Poisson Regression

BIOS 509: Applied Linear Models

Jacob Englert
Department of Biostatistics & Bioinformatics
Rollins School of Public Health
Emory University

EMORY | ROLLINS SCHOOL OF PUBLIC HEALTH

10 April 2023

Outline

Outline

Poisson Regression Examples

We use Poisson regression when the outcome variable is a count. We can use the raw counts themselves or *rates*, which are simply the observed counts per some measure of population, space, time, etc.

- Number of grandchildren
- Number of asthma-related visits to an emergency room per day
- Number of drug overdoses per 100,000 people
- Number of students per square foot

We want to find helpful predictors to help explain these counts

- Number of children
- Amount/concentration of air pollutants, temperature
- Opioid prescribing rates
- Rollins building, department offering the course

Poisson Distribution

We say that $Y \sim \text{Poisson}(\mu), \mu > 0$, if Y takes on non-negative integer value y with probability

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

Some notes:

- Mean and variance are equal (both are μ)
- For small μ, the distribution of Y is right-skewed
- For large μ, $Y \dot{\sim}$ Normal

Linear Regression for Poisson Outcome?

We are interested in modeling μ as a function of predictor variables $\mathbf{X}$.

What if we use the linear regression model?

$$\mu = \mathrm{E}\left[Y \mid \mathbf{X}\right] = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

Problems:

- Does not restrict estimates or confidence intervals for μ to be positive.
- Violates the assumptions of normality and constant variance.

Generalized Linear Model (GLM)

1. Random Component:

$$Y_i \sim f_{Y_i}(y_i; \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\}$$

2. Systematic Component:

$$\eta_i = \sum_{j=1}^{p} \beta_j x_{ij} = \mathbf{X}_i^T \boldsymbol{\beta}$$

3. Link Function:

$$\eta_i = g(\mathrm{E}(Y_i))$$

   - Canonical Link: $g(\mathrm{E}(Y_i)) = \theta_i$

Poisson Log-Linear Model

Suppose $Y_i \sim \text{Poisson}(\mu_i)$. To identify the canonical link function for Poisson regression, write the PDF of $Y_i \sim \text{Poisson}(\mu_i)$ in its exponential family form:

$$f_{Y_i}(y_i; \mu_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} = \exp\{y_i \log \mu_i - \mu_i - \log y_i!\}$$

From here we can identify that the canonical link is $g(\mu_i) = \log \mu_i$. Does this make sense?

- $\mu_i \in (0, \infty) \Rightarrow g(\mu_i) \in (-\infty, \infty)$

With this link, we form the Poisson log-linear model:

$$\log \mu_i = \mathbf{X}_i^T \boldsymbol{\beta}$$

where $\mu_i = \text{E}[Y_i \mid \mathbf{X}_i]$

Estimation

As with any GLM, estimates of $\boldsymbol{\beta}$ are obtained by solving a system of score equations. When using the canonical link, these equations reduce to

$$U(\beta_j) = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{a(\phi)} = 0, \qquad j = 0, 1, \ldots, p$$

In Poisson regression, $a(\phi) = 1$ (constant) so we can ignore it.

The solution to the score equations, $\hat{\boldsymbol{\beta}}$, is obtained numerically:

- Newton-Raphson algorithm
- Fisher scoring algorithm
- Iteratively reweighted least squares (IRLS) - typically used in software

Interpretation: Response

Consider a single predictor model. Remember $\mu = \mathrm{E}[Y \mid X]$.

$$\log(\mathrm{E}[Y \mid X]) = \beta_0 + \beta_1 X$$

We can move back to the response scale by exponentiating both sides:

$$\mathrm{E}[Y \mid X] = e^{\beta_0 + \beta_1 X}$$

If $X = 0$, we have the expected response for the reference category:

$$\mathrm{E}[Y \mid X] = e^{\beta_0}$$

Interpretation: β_1

$$\log(\mathrm{E}[Y \mid X]) = \beta_0 + \beta_1 X$$

What happens on the log scale when we increase X by 1 unit?

$$\begin{aligned}\log(\mathrm{E}[Y \mid X = x+1]) &= \beta_0 + \beta_1(x+1) \\ &= \beta_0 + \beta_1 x + \beta_1 \\ &= \log(\mathrm{E}[Y \mid X = x]) + \beta_1\end{aligned}$$

Rearranging,

$$\beta_1 = \log(\mathrm{E}[Y \mid X = x+1]) - \log(\mathrm{E}[Y \mid X = x])$$

- β_1: change in the log expected count of Y for a one unit increase in X.

