Applied Regression Modeling for Cross-Section Data

Kosuke Imai

Princeton University

POL573 Quantitative Analysis III
Fall 2013

Suggested Readings and References

- King, 5.7–5.9,
- Wooldridge, 19.1–19.4
- Gelman and Hill, Chapter 6
- McCullagh and Nelder. (1989). *Generalized Linear Models*. Chapman & Hall. Especially, Chapters 2, 4, 9, and 12.

Event Count Data

- The number of events: $Y \in \{0, 1, 2, \dots\}$
- Poisson process: $Y(t) = \#$ of events up to time $t \geq 0$
  1. $Y(0) = 0$: zero event at the start
  2. $Y(t_1) \leq Y(t_2)$ for any $t_1 < t_2$: no "negative" event
  3. $\{Y(t_2) - Y(t_1)\} \perp\!\!\!\perp Y(t_1)$ for any $t_1 < t_2$: no contagion or diffusion
  4. For any t and sufficiently small h,

$$\begin{aligned} \Pr(Y(t+h) - Y(t) = 0) &\approx 1 - \mu h \\ \Pr(Y(t+h) - Y(t) = 1) &\approx \mu h \\ \Pr(Y(t+h) - Y(t) \geq 2) &\approx 0 \end{aligned}$$

  No simultaneous events, same probability at any time

- After some math, this leads to the Poisson distribution:

$$\Pr(Y(t) = y) = \frac{(\mu t)^y}{y!} \exp(-\mu t) \quad \text{or} \quad \Pr(Y = y) = \frac{\mu^y}{y!} \exp(-\mu)$$

Poisson Regression Model

- Model: $Y_i \mid X_i \overset{\text{indep.}}{\sim} \text{Poisson}(\mu_i)$ where $\mu_i = \exp(X_i^\top \beta)$
- Moments: $\mathbb{E}(Y_i \mid X_i) = \mathbb{V}(Y_i \mid X_i) = \mu_i$
- Likelihood and log-likelihood functions:

$$L_n(\beta \mid Y, X) \propto \exp\left(-\sum_{i=1}^n \mu_i\right) \prod_{i=1}^n \mu_i^{Y_i}$$

$$l_n(\beta \mid Y, X) = \text{constant} + \sum_{i=1}^n \left\{ Y_i X_i^\top \beta - \exp(X_i^\top \beta) \right\}$$

- Score: $\frac{\partial}{\partial \beta} l_n(\beta \mid Y) = \sum_{i=1}^n \{Y_i - \exp(X_i^\top \beta)\} X_i$
- Hessian: $\frac{\partial^2}{\partial \beta \partial \beta^\top} l_n(\beta \mid Y) = -\sum_{i=1}^n \exp(X_i^\top \beta) X_i X_i^\top$
- Quantities of interest: $\mathbb{E}(Y_i \mid X_i = x)$, $\Pr(Y_i = y \mid X_i = x)$, etc.
- Uncertainty: the Delta method or the Monte Carlo simulation

Poisson as a Limit of Binomial

- In the Poisson process, events are assumed to occur independently with equal probability
- Recall the Binomial distribution:

$$\Pr(Y = y \mid n, p) = \frac{n!}{y!(n-y)!} p^y (1-p)^{n-y}$$

- Rewrite this using $\mu = \mathbb{E}(Y \mid n, p) = np$ as

$$\Pr(Y = y \mid n, \mu) = \frac{n!}{y!(n-y)!} \left(\frac{\mu}{n}\right)^y \left(1 - \frac{\mu}{n}\right)^{n-y}$$

- Taking a limit w.r.t. n (while keeping μ constant), we have

$$\lim_{n \to \infty} \Pr(Y = y \mid n, \mu) = \frac{\mu^y}{y!} \exp(-\mu)$$

Overdispersion

- Definition: Larger variance than what is assumed in a model
- Underdispersion $\Longleftrightarrow$ Overdispersion
- Logistic/Probit regression $\mathbb{V}(Y_i \mid X_i) > \pi_i\{1 - \pi_i\}$
- Poisson regression $\mathbb{V}(Y_i \mid X_i) > \mathbb{E}(Y_i \mid X_i) = \mu_i$
- Potential sources of overdispersion:
  1. unobserved heterogeneity
  2. clustering
  3. contagion or diffusion
  4. (classical) measurement error
- Underdispersion is rare but could occur due to negative correlations induced by contagion and clustering

