More on GLM

Prof. Jacob M. Montgomery

Quantitative Political Methodology II (L32 382)

October 29, 2018

Poisson: This time more confusing

Probability mass function:

$$f(Y|\mu) = \frac{(\mu)^Y e^{-\mu}}{Y!}, \qquad\qquad y = 0, 1, 2, \ldots, \ \mu > 0$$

1) Infinitesimal Interval. The probability of an arrival in the interval: $(t : \delta t)$ equals $\mu\delta t + \circ(\delta t)$ where μ is the "intensity" parameter and $\circ(\delta t)$ is a time interval with the property: $\lim_{\delta t \to 0} \frac{\circ(\delta t)}{\delta t} = 0$.

Poisson: This time more confusing

Probability mass function:

$$f(Y|\mu) = \frac{(\mu)^Y e^{-\mu}}{Y!}, \qquad\qquad y = 0, 1, 2, \ldots, \ \mu > 0$$

1) Infinitesimal Interval. The probability of an arrival in the interval: $(t : \delta t)$ equals $\mu\delta t + \circ(\delta t)$ where μ is the "intensity" parameter and $\circ(\delta t)$ is a time interval with the property: $\lim_{\delta t \to 0} \frac{\circ(\delta t)}{\delta t} = 0$.

- In other words, as the interval δt reduces in size towards zero, $\circ(\delta t)$ is negligible compared to δt.

Poisson: This time more confusing

Probability mass function:

$$f(Y|\mu) = \frac{(\mu)^Y e^{-\mu}}{Y!}, \qquad y = 0, 1, 2, \ldots, \ \mu > 0$$

1) Infinitesimal Interval. The probability of an arrival in the interval: $(t : \delta t)$ equals $\mu \delta t + \circ(\delta t)$ where μ is the "intensity" parameter and $\circ(\delta t)$ is a time interval with the property: $\lim_{\delta t \to 0} \frac{\circ(\delta t)}{\delta t} = 0$.

- In other words, as the interval δt reduces in size towards zero, $\circ(\delta t)$ is negligible compared to δt.
- This assumption is required to establish that μ adequately describes the intensity or expectation of arrivals.

Poisson: This time more confusing

Probability mass function:

$$f(Y|\mu) = \frac{(\mu)^Y e^{-\mu}}{Y!}, \qquad y = 0, 1, 2, \ldots, \ \mu > 0$$

1) Infinitesimal Interval. The probability of an arrival in the interval: $(t : \delta t)$ equals $\mu\delta t + \circ(\delta t)$ where μ is the "intensity" parameter and $\circ(\delta t)$ is a time interval with the property: $\lim_{\delta t \to 0} \frac{\circ(\delta t)}{\delta t} = 0$.

- In other words, as the interval δt reduces in size towards zero, $\circ(\delta t)$ is negligible compared to δt.
- This assumption is required to establish that μ adequately describes the intensity or expectation of arrivals.
- Typically there is no problem meeting this assumption provided that the time measure is adequately granular with respect to arrival rates.

2) Non-Simultaneity of Events. The probability of more than one arrival in the interval: $(t : \delta t)$equals $\circ(\delta t)$. Since $\circ(\delta t)$is negligible with respect to $\mu\delta t$ for sufficiently small $\mu\delta t$, the probability of simultaneous arrivals approaches zero in the limit.

3) I.I.D. Arrivals. The number of arrivals in any two consecutive or non-consecutive intervals are independent and identically distributed. More specifically, $P(Y = y) \ \in \ (T_j : T_{j+1})$ does not depend on $P(Y = y \ \in \ (T_k : T_{k+1})$ for any $j \neq k$.

- The intensity parameter (μ) is both the mean and variance for a single Poisson distributed random variable.

- The intensity parameter (μ) is both the mean and variance for a single Poisson distributed random variable.
- The intensity parameter is tied to a time interval, and rescaling time rescales the intensity parameter.

- The intensity parameter (μ) is both the mean and variance for a single Poisson distributed random variable.
- The intensity parameter is tied to a time interval, and rescaling time rescales the intensity parameter.
- Sums of independent Poisson random variables are themselves Poisson.

- The intensity parameter (μ) is both the mean and variance for a single Poisson distributed random variable.
- The intensity parameter is tied to a time interval, and rescaling time rescales the intensity parameter.
- Sums of independent Poisson random variables are themselves Poisson.
- We can also specifically model time by including it in the intensity parameter: $\mu^* = \mu t$.

Relationship to other models

- Poisson assumption is that there is no upper limit; if there is one use a binomial PMF.

Relationship to other models

- Poisson assumption is that there is no upper limit; if there is one use a binomial PMF.
- If $\mu = np$ as $n \to \infty$, then the Poisson is a good approximation for the binomial.

Relationship to other models

- Poisson assumption is that there is no upper limit; if there is one use a binomial PMF.
- If $\mu = np$ as $n \to \infty$, then the Poisson is a good approximation for the binomial.
- If n is small, then $\text{logit}(p) \approx \log(p)$, so the logit model is close to the Poisson model.