Interpretation: e^{β_1}

$$\log(\mathrm{E}[Y \mid X]) = \beta_0 + \beta_1 X$$

What happens on the response scale when we increase X by one unit?

$$\begin{aligned}
\mathrm{E}[Y \mid X = x+1] &= e^{\beta_0 + \beta_1 (x+1)} \\
&= e^{\beta_0 + \beta_1 x + \beta_1} \\
&= e^{\beta_0 + \beta_1 x} \times e^{\beta_1} \\
&= \mathrm{E}[Y \mid X = x] \times e^{\beta_1}
\end{aligned}$$

Rearranging,

$$e^{\beta_1} = \frac{\mathrm{E}[Y \mid X = x+1]}{\mathrm{E}[Y \mid X = x]}$$

- e^{β_1}: expected count of Y changes by a multiplicative factor of e^{β_1} for a one unit increase in X. e^{β_1} is often called a *risk ratio* or *relative risk* and is sometimes interpreted as a percentage increase/decrease.

Confidence Intervals

A $(100-\alpha)\%$ confidence interval for:

- $c\hat{\beta}_1$ (the effect of a c unit increase in X on the log expected count)

$$c\hat{\beta}_1 \pm z_{1-\alpha/2} \times c \times SE_{\hat{\beta}_1}$$

- $e^{c\hat{\beta}_1}$ (the effect of a c unit increase in X on the expected count)

$$\left(e^{\left[c\hat{\beta}_1 - z_{1-\alpha/2} \times c \times SE_{\hat{\beta}_1}\right]}, \quad e^{\left[c\hat{\beta}_1 + z_{1-\alpha/2} \times c \times SE_{\hat{\beta}_1}\right]} \right)$$

- $\log(\mathrm{E}[Y \mid \mathbf{X}])$ (the log expected count given some covariates)

$$\mathbf{X}\hat{\boldsymbol{\beta}} \pm z_{1-\alpha/2} \times SE_{\mathbf{X}\hat{\boldsymbol{\beta}}}$$

- $\mathrm{E}[Y \mid \mathbf{X}]$ (the expected count given some covariates)

$$\left(e^{\left[\mathbf{X}\hat{\boldsymbol{\beta}} - z_{1-\alpha/2} \times SE_{\mathbf{X}\hat{\boldsymbol{\beta}}}\right]}, \quad e^{\left[\mathbf{X}\hat{\boldsymbol{\beta}} + z_{1-\alpha/2} \times SE_{\mathbf{X}\hat{\boldsymbol{\beta}}}\right]} \right)$$

Including an Offset

So far we have been predicting a "risk", or rate per unit. In many cases, it makes sense to adjust the rate to something more informative. We do this by adding an offset predictor with a coefficient forced to be 1.

$$\log(\mathrm{E}\left[Y \mid X\right]) = \log(t) + \beta_0 + \beta_1 X \quad \Leftrightarrow \quad \mathrm{E}\left[Y \mid X\right] = t \times e^{\beta_0 + \beta_1 X}$$

Note that rearranging this is equivalent to

$$\log\left(\frac{\mathrm{E}\left[Y \mid X\right]}{t}\right) = \beta_0 + \beta_1 X$$

Now we are predicting the expected count of Y per t. You might see the regression coefficients referred to as *relative rates* or *rate ratios*.

It is appropriate to include an offset when you are interested in quantities such as cases per 1,000 people, crimes per year, or adverse events per person-year.

Poisson Regression Assumptions

The assumptions for Poisson regression differ from those for linear regression.

Linear Regression

- $Y \mid X \sim$ Normal
- Homoscedasticity
- Existence
- Independence
- Linearity: $\mathrm{E}[Y \mid X]$ is a linear function of the X's

Poisson Regression

- $Y \mid X \sim$ Poisson
- Mean and variance are equal
- Existence
- Independence
- Linearity: $\log(\mathrm{E}[Y \mid X])$ is a linear function of the X's

Example: Major League Soccer (MLS)

The MLS is the top soccer league in the United States. There are 28 teams, with each team having about 20-25 players.

One important metric of success is goal contributions. A player is credited with a goal contribution (GC) if he either a) scores a goal directly, or b) assists a goal-scorer (provides the final pass before the goal is scored).

In general, not a lot of goals are scored in soccer. So the number of GCs a player accumulated over the course of a season is relatively low. For this reason Poisson regression might perform better than linear regression.

Example: Major League Soccer (MLS)

The dataset we will be working with (*mls.csv*) contains data for players from every team during the 2022 MLS season. Some of the variables are:

- Player: player name
- Pos: position played (FW = Forward, MF = Midfielder, DF = Defender)
- MP: matches played (there are 34 matches in a season)
- Min: minutes played (a single match is 90 minutes)
- Wage: annual wage (in \$100,000s)
- Gls: number of goals scored
- Ast: number of assists provided

Research question:

- How is annual salary associated with number of goal contributions?