Negative-Binomial Distribution as Positive Contagion

- Independent event occurance in the Poisson process is unrealistic
- Pólya's urn for positive contagion:
  - n balls where np balls are white and $n(1-p)$ are black
  - m successive draws
  - after each draw the ball is replaced and nq balls of the color last drawn are added to the urn
- Let Y be the number of white balls in m successive draws:

$$\lim_{m \to \infty} \Pr(Y = y \mid \gamma, \rho) = \left(\begin{array}{c} \gamma\rho + y - 1 \\ y \end{array} \right) \left(\frac{\rho}{1+\rho} \right)^{\gamma\rho} \left(\frac{1}{1+\rho} \right)^{y}$$

  where $p \to 0$ and $q \to 0$ s.t. $mp \to \gamma$ and $mq \to \rho^{-1}$
- This is Negative-Binomial distribution:
  - How many failures must occur in order to have $\gamma\rho$ successes with success probability $\pi = \rho/(1+\rho)$?
  - Mean $\gamma\rho(1-\pi)/\pi \leq$ Variance $\gamma\rho(1-\pi)/\pi^2$

A Poisson Mixture Model

- Yet another way to generate the negative-binomial distribution!
- A Poisson mixture model to account for heterogeneity:

$$
\begin{aligned}
Y_i \mid \phi_i, \mu_i &\overset{\text{indep.}}{\sim} \text{Poisson}(\phi_i \mu_i) \quad \text{where} \quad \mu_i = \exp(X_i^\top \beta) \\
\phi_i &\overset{\text{i.i.d.}}{\sim} \frac{1}{\theta}\text{Gamma}(\theta, 1) \quad \text{with} \quad \theta > 0
\end{aligned}
$$

- Each observation comes from a *different* Poisson with $\phi_i \mu_i$
- Gamma is the mixing distribution
- Hierarchical (multilevel) structure
- The model is:

$$
f(Y_i \mid \mu_i, \theta) = \int_0^\infty f(Y_i \mid \mu_i, \phi_i) f(\phi_i \mid \theta) d\phi_i
$$

A Review of the Gamma Distribution

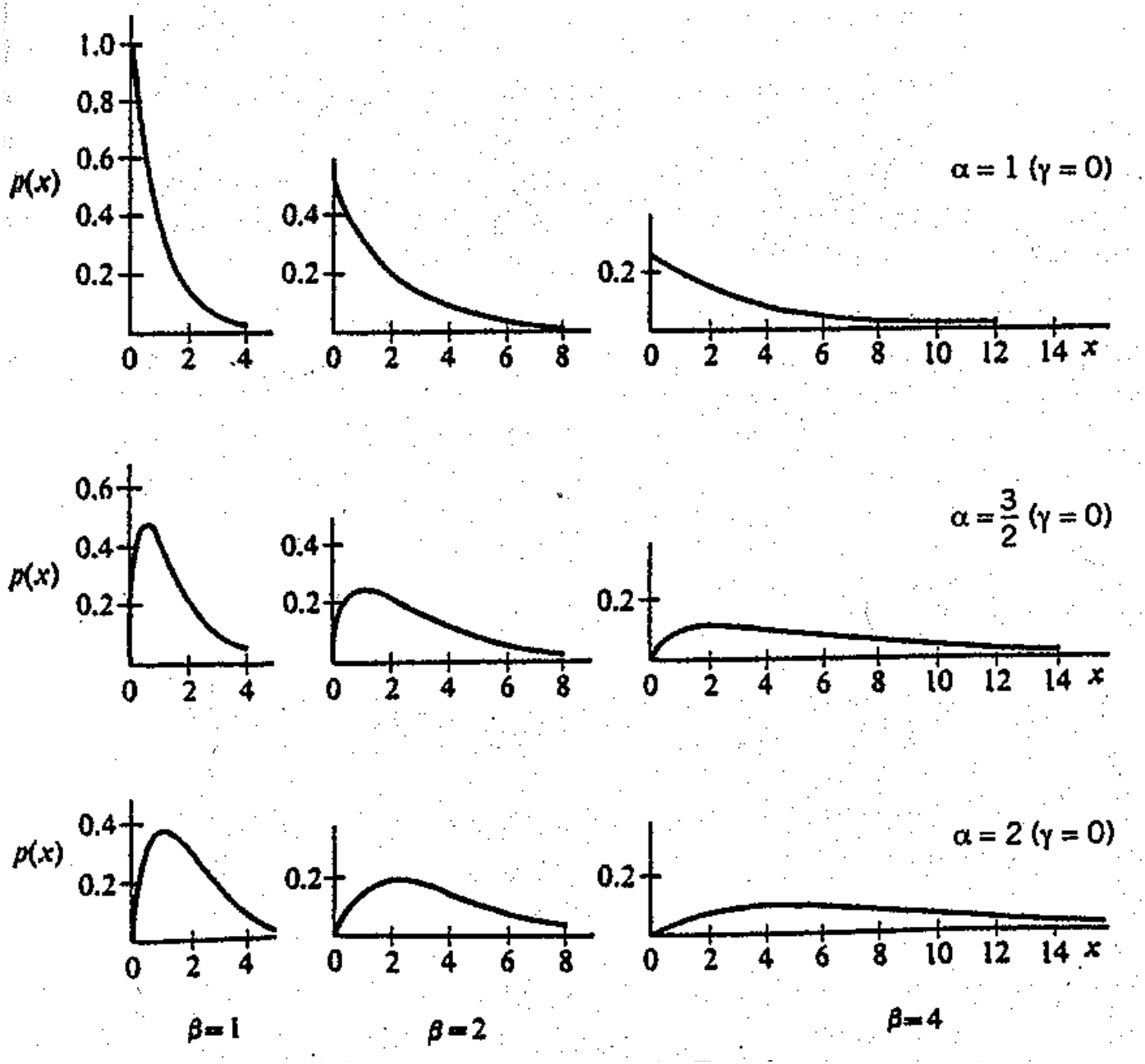
- Probability density function for $0 < x < \infty$:

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} \exp(-x/\beta)$$

  where $\alpha > 0$ (shape parameter) and $\beta > 0$ (scale parameter)
- Gamma function:

$$\Gamma(\alpha) \equiv \int_0^{\infty} x^{\alpha-1} \exp(-x) dx$$

- $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$ for $\alpha > 0$ and $\Gamma(n) = (n-1)!$ for $n \in \mathbb{N}$
- Mean $= \alpha\beta$ and Variance $= \alpha\beta^2$


p(x)
1.0
0.8
0.6
0.4
0.2
0 2 4
0.4
0.2
0 2 4 6 8
0.2
0 2 4 6 8 10 12 14 x
α = 1 (γ = 0)
p(x)
0.6
0.4
0.2
0 2 4
0.4
0.2
0 2 4 6 8
0.2
0 2 4 6 8 10 12 14 x
α = 3/2 (γ = 0)
p(x)
0.4
0.2
0 2 4
0.2
0 2 4 6 8
0.2
0 2 4 6 8 10 12 14 x
α = 2 (γ = 0)
β = 1
β = 2
β = 4

Negative-Binomial as an Overdispersed Poisson

- After some calculations,

$$\begin{aligned} p(Y_i \mid \mu_i, \theta) &= \frac{\mu_i^{Y_i}\theta^{\theta}\Gamma(Y_i+\theta)}{Y_i!\Gamma(\theta)(\mu_i+\theta)^{Y_i+\theta}} \\ &= \frac{(Y_i+\theta-1)!}{Y_i!(\theta-1)!}\left(\frac{\mu_i}{\mu_i+\theta}\right)^{Y_i}\left(\frac{\theta}{\mu_i+\theta}\right)^{\theta} \end{aligned}$$