Relationship to other models

- Poisson assumption is that there is no upper limit; if there is one use a binomial PMF.
- If $\mu = np$ as $n \to \infty$, then the Poisson is a good approximation for the binomial.
- If n is small, then $\text{logit}(p) \approx \log(p)$, so the logit model is close to the Poisson model.
- If counts are bins, then use the multinomial PMF

Derivation of the link

- Definition:

$$\log(\mu_i) = \eta_i \;\Rightarrow\; \mu_i = \exp(\eta_i) = \exp(\mathbf{X}_i\boldsymbol{\beta})$$

Derivation of the link

- Definition:

$$\log(\mu_i) = \eta_i \;\Rightarrow\; \mu_i = \exp(\eta_i) = \exp(\mathbf{X}_i\boldsymbol{\beta})$$

- Start with the substitution:

$$L(\boldsymbol{\beta}|\mathbf{y}) = \prod_{i=1}^{n} \frac{e^{-\mu}\mu^{y_i}}{y_i!}\Big|_{\mu_i=\exp(\mathbf{X}_i\boldsymbol{\beta})} = \prod_{i=1}^{n} e^{-\exp(\mathbf{X}_i\boldsymbol{\beta})}\exp(\mathbf{X}_i\boldsymbol{\beta})^{y_i}/y_i!$$

Derivation of the link

- Definition:

$$\log(\mu_i) = \eta_i \;\Rightarrow\; \mu_i = \exp(\eta_i) = \exp(\mathbf{X}_i\boldsymbol{\beta})$$

- Start with the substitution:

$$L(\boldsymbol{\beta}|\mathbf{y}) = \prod_{i=1}^{n} \frac{e^{-\mu}\mu^{y_i}}{y_i!}\Big|_{\mu_i=\exp(\mathbf{X}_i\boldsymbol{\beta})} = \prod_{i=1}^{n} e^{-\exp(\mathbf{X}_i\boldsymbol{\beta})}\exp(\mathbf{X}_i\boldsymbol{\beta})^{y_i}/y_i!$$

- Take the log:

$$\ell(\boldsymbol{\beta}|\mathbf{y}) = \sum_{i=1}^{n}\left[-\exp(\mathbf{X}_i\boldsymbol{\beta}) + y_i(\mathbf{X}_i\boldsymbol{\beta}) - \log(y_i!)\right]$$

Derivation of the link

- Definition:

$$\log(\mu_i) = \eta_i \;\Rightarrow\; \mu_i = \exp(\eta_i) = \exp(\mathbf{X}_i\boldsymbol{\beta})$$

- Start with the substitution:

$$L(\boldsymbol{\beta}|\mathbf{y}) = \prod_{i=1}^{n} \frac{e^{-\mu}\mu^{y_i}}{y_i!}\Big|_{\mu_i=\exp(\mathbf{X}_i\boldsymbol{\beta})} = \prod_{i=1}^{n} e^{-\exp(\mathbf{X}_i\boldsymbol{\beta})}\exp(\mathbf{X}_i\boldsymbol{\beta})^{y_i}/y_i!$$

- Take the log:

$$\ell(\boldsymbol{\beta}|\mathbf{y}) = \sum_{i=1}^{n}\left[-\exp(\mathbf{X}_i\boldsymbol{\beta}) + y_i(\mathbf{X}_i\boldsymbol{\beta}) - \log(y_i!)\right]$$

- Now take the first derivative:

$$\frac{d}{d\boldsymbol{\beta}}\ell(\boldsymbol{\beta}|\mathbf{y}) = \sum_{i=1}^{n}\left[\exp(\mathbf{X}_i\boldsymbol{\beta})\mathbf{X}_j + \mathbf{y}_i\mathbf{X}_j\right], \qquad \forall j$$

Derivation of the link

- Definition:

$$\log(\mu_i) = \eta_i \;\Rightarrow\; \mu_i = \exp(\eta_i) = \exp(\mathbf{X}_i\boldsymbol{\beta})$$

- Start with the substitution:

$$L(\boldsymbol{\beta}|\mathbf{y}) = \prod_{i=1}^{n} \frac{e^{-\mu}\mu^{y_i}}{y_i!}\Big|_{\mu_i=\exp(\mathbf{X}_i\boldsymbol{\beta})} = \prod_{i=1}^{n} e^{-\exp(\mathbf{X}_i\boldsymbol{\beta})}\exp(\mathbf{X}_i\boldsymbol{\beta})^{y_i}/y_i!$$

- Take the log:

$$\ell(\boldsymbol{\beta}|\mathbf{y}) = \sum_{i=1}^{n}\left[-\exp(\mathbf{X}_i\boldsymbol{\beta}) + y_i(\mathbf{X}_i\boldsymbol{\beta}) - \log(y_i!)\right]$$

- Now take the first derivative:

$$\frac{d}{d\boldsymbol{\beta}}\ell(\boldsymbol{\beta}|\mathbf{y}) = \sum_{i=1}^{n}\left[\exp(\mathbf{X}_i\boldsymbol{\beta})\mathbf{X}_j + \mathbf{y}_i\mathbf{X}_j\right], \qquad \forall j$$

- Problem: there does not exist a closed form solution for $\hat{\boldsymbol{\beta}}$, so we use numerical methods.

- Sub-Saharan Africa has experienced a disproportionately high proportion of regime changes due to the military takeover of government for a variety of reasons, including ethnic fragmentation, arbitrary borders, economic problems, outside intervention, and poorly developed governmental institutions.