MLS Example: Univariate Analysis

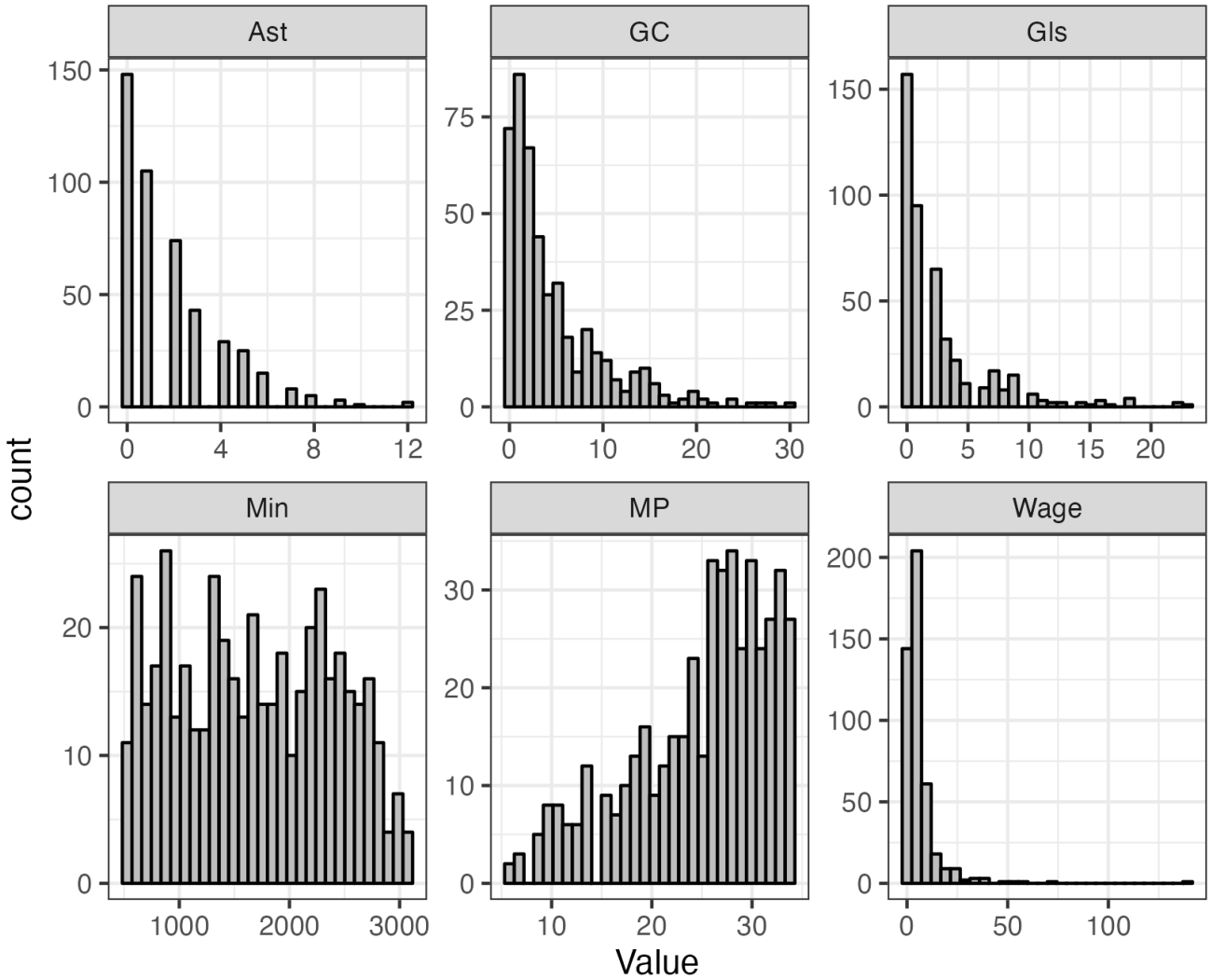

Figure: Histograms of Outcome and Predictor Variables

MLS Example: Data Pre-Processing

- Calculate goal contributions (GC) as goals (Gls) + assists (Ast)
- Restrict the sample to only include field players (no goalkeepers) with more than 500 minutes of play time.
- Log transform the skewed wage variable to improve efficiency.
- Use effective matches played (eMP) as an offset variable
  - Some players might only get 10 minutes per match, and it is not fair to compare their performance to players who play the entire game.
  - We can calculate effective matches played using

$$eMP = \frac{Min}{90}$$

  - We will denote the offset (expected) count as "(expected) goal contributions per 90 minutes", or"(x)GCp90)".