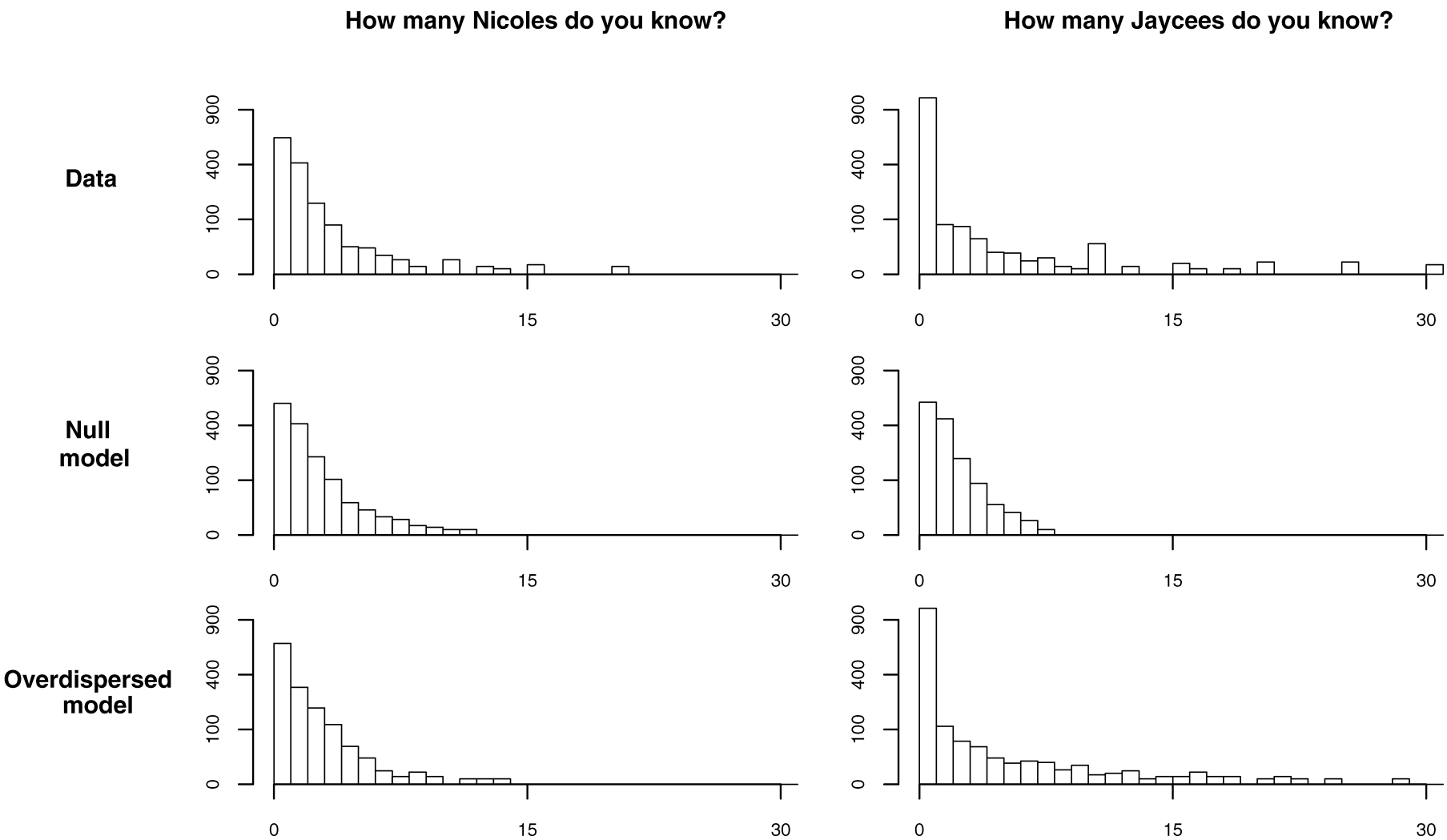
- Mean $= \mu_i$ and Variance $= \mu_i + \mu_i^2/\theta$ (overdispersed!)
- Log-likelihood function:

$$\begin{aligned} l_n(\beta, \theta \mid Y, X) = \sum_{i=1}^{n} \Big[& Y_i X_i^{\top}\beta - (Y_i+\theta)\log\{\exp(X_i^{\top}\beta)+\theta\} \\ & + \log\Gamma(Y_i+\theta)\Big] - n\theta\log\theta + n\log\Gamma(\theta) + \text{constant} \end{aligned}$$

- The model can arise from many forms of overdispersion, heterogeneity, etc.

How Many X's Do You Know?

- "Noisy" social network survey data (Zheng, et al., 2006 *JASA*):



Generalized Linear Models (GLMs)

- GLMs represent a systematic way to make inferences with commonly used regression models
- A GLM has 3 components
  1. Systematic component:
     - $\eta_i = X_i^\top \beta$
     - linear predictor
  2. Random component:
     - $f(Y; \theta, \phi)$
     - exponential family, (conditionally) independent across units
  3. Link component:
     - $g(\mu_i) = \eta_i$ with $\mu_i = \mathbb{E}(Y_i \mid X_i)$
     - monotonic and differentiable
- Advantages of GLM: flexibility and generalizability