- Sub-Saharan Africa has experienced a disproportionately high proportion of regime changes due to the military takeover of government for a variety of reasons, including ethnic fragmentation, arbitrary borders, economic problems, outside intervention, and poorly developed governmental institutions.
- These data, selected from a larger set given by Bratton and Van De Walle (1994), look at potential causal factors for counts of military coups (ranging from 0 to 6 events) in 33 sub-Saharan countries over the period from each country's colonial independence to 1989.

- Sub-Saharan Africa has experienced a disproportionately high proportion of regime changes due to the military takeover of government for a variety of reasons, including ethnic fragmentation, arbitrary borders, economic problems, outside intervention, and poorly developed governmental institutions.
- These data, selected from a larger set given by Bratton and Van De Walle (1994), look at potential causal factors for counts of military coups (ranging from 0 to 6 events) in 33 sub-Saharan countries over the period from each country's colonial independence to 1989.
- Seven explanatory variables are chosen here to model the count of military coups: **Military Oligarchy** (the number of years of this type of rule); **Political Liberalization** (0 for no observable civil rights for political expression, 1 for limited, and 2 for extensive); **Parties** (number of legally registered political parties); **Percent Legislative Voting**; **Percent Registered Voting**; **Size** (in one thousand square kilometer units); and **Population** (given in millions).

- A generalized linear model for these data with the Poisson link function is specified as:

$$g^{-1}(\boldsymbol{\theta}) = g^{-1}(\mathbf{X}\boldsymbol{\beta}) = \exp\left[\mathbf{X}\boldsymbol{\beta}\right] = E[\mathbf{Y}] = E[\textbf{Military Coups}].$$

- A generalized linear model for these data with the Poisson link function is specified as:

$$g^{-1}(\theta) = g^{-1}(\mathbf{X}\beta) = \exp[\mathbf{X}\beta] = E[\mathbf{Y}] = E[\textbf{Military Coups}].$$

- In this specification, the systematic component is $\mathbf{X}\beta$, the stochastic component is $\mathbf{Y} = \textbf{Military Coups}$, and the link function is $\theta = \log(\mu)$.

- A generalized linear model for these data with the Poisson link function is specified as:

$$g^{-1}(\boldsymbol{\theta}) = g^{-1}(\mathbf{X}\boldsymbol{\beta}) = \exp\left[\mathbf{X}\boldsymbol{\beta}\right] = E[\mathbf{Y}] = E[\textbf{Military Coups}].$$

- In this specification, the systematic component is $\mathbf{X}\boldsymbol{\beta}$, the stochastic component is $\mathbf{Y} = \textbf{Military Coups}$, and the link function is $\boldsymbol{\theta} = \log(\boldsymbol{\mu})$.
- We can re-express this model by moving the link function to the left-hand side exposing the linear predictor: $g(\boldsymbol{\mu}) = \log(E[\mathbf{Y}]) = \mathbf{X}\boldsymbol{\beta}$ (although this is now a less intuitive form for understanding the outcome variable).

- A generalized linear model for these data with the Poisson link function is specified as:

$$g^{-1}(\boldsymbol{\theta}) = g^{-1}(\mathbf{X}\boldsymbol{\beta}) = \exp\left[\mathbf{X}\boldsymbol{\beta}\right] = E[\mathbf{Y}] = E[\textbf{Military Coups}].$$

- In this specification, the systematic component is $\mathbf{X}\boldsymbol{\beta}$, the stochastic component is $\mathbf{Y} = \textbf{Military Coups}$, and the link function is $\boldsymbol{\theta} = \log(\boldsymbol{\mu})$.
- We can re-express this model by moving the link function to the left-hand side exposing the linear predictor: $g(\boldsymbol{\mu}) = \log(E[\mathbf{Y}]) = \mathbf{X}\boldsymbol{\beta}$ (although this is now a less intuitive form for understanding the outcome variable).
- The R language GLM call for this model is:

```
africa.out<-glm(MILTCOUP ~ MILITARY+POLLIB+PARTY93+PCTVOTE+PCTTURN
                           +SIZE*POP+NUMREGIM*NUMELEC, family=poisson)
```

- A generalized linear model for these data with the Poisson link function is specified as:

$$g^{-1}(\boldsymbol{\theta}) = g^{-1}(\mathbf{X}\boldsymbol{\beta}) = \exp\left[\mathbf{X}\boldsymbol{\beta}\right] = E[\mathbf{Y}] = E[\textbf{Military Coups}].$$

- In this specification, the systematic component is $\mathbf{X}\boldsymbol{\beta}$, the stochastic component is $\mathbf{Y} = \textbf{Military Coups}$, and the link function is $\boldsymbol{\theta} = \log(\boldsymbol{\mu})$.
- We can re-express this model by moving the link function to the left-hand side exposing the linear predictor: $g(\boldsymbol{\mu}) = \log(E[\mathbf{Y}]) = \mathbf{X}\boldsymbol{\beta}$ (although this is now a less intuitive form for understanding the outcome variable).
- The `R` language GLM call for this model is:

```
africa.out<-glm(MILTCOUP ~ MILITARY+POLLIB+PARTY93+PCTVOTE+PCTTURN
                           +SIZE*POP+NUMREGIM*NUMELEC, family=poisson)
```

- The new part is `family=poisson`, where poisson is not capitalized.