MLS Example: Goal Contributions

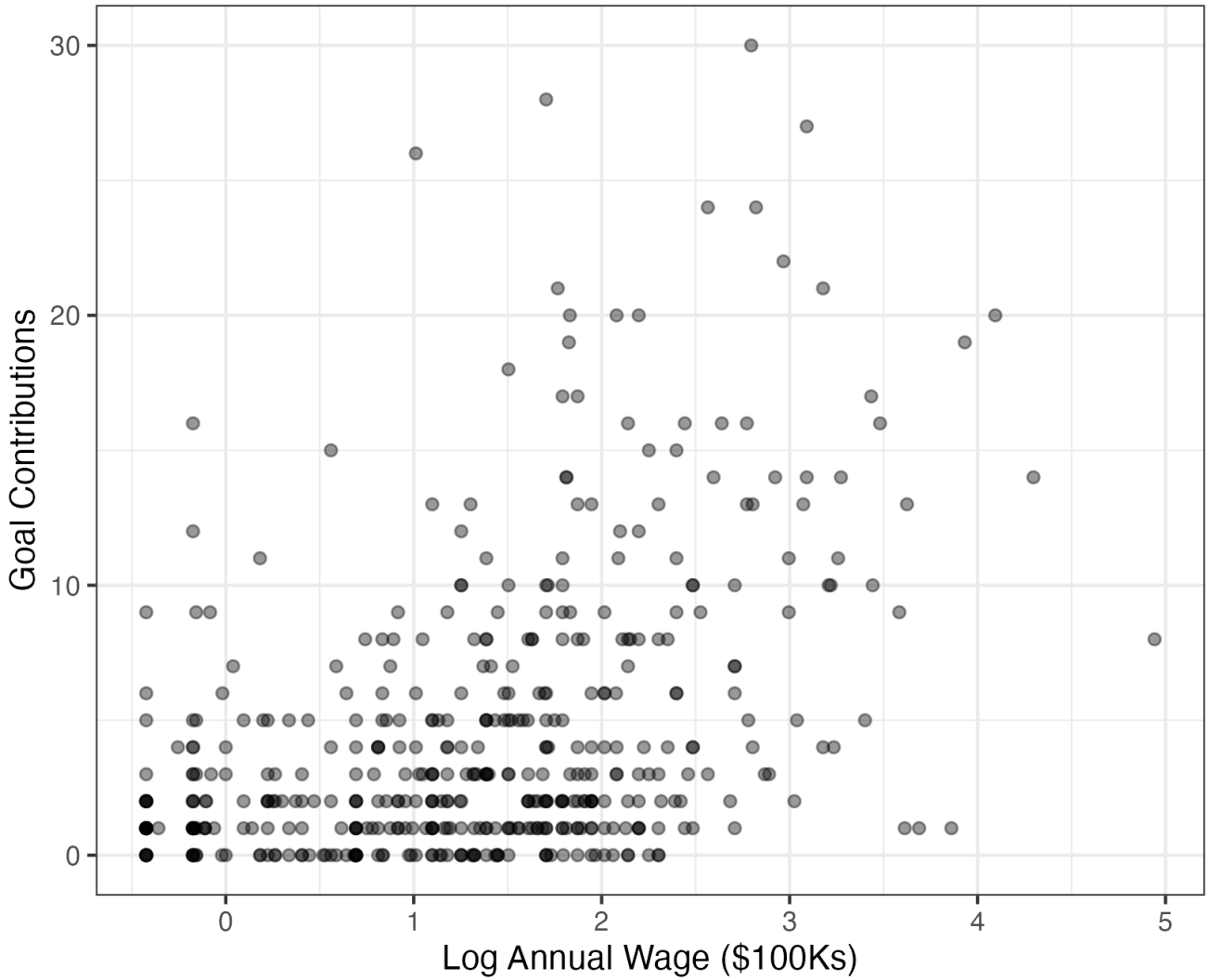

Figure: Scatter plot of Goal Contributions

MLS Example: Effective Matches Played

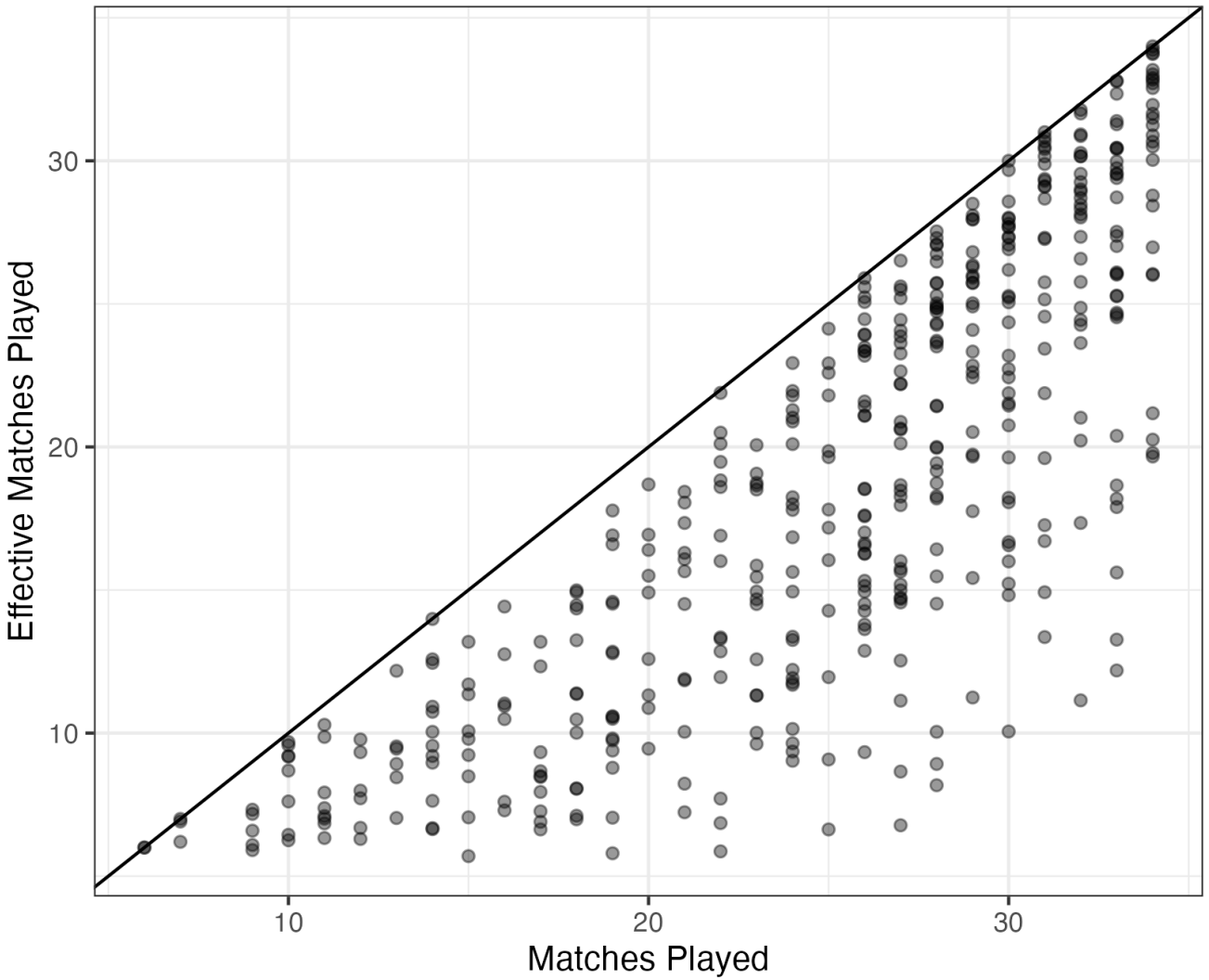

Figure: Matches Played vs. Effective Matches Played

MLS Example: Initial Results

We will start with the following model:

$$\log(\mathrm{E}\left[GC \mid \cdot\right]) = \log(eMP) + \beta_0 + \beta_1 lWage$$

	Log Scale			Response Scale	
	Estimate	SE	95% CI	Estimate	95% CI
β_0	-2.042	0.046	(-2.133, -1.951)	0.130	(0.119, 0.142)
β_1	0.383	0.022	(0.340, 0.427)	1.467	(1.405, 1.533)

Table: Parameter Estimates for Single Predictor Model

- Players earning \$100,000 ($lWage = 0$) have an xGCp90 of 0.13.
- For a 1-unit increase in log annual wage (in \$100,000s), xGCp90 increases by 46.7%.

MLS Example: Initial Results

We will start with the following model:

$$\log(\mathrm{E}\left[GC \mid \cdot\right]) = \log(eMP) + \beta_0 + \beta_1 lWage$$

	Log Scale			Response Scale	
	Estimate	SE	95% CI	Estimate	95% CI
β_0	-2.042	0.046	(-2.133, -1.951)	0.130	(0.119, 0.142)
β_1	0.383	0.022	(0.340, 0.427)	1.467	(1.405, 1.533)

Table: Parameter Estimates for Single Predictor Model

- For a 10% increase in a player's wage, xGCp90 increases by a factor of $\exp(\log(1.1) \times 0.383) = 1.037$.
- Players making \$5,000,000 ($lWage = 3.91$) have an xGCp90 of $\exp(-2.042 + 3.91 \times 0.383) = 0.58$.