Exponential Family of Distributions

- Exponential family:

$$f_Y(y; \theta, \phi) = \exp\{(y\theta - b(\theta))/a(\phi) + c(y, \phi)\}$$

1. $\mathcal{N}(\mu, \sigma^2)$: $\theta = \mu$ and $\phi = \sigma^2$ with $a(\phi) = \phi$, $b(\theta) = \theta^2/2$, and $c(y, \phi) = -\{y^2/\sigma^2 + \log(2\pi\sigma^2)\}/2$
2. Poisson(μ): $\theta = \log \mu$ and $\phi = 1$ with $a(\phi) = \phi$, $b(\theta) = \exp(\theta)$, and $c(y, \phi) = -\log y!$
3. Binomial(π, n)/n: $\theta = \log\{\pi/(1-\pi)\}$ and $\phi = n$ with $a(\phi) = 1/\phi$, $b(\theta) = \log\{1 + \exp(\theta)\}$, and $c(y, \phi) = \log \binom{n}{ny}$

- Mean and variance:

$$\mathbb{E}(Y) = b'(\theta) \quad \text{and} \quad \mathbb{V}(Y) = \underbrace{b''(\theta)}_{\textit{variance function}} a(\phi)$$

- A common form: $a(\phi) = \phi/\omega$ where ϕ is the dispersion parameter and ω is the prior weight varying across observations

Link Functions

- Relates the linear predictor $\eta_i = X_i^\top \beta$ with the mean μ_i
- Must map the real line onto the range of μ_i
- Canonical link functions $\eta = \theta$:
  - Normal: $\eta = \mu$
  - Poisson: $\eta = \log \mu$
  - binomial: $\eta = \log\{\pi/(1-\pi)\}$
- Monotonic and differentiable; $\mu_i = g^{-1}(\eta_i)$
- No reason to believe that canonical links work (just like we cannot believe in any model a priori!)

Likelihood Function, Score, and Information Matrix

- Log-likelihood function with $a(\phi) = \phi$:

$$l_n(\theta, \phi; Y) = \sum_{i=1}^{n} \left\{ \frac{Y_i \theta_i - b(\theta_i)}{\phi} + c(Y_i, \phi) \right\}$$

- Score statistic:

$$s_n(\beta) = \frac{\partial l_n(\theta, \phi; Y)}{\partial \beta} = \sum_{i=1}^{n} \frac{Y_i - b'(\theta_i)}{\phi} \frac{\partial \theta_i}{\partial \beta} = \sum_{i=1}^{n} (Y_i - \mu_i) w_i g'(\mu_i) X_i$$

  where $w_i = \frac{1}{\mathbb{V}(Y_i|X_i)} \frac{\partial \eta_i}{\partial \mu_i} = \frac{1}{\phi b''(\theta_i)\{g'(\mu_i)\}^2}$

- Expected Fisher information:

$$\Omega(\beta) = -\mathbb{E}\left(\frac{\partial^2 l_n(\theta, \phi; Y)}{\partial \beta \partial \beta^\top} \right) = \sum_{i=1}^{n} w_i X_i X_i^\top$$

- Canonical link (i.e., $\eta_i = \theta_i$): the expected information $=$ the observed information

Inverse Function Theorem

- **Theorem:** If $f(x)$ is a continuous and differentiable function and $f'(x) \neq 0$ for all x, then

$$(f^{-1})'(y) = \frac{1}{f'(x)}$$

  where $y = f(x)$

- Recall $\mu_i = b'(\theta_i)$, which implies $\theta_i = (b')^{-1}(\mu_i)$
- Applying the theorem yields

$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b''(\theta_i)}$$

- Similarly, $\eta_i = X_i^\top \beta = g(\mu_i)$, which implies $\mu_i = g^{-1}(\eta_i)$
- Applying the theorem yields

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{1}{g'(\mu_i)}$$

Iterated Weighted Least Squares Algorithm

- Recall the Fisher scoring algorithm from POL 572:

$$
\begin{aligned}
& \beta^{(t+1)} \\
= \; & \beta^{(t)} + \Omega(\beta^{(t)})^{-1} s_n(\beta^{(t)}) \\
= \; & \beta^{(t)} + \left(\sum_{i=1}^{n} w_i^{(t)} X_i X_i^\top \right)^{-1} \sum_{i=1}^{n} (Y_i - \mu_i^{(t)}) g'(\mu_i^{(t)}) w_i^{(t)} X_i \\
= \; & \left(\sum_{i=1}^{n} w_i^{(t)} X_i X_i^\top \right)^{-1} \sum_{i=1}^{n} w_i^{(t)} X_i \left\{ (Y_i - \mu_i^{(t)}) g'(\mu_i^{(t)}) + X_i^\top \beta^{(t)} \right\}
\end{aligned}
$$