	Parameter Estimate	Standard Error	95% Confidence Interval
(Intercept)	2.9209	1.3368	[0.3008: 5.5410]
Military Oligarchy	0.1709	0.0509	[0.0711: 0.2706]
Political Liberalization	-0.4654	0.3319	[-1.1160: 0.1851]
Parties	0.0248	0.0109	[0.0035: 0.0460]
Percent Legislative Voting	0.0613	0.0218	[0.0187: 0.1040]
Percent Registered Voting	-0.0361	0.0137	[-0.0629:-0.0093]
Size	-0.0018	0.0007	[-0.0033:-0.0004]
Population	-0.1188	0.0397	[-0.1965:-0.0411]
Regimes	-0.8662	0.4571	[-1.7621: 0.0298]
Elections	-0.4859	0.2118	[-0.9010:-0.0709]
(Size)(Population)	0.0001	0.0001	[0.0001: 0.0002]
(Regimes)(Elections)	0.1810	0.0689	[0.0459: 0.3161]

- Note that the two interaction terms are specified by using the multiplication character. The iteratively weighted least squares algorithm converged in only four iterations using Fisher scoring, and the results are provided in the table.

- Note that the two interaction terms are specified by using the multiplication character. The iteratively weighted least squares algorithm converged in only four iterations using Fisher scoring, and the results are provided in the table.
- The model appears to fit the data quite well:

- Note that the two interaction terms are specified by using the multiplication character. The iteratively weighted least squares algorithm converged in only four iterations using Fisher scoring, and the results are provided in the table.
- The model appears to fit the data quite well:
- an improvement from the null deviance of 62 on 32 degrees of freedom to a residual deviance of 7.5 on 21 degrees of freedom

- Note that the two interaction terms are specified by using the multiplication character. The iteratively weighted least squares algorithm converged in only four iterations using Fisher scoring, and the results are provided in the table.
- The model appears to fit the data quite well:
- an improvement from the null deviance of 62 on 32 degrees of freedom to a residual deviance of 7.5 on 21 degrees of freedom
- evidence that the model does not fit would be supplied by a model deviance value in the tail of a χ^2_{n-k} distribution

- Note that the two interaction terms are specified by using the multiplication character. The iteratively weighted least squares algorithm converged in only four iterations using Fisher scoring, and the results are provided in the table.
- The model appears to fit the data quite well:
- an improvement from the null deviance of 62 on 32 degrees of freedom to a residual deviance of 7.5 on 21 degrees of freedom
- evidence that the model does not fit would be supplied by a model deviance value in the tail of a χ^2_{n-k} distribution
- and nearly all the coefficients have 95% confidence intervals bounded away from zero and therefore appear reliable in the model.

Back to deviance and model fit

- General Deviance Notation: $D = \sum_{i=1}^{n} d(\boldsymbol{\eta}, y_i)$, where the individual deviance function is defined as:
  $d(\boldsymbol{\eta}, y_i) = -2\left[\ell(\hat{\boldsymbol{\eta}}, \psi|y_i) - \ell(\tilde{\boldsymbol{\eta}}, \psi|y_i)\right]$, where $\tilde{\boldsymbol{\eta}}$is the saturated estimate.

Back to deviance and model fit

- General Deviance Notation: $D = \sum_{i=1}^{n} d(\boldsymbol{\eta}, y_i)$, where the individual deviance function is defined as: $d(\boldsymbol{\eta}, y_i) = -2\left[\ell(\hat{\boldsymbol{\eta}}, \psi | y_i) - \ell(\tilde{\boldsymbol{\eta}}, \psi | y_i)\right]$, where $\tilde{\boldsymbol{\eta}}$is the saturated estimate.
- Linear Model Residual Vector: $\mathbf{R}_{standard} = \mathbf{Y} - \mathbf{X}\beta$.

Back to deviance and model fit

- General Deviance Notation: $D = \sum_{i=1}^{n} d(\boldsymbol{\eta}, y_i)$, where the individual deviance function is defined as: $d(\boldsymbol{\eta}, y_i) = -2\left[\ell(\hat{\boldsymbol{\eta}}, \psi|y_i) - \ell(\tilde{\boldsymbol{\eta}}, \psi|y_i)\right]$, where $\tilde{\boldsymbol{\eta}}$is the saturated estimate.
- Linear Model Residual Vector: $\mathbf{R}_{standard} = \mathbf{Y} - \mathbf{X}\beta$.
- Response Residual Vector: $\mathbf{R}_{Response} = \mathbf{Y} - g^{-1}(\boldsymbol{X}\boldsymbol{\beta}) = \mathbf{Y} - \hat{\boldsymbol{\mu}}$.

