(\beta)$, $\mathrm{Var}(Y) = \phi V\{\mu(\beta)\}$

- Construct an estimating equation that is analogous to the score function: $E_{\text{true } \beta}(\text{score}) = 0$.
- Consider the function $f(\beta) = H^T\{Y - \mu(\beta)\}$, where $H \in \mathbb{R}^{n \times p}$ can be a function of β but not Y.
  - $E\{f(\beta)\} = 0$.
  - $\mathrm{Var}\{f(\beta)\} = \phi H^T V\{\mu(\beta)\} H$
- Estimator: $\hat{\beta}$ satisfies $f\left(\hat{\beta}\right) = 0$. How do we choose H?
- "Best" H is the one with the smallest asymptotic variance.
- Let $D(\beta) = \nabla_\beta \mu(\beta) \in \mathbb{R}^{n \times p}$.

$$
\begin{aligned}
0 =& n^{-1} f\left(\hat{\beta}\right) \approx n^{-1} f(\beta) \\
&+ \left[-n^{-1} H^T D(\beta) + \begin{pmatrix} n^{-1}\{Y - \mu(\beta)\}^T (\nabla_\beta H_1) \\ \vdots \\ n^{-1}\{Y - \mu(\beta)\}^T (\nabla_\beta H_p) \end{pmatrix}\right] \left(\hat{\beta} - \beta\right)
\end{aligned}
$$

Estimation

$EY = \mu(\beta)$, $\mathrm{Var}(Y) = \phi V\{\mu(\beta)\}$

- Construct an estimating equation that is analogous to the score function: $E_{\text{true }\beta}(\text{score}) = 0$.
- Consider the function $f(\beta) = H^T\{Y - \mu(\beta)\}$, where $H \in \mathbb{R}^{n\times p}$ can be a function of β but not Y.
  - $E\{f(\beta)\} = 0$.
  - $\mathrm{Var}\{f(\beta)\} = \phi H^T V\{\mu(\beta)\} H$
- Estimator: $\hat{\beta}$ satisfies $f\left(\hat{\beta}\right) = 0$. How do we choose H?
- "Best" H is the one with the smallest asymptotic variance.
- Let $D(\beta) = \nabla_\beta \mu(\beta) \in \mathbb{R}^{n\times p}$.

$$
\begin{aligned}
0 =& n^{-1} f\left(\hat{\beta}\right) \approx n^{-1} f(\beta) \\
&+ \left[-n^{-1}H^T D(\beta) + \begin{pmatrix} n^{-1}\{Y-\mu(\beta)\}^T(\nabla_\beta H_1) \\ \vdots \\ n^{-1}\{Y-\mu(\beta)\}^T(\nabla_\beta H_p) \end{pmatrix}\right]\left(\hat{\beta}-\beta\right)
\end{aligned}
$$

Estimation

$EY = \mu(\beta)$, $\mathrm{Var}(Y) = \phi V\{\mu(\beta)\}$

- Construct an estimating equation that is analogous to the score function: $E_{\text{true }\beta}(\text{score}) = 0$.
- Consider the function $f(\beta) = H^T\{Y - \mu(\beta)\}$, where $H \in \mathbb{R}^{n\times p}$ can be a function of β but not Y.
  - $E\{f(\beta)\} = 0$.
  - $\mathrm{Var}\{f(\beta)\} = \phi H^T V\{\mu(\beta)\} H$
- Estimator: $\hat{\beta}$ satisfies $f\left(\hat{\beta}\right) = 0$. How do we choose H?
- "Best" H is the one with the smallest asymptotic variance.
- Let $D(\beta) = \nabla_\beta \mu(\beta) \in \mathbb{R}^{n\times p}$.

$$
\begin{aligned}
0 =& n^{-1} f\left(\hat{\beta}\right) \approx n^{-1} f(\beta) \\
&+ \left[-n^{-1} H^T D(\beta) + \begin{pmatrix} n^{-1}\{Y - \mu(\beta)\}^T (\nabla_\beta H_1) \\ \vdots \\ n^{-1}\{Y - \mu(\beta)\}^T (\nabla_\beta H_p) \end{pmatrix} \right] \left(\hat{\beta} - \beta\right)
\end{aligned}
$$

Estimation (cont.)

$D(\beta) = \nabla_\beta \mu(\beta) \in \mathbb{R}^{n \times p}$

- $n^{-1}\{Y - \mu(\beta)\}^T (\nabla_\beta H_j)$ is an average of n independent, mean 0 r.v.s $\Rightarrow \approx 0$, and can be ignored.
-
$$0 \approx n^{-1} f(\beta) - n^{-1} H^T D \left(\hat{\beta} - \beta\right)$$
- Therefore, $\text{Var}\left(\hat{\beta}\right) \underbrace{=}_{\text{asy.}} \phi \left(H^T D\right)^{-1} H^T V(\mu) H \left(D^T H\right)^{-1}$
- Some fancy linear algebra $\Rightarrow \left(H^T D\right)^{-1} H^T V(\mu) H \left(D^T H\right)^{-1} \succeq \left\{D^T V^{-1} D\right\}^{-1}$
- Therefore, "best" $H = V\{\mu(\beta)\}^{-1} D(\beta)$