Outline

The Saturated Model

For any regression model M, after finding the MLE $\hat{\boldsymbol{\beta}}$, we can obtain predictions for each observation as $\hat{\mu}_i = g^{-1}(\mathbf{X}_i^T \hat{\boldsymbol{\beta}})$, where $g(\cdot)$ is the link function.

The *saturated model* M_s is the model that perfectly fits the data

- $\tilde{\mu}_i = g^{-1}(\mathbf{X}_i^T \tilde{\boldsymbol{\beta}}) = y_i \quad \forall \quad i = 1, \ldots, n$
- The number of parameters equals the number of observations
- This model is overfit
- The observed data log-likelihood $l(\boldsymbol{\mu})$ is maximized by $\tilde{\boldsymbol{\mu}}$

Deviance

The *deviance* compares the fit of a given model M to the saturated model M_s.

The deviance of M with predictions $\hat{\mu}_i = g^{-1}(\mathbf{X}_i^T \hat{\boldsymbol{\beta}})$ is defined as:

$$\begin{aligned} D_M &\equiv -2\phi \left\{ l(\hat{\boldsymbol{\mu}}) - l(\tilde{\boldsymbol{\mu}}) \right\} \\ &= \frac{-2\phi}{a(\phi)} \sum_{i=1}^{N} \left\{ y_i \left(\tilde{\theta}_i - \hat{\theta}_i \right) - \left[b\left(\tilde{\theta}_i \right) - b\left(\hat{\theta}_i \right) \right] \right\} \end{aligned}$$

- In many cases, $a(\phi) \propto \phi$, and so D does not depend on ϕ

A related quantity is the *scaled deviance*

$$D_M^* = \frac{D_M}{\phi} = -2 \left\{ l(\hat{\boldsymbol{\mu}}) - l(\tilde{\boldsymbol{\mu}}) \right\}$$

- If $\phi = 1$ as in Poisson regression, $D_M = D_M^*$.
- This is equivalent to the likelihood ratio test statistic.

Scaled Deviance: Examples

Poisson Regression (log link)

$$
\begin{aligned}
D_M^* &= -2\left\{l(\hat{\boldsymbol{\mu}}) - l(\tilde{\boldsymbol{\mu}})\right\} \\
&= -2\sum_{i=1}^{n}\left\{\left[y_i \log(\hat{\mu}_i) - \hat{\mu}_i\right] - \left[y_i \log(y_i) - y_i\right]\right\} \\
&= -2\sum_{i=1}^{n}\left\{y_i\left[\log(\hat{\mu}_i) - \log(y_i)\right] - (\hat{\mu}_i - y_i)\right\} \\
&= -2\sum_{i=1}^{n}\left\{y_i \log\left(\frac{\hat{\mu}_i}{y_i}\right) - (\hat{\mu}_i - y_i)\right\}
\end{aligned}
$$

Other common examples:

- Normal: $D_M^* = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(y_i - \hat{\mu}_i\right)^2$
- Binomial: $D_M^* = -2\sum_{i=1}^{n}\left\{y_i \log\left(\frac{\hat{\mu}_i}{y_i}\right) + (m_i - y_i)\log\left(\frac{m_i - \hat{\mu}_i}{m_i - y_i}\right)\right\}$

Deviance Goodness of Fit Test

When ϕ is known, the scaled deviance has an asymptotic distribution

$$D_M^* \dot{\sim} \chi^2_{n-p-1}$$

where p is the number of predictors, and n is the number of observations.

Suppose we wish to test the fit of a model M:

$$H_0 : M \text{ fits the data well} \quad H_1 : M \text{ fits the data poorly}$$

$$D_M^* \sim \chi^2_{n-p-1} \quad \text{under } H_0$$

If $\Pr(\chi^2_{n-p-1} > D_M^*) < \alpha$, then there is evidence of a lack of fit.

Using Deviance for Model Comparison

The deviance GOF test is simply a comparison between the current and saturated models. We can just as easily compare any two nested models using a drop-in-deviance test, using the fact that the saturated model doesn't change.