Back to deviance and model fit

- General Deviance Notation: $D = \sum_{i=1}^{n} d(\boldsymbol{\eta}, y_i)$, where the individual deviance function is defined as:
  $d(\boldsymbol{\eta}, y_i) = -2\left[\ell(\hat{\boldsymbol{\eta}}, \psi | y_i) - \ell(\tilde{\boldsymbol{\eta}}, \psi | y_i)\right]$, where $\tilde{\boldsymbol{\eta}}$is the saturated estimate.
- Linear Model Residual Vector: $\mathbf{R}_{standard} = \mathbf{Y} - \mathbf{X}\beta$.
- Response Residual Vector: $\mathbf{R}_{Response} = \mathbf{Y} - g^{-1}(\boldsymbol{X\beta}) = \mathbf{Y} - \hat{\boldsymbol{\mu}}$.
- Pearson Residual Vector: $\mathbf{R}_{Pearson} = \frac{\mathbf{Y} - \hat{\mu}}{\sqrt{VAR[\hat{\boldsymbol{\mu}}]}}$ (the sum of the Pearson residuals for a Poisson generalized linear model is the Pearson χ^2 goodness-of-fit measure).

Back to deviance and model fit

- General Deviance Notation: $D = \sum_{i=1}^{n} d(\boldsymbol{\eta}, y_i)$, where the individual deviance function is defined as:
  $d(\boldsymbol{\eta}, y_i) = -2\left[\ell(\hat{\boldsymbol{\eta}}, \psi | y_i) - \ell(\tilde{\boldsymbol{\eta}}, \psi | y_i)\right]$, where $\tilde{\boldsymbol{\eta}}$is the saturated estimate.
- Linear Model Residual Vector: $\mathbf{R}_{standard} = \mathbf{Y} - \mathbf{X}\beta$.
- Response Residual Vector: $\mathbf{R}_{Response} = \mathbf{Y} - g^{-1}(\boldsymbol{X}\boldsymbol{\beta}) = \mathbf{Y} - \hat{\boldsymbol{\mu}}$.
- Pearson Residual Vector: $\mathbf{R}_{Pearson} = \frac{\mathbf{Y} - \hat{\mu}}{\sqrt{VAR[\hat{\boldsymbol{\mu}}]}}$ (the sum of the Pearson residuals for a Poisson generalized linear model is the Pearson χ^2 goodness-of-fit measure).
- Working Residual Vector: $\mathbf{R}_{Working} = (\mathbf{y} - \hat{\boldsymbol{\mu}}) \frac{\partial}{\partial \eta} \hat{\boldsymbol{\mu}}$ (from the last step of Iteratively Reweighted Least Squares algorithm).

In the Poisson model

- The "G-statistic" (summed deviance) for this model is:

$$D_{\text{Poisson}} = 2\sum_{i=1}^{n} \left(y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)\right) \underset{a}{\sim} \chi^2_{n-p},$$

where p is the number of explanatory variables including the constant, and $\hat{\mu}_i$) is the predicted outcome for the ith case.

In the Poisson model

- The "G-statistic" (summed deviance) for this model is:

$$D_{\text{Poisson}} = 2 \sum_{i=1}^{n} \left(y_i \log(y_i / \hat{\mu}_i) - (y_i - \hat{\mu}_i) \right) \underset{a}{\sim} \chi^2_{n-p},$$

  where p is the number of explanatory variables including the constant, and $\hat{\mu}_i)$ is the predicted outcome for the ith case.

- Individual Deviance Function:

$$R_{Deviance} = \frac{(y_i - \hat{\mu}_i)}{|y_i - \hat{\mu}_i|} \sqrt{|d(\boldsymbol{\eta}, y_i)|}$$

  where

$$d(\boldsymbol{\eta}, y_i) = -2 \left[\ell(\hat{\boldsymbol{\eta}}, \psi | y_i) - \ell(\tilde{\boldsymbol{\eta}}, \psi | y_i) \right].$$

- Recall also the Pearson's statistic:

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \underset{a}{\sim} \chi^2_{n-p}.$$

- Recall also the Pearson's statistic:

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \underset{\text{a}}{\sim} \chi^2_{n-p}.$$

- Generally the summed deviance is more robust.

New and Old Ways to Look at Model Fit

- Approximation to Pearson's Statistic.

$$X^2 = \sum_{i=1}^{n} \mathbf{R}^2_{Pearson} = \sum_{i=1}^{n} \left[\frac{\mathbf{Y} - \mu}{\sqrt{VAR[\mu]}} \right]^2 .$$

New and Old Ways to Look at Model Fit

- Approximation to Pearson's Statistic.

$$X^2 = \sum_{i=1}^{n} \mathbf{R}^2_{Pearson} = \sum_{i=1}^{n} \left[\frac{\mathbf{Y} - \boldsymbol{\mu}}{\sqrt{VAR[\boldsymbol{\mu}]}} \right]^2 .$$

- If the sample size is sufficiently large, then $\frac{X^2}{a(\psi)} \sim \chi^2_{n-p}$ where n is the sample size, p is the number of explanatory variables including the constant, and $a(\psi)$ is the scale function that we'll see in Chapter 6.

New and Old Ways to Look at Model Fit

- Approximation to Pearson's Statistic.

$$X^2 = \sum_{i=1}^{n} \mathbf{R}^2_{Pearson} = \sum_{i=1}^{n} \left[\frac{\mathbf{Y} - \boldsymbol{\mu}}{\sqrt{VAR[\boldsymbol{\mu}]}} \right]^2 .$$

- If the sample size is sufficiently large, then $\frac{X^2}{a(\psi)} \sim \chi^2_{n-p}$ where n is the sample size, p is the number of explanatory variables including the constant, and $a(\psi)$ is the scale function that we'll see in Chapter 6.
- For the summed deviance with sufficient sample size it is also true that $D(\boldsymbol{\eta}, \mathbf{y}) / a(\psi) \sim \chi^2_{n-p}$.