Estimation and inference

$$EY = \mu(\beta), \mathrm{Var}(Y) = \phi V\{\mu(\beta)\}, D(\beta) = \nabla_\beta \mu(\beta) \in \mathbb{R}^{n\times p}.$$

- $f(\beta) = D(\beta)^T V\{\mu(\beta)\}^{-1}\{Y - \mu(\beta)\}$. $\hat\beta$ satisfies $f(\hat\beta) = 0$.
  - Estimation: update with Fisher scoring:

$$\hat\beta_1 = \left[D(\hat\beta_0)^T V\left\{\mu\left(\hat\beta_0\right)\right\}^{-1} D(\hat\beta_0)\right]^{-1} D(\hat\beta_0)^T V\left\{\mu\left(\hat\beta_0\right)\right\}^{-1} \times$$
$$\times \left\{Y - \mu\left(\hat\beta_0\right)\right\} + \hat\beta_0$$

  - Asymptotic variance:

$$\mathrm{Var}\left(\hat\beta\right) \underbrace{=}_{asy.} \phi\left[D(\beta)^T V\{\mu(\beta)\}^{-1} D(\beta)\right]^{-1}$$

  - How to estimate ϕ?

Connection to quasi-likelihood

- Assume entries Y are independent, i.e. $V(\mu) = \operatorname{diag}\{V_1(\mu_1), \ldots, V_n(\mu_n)\}$

$$f(\beta) = D(\beta)^T V\{\mu(\beta)\}^{-1}\{Y - \mu(\beta)\}$$
$$= \sum_{i=1}^{n} \{\nabla_\beta \mu_i(\beta)\} \frac{Y - \mu_i(\beta)}{V_i\{\mu_i(\beta)\}}$$

- Note $\operatorname{Var}\{f(\beta)\} = -\phi E\{\nabla_\beta f(\beta)\}$, which looks like equality $\operatorname{Var}\{\nabla_\beta \ell(\beta)\} = -E\left\{\nabla_\beta^2 \ell(\beta)\right\}$
- What is $f(\beta)$ in logistic regression? Poisson regression (with log-link)? For arbitrary GLM?
- is EXACTLY the score function (up to the multiplicative constant ϕ) for the log-likelihood

$$\ell_i(\beta; Y_i) = \int_{Y_i}^{\mu_i(\beta)} \frac{Y_i - t}{\phi V_i(t)} dt$$

Connection to quasi-likelihood

- Assume entries Y are independent, i.e. $V(\mu) = \text{diag}\{V_1(\mu_1), \ldots, V_n(\mu_n)\}$

$$
\begin{aligned}
f(\beta) &= D(\beta)^T \, V\{\mu(\beta)\}^{-1}\{Y - \mu(\beta)\} \\
&= \sum_{i=1}^{n} \{\nabla_\beta \mu_i(\beta)\} \frac{Y - \mu_i(\beta)}{V_i\{\mu_i(\beta)\}}
\end{aligned}
$$

- Note $\text{Var}\{f(\beta)\} = -\phi E\{\nabla_\beta f(\beta)\}$, which looks like equality $\text{Var}\{\nabla_\beta \ell(\beta)\} = -E\left\{\nabla_\beta^2 \ell(\beta)\right\}$
- What is $f(\beta)$ in logistic regression? Poisson regression (with log-link)? For arbitrary GLM?
- is EXACTLY the score function (up to the multiplicative constant ϕ) for the log-likelihood

$$
\ell_i(\beta; Y_i) = \int_{Y_i}^{\mu_i(\beta)} \frac{Y_i - t}{\phi V_i(t)} dt
$$

Connection to quasi-likelihood

- Assume entries Y are independent, i.e. $V(\mu) = \text{diag}\left\{V_1(\mu_1), \ldots, V_n(\mu_n)\right\}$

$$f(\beta) = D(\beta)^T \, V\left\{\mu(\beta)\right\}^{-1} \left\{Y - \mu(\beta)\right\}$$
$$= \sum_{i=1}^{n} \left\{\nabla_\beta \mu_i(\beta)\right\} \frac{Y - \mu_i(\beta)}{V_i\left\{\mu_i(\beta)\right\}}$$

- Note $\text{Var}\left\{f(\beta)\right\} = -\phi E\left\{\nabla_\beta f(\beta)\right\}$, which looks like equality $\text{Var}\left\{\nabla_\beta \ell(\beta)\right\} = -E\left\{\nabla^2_\beta \ell(\beta)\right\}$
- What is $f(\beta)$ in logistic regression? Poisson regression (with log-link)? For arbitrary GLM?
- is EXACTLY the score function (up to the multiplicative constant ϕ) for the log-likelihood

$$\ell_i(\beta; Y_i) = \int_{Y_i}^{\mu_i(\beta)} \frac{Y_i - t}{\phi V_i(t)} dt$$