- Weighted least squares!
  1. Calculate $Z_i^{(t)} = (Y_i - \mu_i^{(t)}) g'(\mu_i^{(t)}) + X_i^\top \beta^{(t)}$
  2. Regress $Z_i^{(t)}$ on X_i with the weight $w_i^{(t)}$ which gives $\beta^{(t+1)}$
  3. Repeat until convergence

Analysis of Deviance

- Null model (one parameter) $\Longleftrightarrow$ Saturated model (n parameters)
- Scaled deviance: Likelihood-ratio test statistic against the saturated model where $\tilde{\mu}_i = Y_i$ and $\tilde{\theta}_i = \theta(\tilde{\mu}_i) = \theta_i(Y_i)$

$$D^*(Y;\theta) \equiv -2\{l_n(\hat{\theta};Y) - l_n(\tilde{\theta};Y)\} = D(Y;\theta)/\phi$$

  where $D(Y;\theta) \equiv -2\sum_{i=1}^{n}[Y_i(\hat{\theta}_i - \tilde{\theta}_i) - \{b(\hat{\theta}_i) - b(\tilde{\theta}_i)\}]$ is the deviance
- Goodness-of-fit: If the model fits the data well, $D^*(Y;\theta) \overset{approx.}{\sim} \chi^2_{n-p}$
- Test nested models: $D^*_1 - D^*_2$ =LRT statistic
- Not useful as a general model selection algorithm since it depends on the sequence of models being tested
- A large literature on model/variable selection (e.g., Lasso)

Estimation of the Overdispersion Parameter

- Generalized Pearson χ^2 statistic:

$$\mathcal{X}^2 \equiv \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2}{b''(\hat{\theta}_i)}$$

- For normal distribution, $\mathcal{X}^2 = D = \sum_{i=1}^{n} (Y_i - \hat{\mu}_i)^2/\phi \sim \chi^2_{n-p}$
- Moment-based estimates of ϕ:

$$\hat{\phi}_{\mathcal{X}^2} = \frac{\mathcal{X}^2}{n-p} \quad \text{and} \quad \hat{\phi}_D = \frac{D}{n-p}$$

- For normal distribution, it corresponds to the MLE
- These estimates can be used to account for overdispersion in standard error calculation

Residuals

- Pearson residual:

$$\hat{\epsilon}_i^P \equiv \frac{Y_i - \hat{\mu}_i}{\sqrt{b''(\hat{\theta}_i)}}$$

- Often heavily skewed for non-Normal distribution
- Deviance residual:

$$\hat{\epsilon}_i^D \equiv \text{sgn}(Y_i - \hat{\mu}_i)\sqrt{d_i}$$

  where d_i is the deviance for the ith observation
- $\hat{\epsilon}_i^D$ is often similar to the Anscombe residual, which is distributed more closely to the normal distribution

Influential Observations and Leverage Points

- Recall the leverage points from linear regression

  $p_i \equiv X_i^\top (X^\top X)^{-1} X_i = \text{the } i\text{th diagonal element of } P_X = X(X^\top X)^{-1} X^\top$

- For GLM, we can use the weighted least squares

$$\begin{aligned} p_i &\equiv w_i X_i^\top (X^\top W X)^{-1} X_i \\ &= \text{the } i\text{th diagonal element of } W^{1/2} X (X^\top W X)^{-1} X^\top W^{1/2} \end{aligned}$$

  where $W = \text{diag}(w_i)$

- $0 \leq p_i \leq 1$
- $\sum_{i=1}^n p_i = \text{trace}(P_{W^{1/2}X}) = K$

- Asymptotically, $\mathbb{V}(Y_i - \hat{\mu}_i) \approx \phi b''(\theta_i)(1 - p_i)$
- Studentized Pearson residual: $\hat{\epsilon}_i^{P*} \equiv \hat{\epsilon}_i^P / \sqrt{\hat{\phi}(1 - p_i)}$
- Studentized Deviance residual: $\hat{\epsilon}_i^{D*} \equiv \hat{\epsilon}_i^D / \sqrt{\hat{\phi}(1 - p_i)}$