New and Old Ways to Look at Model Fit

- Approximation to Pearson's Statistic.

$$X^2 = \sum_{i=1}^{n} \mathbf{R}^2_{Pearson} = \sum_{i=1}^{n} \left[\frac{\mathbf{Y} - \boldsymbol{\mu}}{\sqrt{VAR[\boldsymbol{\mu}]}} \right]^2 .$$

- If the sample size is sufficiently large, then $\frac{X^2}{a(\psi)} \sim \chi^2_{n-p}$ where n is the sample size, p is the number of explanatory variables including the constant, and $a(\psi)$ is the scale function that we'll see in Chapter 6.
- For the summed deviance with sufficient sample size it is also true that $D(\boldsymbol{\eta}, \mathbf{y})/a(\psi) \sim \chi^2_{n-p}$.
- Recall that it is also common to contrast this with the *null deviance*: the deviance function calculated for a model with no covariates (mean function only).

New and Old Ways to Look at Model Fit

- Akaike Information Criterion.
  minimizes the negative likelihood penalized by the number of parameters:
  $$\text{AIC} = -2\ell(\hat{\boldsymbol{\beta}}|\mathbf{y}) + 2p$$
  where $\ell(\hat{\boldsymbol{\beta}}|\mathbf{y})$ is the maximized model log likelihood value and p is the number of explanatory variables in the model (including the constant). (AIC has a bias towards models that overfit with extra parameters since the penalty component is obviously linear with increases in the number of explanatory variables, and the log likelihood often increases more rapidly.)

New and Old Ways to Look at Model Fit

- Akaike Information Criterion.
  minimizes the negative likelihood penalized by the number of parameters:
  $$\text{AIC} = -2\ell(\hat{\boldsymbol{\beta}}|\mathbf{y}) + 2p$$
  where $\ell(\hat{\boldsymbol{\beta}}|\mathbf{y})$ is the maximized model log likelihood value and p is the number of explanatory variables in the model (including the constant). (AIC has a bias towards models that overfit with extra parameters since the penalty component is obviously linear with increases in the number of explanatory variables, and the log likelihood often increases more rapidly.)
- Schwartz Criterion/Bayesian Information Criterion (BIC).
  $$\text{BIC} = -2\ell(\hat{\boldsymbol{\beta}}|\mathbf{y}) + p\log(n)$$
  where n is the sample size.

New and Old Ways to Look at Model Fit

- Akaike Information Criterion.
  minimizes the negative likelihood penalized by the number of parameters:
  $$\text{AIC} = -2\ell(\hat{\boldsymbol{\beta}}|\mathbf{y}) + 2p$$
  where $\ell(\hat{\boldsymbol{\beta}}|\mathbf{y})$ is the maximized model log likelihood value and p is the number of explanatory variables in the model (including the constant). (AIC has a bias towards models that overfit with extra parameters since the penalty component is obviously linear with increases in the number of explanatory variables, and the log likelihood often increases more rapidly.)
- Schwartz Criterion/Bayesian Information Criterion (BIC).
  $$\text{BIC} = -2\ell(\hat{\boldsymbol{\beta}}|\mathbf{y}) + p\log(n)$$
  where n is the sample size.
- There is also a Deviance Information Criterion (DIC) used in Bayesian MCMC estimation.

Congressional Cosponsoring of Bills

Fowler (2006) looks at patterns of sponsorship and cosponsorship in Congress from 1973 to 2004.

```
cosponsor <- read.table("fowler.dat", header=TRUE);  head(cosponsor,4)
   Congress    Period Total.Sponsors Total.Bills
1 93rd 19731974            442       20994
2 94th 19751976            439       19275
3 95th 19771978            437       18578
4 96th 19791980            436       10478
  Mean.Bills.Per.Leg Mean.Cos.Per.Leg Mean.Cos.Per.Bill
1                 48              129                 3
2                 44              151                 3
3                 42              170                 4
4                 24              187                 8
  Cos.Per.Leg Mean.Dist Senate
1          70      1.95      0
2          79      1.89      0
3          93      1.83      0
4         111      1.76      0
```

Look at summary statistics:

```
mean(cosponsor$Mean.Bills.Per.Leg)
[1] 47.625
var(cosponsor$Mean.Bills.Per.Leg)
[1] 828.24

mean(cosponsor$Mean.Cos.Per.Leg)
[1] 247.5
var(cosponsor$Mean.Cos.Per.Leg)
[1] 6134.7
```

This is clear evidence of *overdispersion* in count data.

- For Poisson models the mean and the variance of a single random variable are assumed to be the same.

- For Poisson models the mean and the variance of a single random variable are assumed to be the same.
- For the likelihood function as a statistic, the variance is scaled by n.

- For Poisson models the mean and the variance of a single random variable are assumed to be the same.
- For the likelihood function as a statistic, the variance is scaled by n.
- Overdispersion, $\text{Var}(Y) > E(Y)$, is relatively common, whereas underdispersion, $\text{Var}(Y) < E(Y)$ is rare.