Suppose we have fit two models, M_1 and M_2, where M_1 has the parameters $\beta_0, \ldots, \beta_{p_1}$ and M_2 has $\beta_0, \ldots, \beta_{p_1}, \beta_{p_1+1}, \ldots, \beta_{p_2}$. We can test the following:

$$H_0 : \beta_{p_1+1} = \ldots = \beta_{p_2} = 0 \quad H_1 : \text{ At least one of } \beta_{p_1+1}, \ldots, \beta_{p_2} \neq 0$$

$$\begin{aligned} D^*_{M_1} - D^*_{M_2} &= -2\left\{l(\hat{\boldsymbol{\mu}}_1) - l(\tilde{\boldsymbol{\mu}})\right\} + 2\left\{l(\hat{\boldsymbol{\mu}}_2) - l(\tilde{\boldsymbol{\mu}})\right\} \\ &= -2\left\{l(\hat{\boldsymbol{\mu}}_1) - l(\hat{\boldsymbol{\mu}}_2)\right\} \\ &\sim \chi^2_{p_2-p_1} \end{aligned}$$

We can reject H_0 if $\Pr(\chi^2_{p_2-p_1} > D^*_{M_1} - D^*_{M_2}) < \alpha$, and conclude that at least one of the additional terms in M_2 improves upon the fit of M_1.

Pearson Goodness of Fit Test

As an alternative to deviance, we can use the Pearson X^2 statistic:

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\widehat{\text{Var}}\,(Y_i)} \sim \chi^2_{n-p-1}$$

Some Examples:

- $Y_i \sim \text{Poisson}(\mu_i)$

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

- $Y_i \sim \text{Binomial}(m_i, p_i)$

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

Pearson vs. Deviance GOF Test

- Both the scaled deviance and Pearson X^2 have an asymptotic χ^2_{n-p-1} distribution, and the two are often similar in magnitude.
- If D^* or $X^2 > \chi^2_{n-p-1,1-\alpha}$, there is sufficient evidence at the α level that the candidate model has a poor fit
- Deviance can be used to compare nested models, while Pearson X^2 cannot
- In cases where ϕ is known, deviance-based tests are equivalent to likelihood ratio tests

MLS Example: Model Comparison

We have established that wage is an important predictor of goal contributions, but it is also worth considering position. For example, forwards are more likely to score than defenders. Consider the model:

$$\log(\mathrm{E}\left[GC \mid \cdot\right]) = \log(eMP) + \beta_0 + \beta_1 lWage + \beta_2 MF + \beta_3 FW$$

We should also consider the possibility of interaction between wage and position.

$$\log(\mathrm{E}\left[GC \mid \cdot\right]) = \log(eMP) + \beta_1 lWage + \beta_2 MF + \beta_3 FW + \beta_4 lWage \times MF + \beta_5 lWage \times FW$$

MLS Example: Model Comparison

Model	Residual DF	Residual Deviance
M_0: NULL	457	1884.82
M_1: Wage	456	1592.36
M_2: Wage + Pos	454	1003.27
M_3: Wage $\times$ Pos	452	984.27

Table: Poisson Model Comparison with Deviance

- Test for significance of position given wage is already in the model:

$$H_0 : \beta_2 = \beta_3 = 0 \quad H_1 : \text{ At least one of } \beta_2, \beta_3 \neq 0$$

$$D^*_{M_1} - D^*_{M_2} = 1592.36 - 1003.27 = 589.09 \sim \chi^2_2$$
$$\Rightarrow \Pr(\chi^2_2 > 589.09) < 0.0001$$

Conclusion: The addition of position improves the fit of a model already containing annual wage.

MLS Example: Goodness of Fit

- Test for significance of interaction between wage and position:

$$H_0 : \beta_4 = \beta_5 = 0 \quad H_1 : \text{ At least one of } \beta_4, \beta_5 \neq 0$$

$$D^*_{M_2} - D^*_{M_3} = 1003.27 - 984.27 = 19 \sim \chi^2_2$$
$$\Rightarrow \Pr(\chi^2_2 > 19) < 0.0001$$

  Conclusion: The association between player wage and xGCp90 depends on position.

So the interaction model fits best, but now let's check how good it fits:

$$H_0 : M_3 \text{ fits the data well} \quad H_1 : M_3 \text{ fits the data poorly}$$

$$D^*_{M_3} = 984.27 \Rightarrow \Pr(\chi^2_{452} > 984.27) < 0.0001$$

$$X^2_{M_3} = 972.20 \Rightarrow \Pr(\chi^2_{452} > 972.20) < 0.0001$$

Both tests suggest there is still significant lack of fit!

Outline

Overdispersion

When using Poisson regression we assume that the mean is equal to the variance. However in many real-life applications, count data have variance larger than the mean and are said to be *overdispersed*.

- *Underdispersion* also exists, but is far less common in practice.

The two most popular methods for modeling overdispersed count data are Quasi-Poisson and Negative Binomial regression.