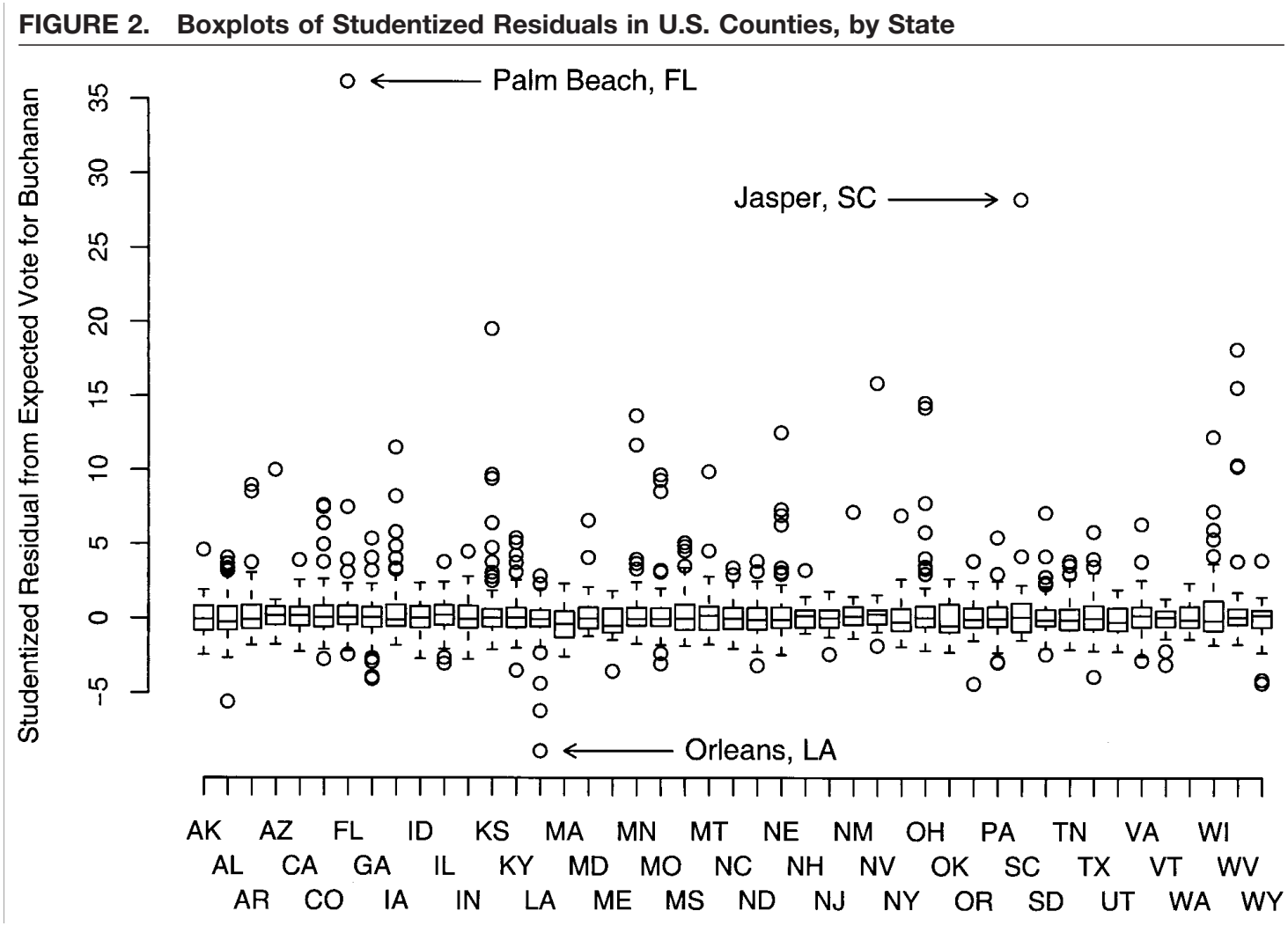
- Influential observations: what happens if we delete the ith observation?

$$\hat{\beta}_{(i)} \approx \hat{\beta} - \sqrt{w_i}(X^\top W X)^{-1} X_i \frac{\hat{\epsilon}_i^P}{(1 - p_i)}$$

- Cook's distance:

$$D_i \equiv \frac{(\hat{\beta}_{(i)} - \hat{\beta})(X^\top W X)^{-1}(\hat{\beta}_{(i)} - \hat{\beta})}{\hat{\phi} K}$$