- For Poisson models the mean and the variance of a single random variable are assumed to be the same.
- For the likelihood function as a statistic, the variance is scaled by n.
- Overdispersion, $\text{Var}(Y) > E(Y)$, is relatively common, whereas underdispersion, $\text{Var}(Y) < E(Y)$ is rare.
- Biggest effect is to make the standard errors wrong.

- For Poisson models the mean and the variance of a single random variable are assumed to be the same.
- For the likelihood function as a statistic, the variance is scaled by n.
- Overdispersion, $\text{Var}(Y) > E(Y)$, is relatively common, whereas underdispersion, $\text{Var}(Y) < E(Y)$ is rare.
- Biggest effect is to make the standard errors wrong.
- One diagnostic: plot $\hat{\mu}$ versus $(y - \hat{\mu})^2$.

- For Poisson models the mean and the variance of a single random variable are assumed to be the same.
- For the likelihood function as a statistic, the variance is scaled by n.
- Overdispersion, $\mathrm{Var}(Y) > E(Y)$, is relatively common, whereas underdispersion, $\mathrm{Var}(Y) < E(Y)$ is rare.
- Biggest effect is to make the standard errors wrong.
- One diagnostic: plot $\hat{\mu}$ versus $(y - \hat{\mu})^2$.
- Solution: make μ a random variable rather than a fixed constant to be estimated, with a gamma distribution: $G[\mu\alpha, \alpha]$. So

$$E[Y] = \mu \qquad\qquad \mathrm{Var}[Y] = \frac{\mu}{\phi}$$

- For Poisson models the mean and the variance of a single random variable are assumed to be the same.
- For the likelihood function as a statistic, the variance is scaled by n.
- Overdispersion, $\text{Var}(Y) > E(Y)$, is relatively common, whereas underdispersion, $\text{Var}(Y) < E(Y)$ is rare.
- Biggest effect is to make the standard errors wrong.
- One diagnostic: plot $\hat{\mu}$ versus $(y - \hat{\mu})^2$.
- Solution: make μ a random variable rather than a fixed constant to be estimated, with a gamma distribution: $G[\mu\alpha, \alpha]$. So

$$E[Y] = \mu \qquad\qquad \text{Var}[Y] = \frac{\mu}{\phi}$$

- This is called the “Poisson-Gamma” model and it means that Y is distributed *negative binomial*.

- Negative binomial distribution has the same sample space as the Poisson, but contains an additional parameter which can be thought of as gamma distributed and therefore used to model a variance function.

- Negative binomial distribution has the same sample space as the Poisson, but contains an additional parameter which can be thought of as gamma distributed and therefore used to model a variance function.
- Used by many to fit a count model with overdispersion.

- Negative binomial distribution has the same sample space as the Poisson, but contains an additional parameter which can be thought of as gamma distributed and therefore used to model a variance function.
- Used by many to fit a count model with overdispersion.
- The binomial distribution measures the number of successes in a given number of fixed trials, whereas the negative binomial distribution measures *the number of failures, y before the r^{th} success*.

- Negative binomial distribution has the same sample space as the Poisson, but contains an additional parameter which can be thought of as gamma distributed and therefore used to model a variance function.
- Used by many to fit a count model with overdispersion.
- The binomial distribution measures the number of successes in a given number of fixed trials, whereas the negative binomial distribution measures *the number of failures, y before the r^{th} success*.
- An alternative but equivalent form,

$$f(y|r,p) = \binom{y-1}{r-1} p^r (1-p)^{y-r},$$

measures the number of trials necessary to get r successes.

- Negative binomial distribution has the same sample space as the Poisson, but contains an additional parameter which can be thought of as gamma distributed and therefore used to model a variance function.
- Used by many to fit a count model with overdispersion.
- The binomial distribution measures the number of successes in a given number of fixed trials, whereas the negative binomial distribution measures *the number of failures, y before the r^{th} success*.
- An alternative but equivalent form,

$$f(y|r,p) = \binom{y-1}{r-1} p^r (1-p)^{y-r},$$

measures the number of trials necessary to get r successes.
- An important application of the negative binomial distribution is in survey research design. If the researcher knows the value of p from previous surveys, then the negative binomial can provide the number of subjects to contact in order to get the desired number of responses for analysis.

Negative binomial

- The PMF is:

$$f(Y|k,p) = \binom{y-1}{k-1} p^k (1-p)^{y-k},$$

$$y = 0, 1, 2, \ldots,$$

$$0 \leq p \leq 1.$$

Negative binomial

- The PMF is:

$$f(Y|k,p) = \binom{y-1}{k-1} p^k (1-p)^{y-k},$$

$$y = 0, 1, 2, \ldots,$$

$$0 \leq p \leq 1.$$

- For this parameterization, we get:

$$E[Y] = \mu, \qquad \mathrm{Var}[Y] = \frac{\mu(1+\phi)}{\phi}.$$

- If ϕ (the dispersion parameter) is unknown, use the estimate:

$$\hat{\phi} = \frac{X^2}{n-p} = \frac{\sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}}{n-p}.$$

- If ϕ (the dispersion parameter) is unknown, use the estimate:

$$\hat{\phi} = \frac{X^2}{n-p} = \frac{\sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}}{n-p}.$$

- This gives an F-test for comparing models (big values implies a difference in models).

Two interpretations

- (1) as a generalized Poisson,

Two interpretations

- (1) as a generalized Poisson,
- (2) with probability p, modeling the number of trials, Y, before the kth success (alternatively failure) where k is fixed in advance.

Two interpretations

- (1) as a generalized Poisson,
- (2) with probability p, modeling the number of trials, Y, before the kth success (alternatively failure) where k is fixed in advance.
- For estimation, use `library(MASS)`, which has `glm.nb`.

- Compare the number of bills assigned to committee in the first 100 days of the 103^{rd} and 104^{th} Houses as a function of the number of members on the committee, the number of subcommittees, the number of staff assigned to the committee, and a dummy variable indicating whether or not it is a high prestige committee.

- Compare the number of bills assigned to committee in the first 100 days of the 103^{rd} and 104^{th} Houses as a function of the number of members on the committee, the number of subcommittees, the number of staff assigned to the committee, and a dummy variable indicating whether or not it is a high prestige committee.
- The model is developed with the link function:

$$\eta = g(\mu) = \log\left(\frac{\mu}{\mu + \frac{1}{k}}\right) \quad \longrightarrow \quad \mu = g^{-1}(\eta) = \frac{\exp(\eta)}{k(1 - \exp(\eta))},$$

where $\eta = \mathbf{X}\boldsymbol{\beta}$, and $k \geq 1$ is the overdispersion term.

Committee	Size	Subcommittees	Staff	Prestige	Bills–103rd
Appropriations	58	13	109	1	9
Budget	42	0	39	1	101
Rules	13	2	25	1	54
Ways and Means	39	5	23	1	542
Banking	51	5	61	0	101
Economic/Educ. Opportunities	43	5	69	0	158
Commerce	49	4	79	0	196
International Relations	44	3	68	0	40
Government Reform	51	7	99	0	72
Judiciary	35	5	56	0	168
Agriculture	49	5	46	0	60
National Security	55	7	48	0	75
Resources	44	5	58	0	98
Transport./Infrastructure	61	6	74	0	69
Science	50	4	58	0	25
Small Business	43	4	29	0	9
Veterans Affairs	33	3	36	0	41
House Oversight	12	0	24	0	233
Standards of Conduct	10	0	9	0	0
Intelligence	16	2	24	0	2

```
committee.poisson <- glm(BILLS104 ~ SIZE + SUBS * (log(STAFF))
                        + PRESTIGE + BILLS103, family=poisson,
                         data=committee.dat)
1 - pchisq(summary(committee.poisson)$deviance,
           summary(committee.poisson)$df.residual)
[1] 0   # IN THE TAIL INDICATES OVERDISPERSION

committee.out <- glm.nb(BILLS104 ~ SIZE + SUBS * (log(STAFF))
                      + PRESTIGE + BILLS103, data=committee.dat)

resp <- resid(committee.out,type="response")
pears <- resid(committee.out,type="pearson")
working <- resid(committee.out,type="working")
devs <- resid(committee.out,type="deviance")
cbind(resp,pears,working,devs)
```

	resp	pears	working	devs
Appropriations	-7.38308	-0.99451	-0.55167	-1.22671
Budget	-6.17325	-0.40931	-0.21161	-0.43997
Rules	22.54158	1.98665	1.05048	1.56745
Ways_and_Means	-135.06135	-0.56848	-0.27560	-0.63081
Banking	21.00117	0.40998	0.20194	0.38568
Economic_Educ_Oppor	-93.92104	-0.85695	-0.41757	-1.01572
Commerce	-58.03818	-0.36306	-0.17639	-0.38675
International_Relations	-49.33480	-0.89295	-0.43918	-1.06810
Government_Reform	32.60986	0.57003	0.28018	0.52480
Judiciary	27.80878	0.25343	0.12349	0.24378
Agriculture	24.21181	0.85168	0.42635	0.75680
National_Security	27.14348	0.87911	0.43881	0.77861
Resources	26.13708	0.45893	0.22559	0.42884
TransInfrastructure	79.10378	2.10068	1.04226	1.64133
Science	-34.35454	-1.12146	-0.55993	-1.43001
Small_Business	-12.50419	-1.14887	-0.60984	-1.48074
Veterans_Affairs	-14.18802	-0.66378	-0.33630	-0.75200
House_Oversight	16.14917	0.62009	0.31145	0.56716
Stds_of_Conduct	0.37836	0.44850	0.60864	0.40700
Intelligence	-13.58498	-1.43490	-0.77253	-2.05981

Results

	Coefficient	Standard Error	95% Confidence Interval
(Intercept)	-6.80543	2.54651	[-12.30683:-1.30402]
Size	-0.02825	0.02093	[-0.07345: 0.01696]
Subcommittees	1.30159	0.54370	[0.12701: 2.47619]
log(Staff)	3.00971	0.79450	[1.29329: 4.72613]
Prestige	-0.32367	0.44102	[-1.27644: 0.62911]
Bills in 103^{rd}	0.00656	0.00139	[0.00355: 0.00957]
Subcommittees:log(STAFF)	-0.32364	0.12489	[-0.59345:-0.05384]
Null deviance: 107.314, $df = 19$			Maximized $\ell()$: 10559
Summed deviance: 20.948, $df = 13$			AIC: 121130