Butterfly Ballot



Wand *et al.* (2001) *APSR*

Informal Graphical Model Checking

- Plot $\hat{\epsilon}_i^{D^*}$ against the normal quantile

- Plot $\hat{\epsilon}_i^{D^*}$ against $\hat{\eta}_i$
- Plot $\hat{\epsilon}_i^{D^*}$ against x_{ij}
- Plot $\hat{\epsilon}_i^{D^*}$ against an omitted variable
- Null pattern is mean zero and constant range

- Plot $|\hat{\epsilon}_i^{D^*}|$ against $\hat{\mu}_i$ to check the variance function
- Null pattern is no trend

- Plot the adjusted dependent variable Z_i against $\hat{\eta}_i$ to check the link function
- Null pattern is a straight line

Relaxing the Distributional Assumption

- How can we avoid the arbitrary choice of distribution?
  1. Systematic component: $\eta_i = X_i^\top \beta$
  2. Link fucntion: $\eta_i = g(\mu_i)$ where $\mu_i \equiv \mathbb{E}(Y_i \mid X)$
  3. Variance function: $\mathbb{V}(Y_i \mid X) = \phi\psi(\mu_i)$
- Only need to get the mean (and variance) right!
- Assume conditional independence (for now): $\mathrm{Cov}(Y_i, Y_j \mid X) = 0$
- Consider the following quasi-score statistic:

$$s_n^*(\beta; Y, X) \;=\; \sum_{i=1}^{n} \frac{Y_i - \mu_i}{\phi\psi(\mu_i)} \frac{\partial \mu_i}{\partial \beta} \;=\; D^\top \Psi^{-1}(Y - \mu)/\phi$$

  where D is a $n \times k$ matrix whose ith row is $\frac{\partial \mu_i}{\partial \beta^\top}$ and $\Psi = \mathrm{diag}(\psi(\mu_i))$
- Inherits two key properties of the GLM score statistic:

$$\mathbb{E}(s_n^* \mid X) \;=\; 0, \quad \mathbb{V}(s_n^* \mid X) \;=\; D^\top \Psi^{-1} D/\phi \;=\; -\mathbb{E}\left(\frac{\partial s_n^*}{\partial \beta^\top}\right)$$

Quasi-Likelihood Function

- The (log) quasi-likelihood function:

$$Q(\beta; Y, X) \equiv \sum_{i=1}^{n} \int_{Y_i}^{\mu_i} \frac{Y_i - u}{\phi\psi(u)} du$$

- The estimator $\hat{\beta}_Q$ solves for $s_n^*(\beta; Y, X) = 0$
- (Asymptotic) Sampling Theory (same as MLE!):

$$\sqrt{n}(\hat{\beta}_Q - \beta_0) \xrightarrow{D} \mathcal{N}\left(0,\ \phi\left(D^\top \Psi^{-1} D\right)^{-1}\right)$$

- Estimation of dispersion parameter:

$$\hat{\phi} = \frac{1}{n-k} \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2}{\psi(\hat{\mu}_i)} = \frac{\mathcal{X}^2}{n-k}$$

- Fisher scoring algorithm can be used

Dealing with Dependent Observations

- $\hat{\beta}_Q$ is still consistent
- Treat the lack of independence as a nuisance
- Familiar sandwitch estimator:

$$\mathbb{V}(\hat{\beta}_Q) \approx \frac{1}{n}\left(D^\top \Psi^{-1} D\right)^{-1}\left(D^\top \Psi^{-1} \Sigma_Y \Psi^{-1} D\right)\left(D^\top \Psi^{-1} D\right)^{-1}$$

where the (i, j) entry of the covariance matrix of Y can be estimated as

$$\widehat{\Sigma}_Y(i,j) = (Y_i - \hat{\mu}_i)(Y_j - \hat{\mu}_j)$$

- When independent, $\Sigma_Y = \text{diag}(\phi\psi(\mu_i))$
- $\hat{\beta}_Q$ is not efficient since it ignores dependence
- We will cover better alternatives later in the course

Concluding Remarks

- GLM generalizes many key concepts of linear regression to non-linear models in a unified framework
- Quasi-likelihood inference relaxes the distributional assumption
- Regression is not necessarily causal inference: omitted variables, post-treatment bias, etc.
- Regression can always give you descriptive or predictive inference
- Additional topics:
  1. Generalized Additive Models (GAMs)
  2. Variable selection