If the mean model is correct but you do not account for overdispersion,

- The estimates of $\hat{\boldsymbol{\beta}}$ are still consistent
- The naive standard errors of $\hat{\boldsymbol{\beta}}$ are underestimated
  - Underestimated S.E. $\Rightarrow$ large test statistics / narrow C.I. $\Rightarrow$ small p-value $\Rightarrow$ wrong decisions

Outline

Quasi-Poisson Regression

$$\text{Assumption: } \text{Var}\,(Y_i) = \phi\mu_i$$

The Poisson score equations are replaced by "estimating" equations:

$$U(\boldsymbol{\beta}_j) = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\phi} = 0, \qquad j = 0, 1, \ldots, p$$

- $\hat{\boldsymbol{\beta}}$ will stay the same, since ϕ is still a constant.
- These are not "score" equations because they do not come from a true likelihood.

But how do we estimate ϕ?

Estimating $\hat{\phi}$

Recall the Pearson χ^2 test statistic:

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\widehat{\text{Var}}(Y_i)} = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\phi \hat{\mu}_i} \sim \chi^2_{n-p-1}$$

According to the expectation of χ^2 random variables, $\text{E}\left[X^2\right] = n - p - 1$

Rearranging,

$$\text{E}\left[\frac{1}{n-p-1} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}\right] = \phi$$

And so we can estimate ϕ with

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

Fitting a Quasi-Poisson Model

The previous slide suggests we can fit a Quasi-Poisson model with these steps:

1. Fit the original Poisson model (with no overdispersion)
2. Calculate the Pearson X^2 test statistic
3. Estimate $\hat{\phi} = \frac{X^2}{n-p-1}$
4. Compute the correct standard errors for $\hat{\boldsymbol{\beta}}$ as the square root of the diagonal elements of $\hat{\phi} \times \widehat{\text{Var}}(\hat{\boldsymbol{\beta}})$, where $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})$ is the estimated variance-covariance matrix obtained from step 1

Thankfully, most software has a shortcut to fit this model

- family = 'quasipoisson' in the glm() function in R
- 'scale = pearson' option in PROC GENMOD
- Since the standard errors need to be estimated, t- and F-distributions are used for testing regression coefficients and comparing nested models. This is analogous to the linear model setting, for which the MSE needs to be estimated.

Outline

Poisson-Gamma Mixture Model

Suppose we formulate the problem as a hierarchy such that:

$$Y \mid \mu \sim \text{Poisson}(\mu) \qquad \mu \sim \text{Gamma}\,(\alpha, \beta)$$

Thinking about the marginal distribution of Y...

1. What distribution does Y follow?
2. What is $\text{E}\,[Y]$?
3. What is $\text{Var}\,[Y]$?

Poisson-Gamma Mixture Model

$$Y \mid \mu \sim \text{Poisson}(\mu) \qquad \mu \sim \text{Gamma}\,(\alpha, \beta)$$

$$\begin{aligned}
f_Y(y) &= \int_\mu f_{Y|\mu}(y) \cdot f_\mu(\mu) d\mu \\
&= \int_0^\infty \frac{\mu^y e^{-\mu}}{y!} \cdot \frac{1}{\Gamma(\alpha)\beta^\alpha} \mu^{\alpha-1} e^{-\frac{\mu}{\beta}} d\mu \\
&= \frac{1}{y!\Gamma(\alpha)\beta^\alpha} \int_0^\infty \mu^{(y+\alpha)-1} e^{-\mu\left(\frac{\beta}{\beta+1}\right)^{-1}} d\mu \\
&= \frac{\Gamma(y+\alpha)\left(\frac{\beta}{\beta+1}\right)^{y+\alpha}}{\Gamma(y+1)\Gamma(\alpha)\beta^\alpha} \int_0^\infty \frac{1}{\Gamma(y+\alpha)\left(\frac{\beta}{\beta+1}\right)^{y+\alpha}} \mu^{(y+\alpha)-1} e^{-\mu\left(\frac{\beta}{\beta+1}\right)^{-1}} d\mu \\
&= \frac{\Gamma(y+\alpha)}{\Gamma(y+1)\Gamma(\alpha)} \left(\frac{1}{\beta+1}\right)^\alpha \left(1 - \frac{1}{\beta+1}\right)^y
\end{aligned}$$

Negative Binomial Model

The marginal distribution of Y is Negative Binomial!

$$Y \sim \text{NB}\left(y; r = \alpha, p = \frac{1}{\beta + 1}\right)$$

where

- y: number of failures in an experiment
- r: number of successes until experiment is stopped
- p: probability of success
- $\text{E}[Y] = \frac{r}{p}(1-p) = \alpha\beta = \mu$
- $\text{Var}[Y] = \frac{r}{p^2}(1-p) = \alpha\beta(1+\beta) = \alpha\beta + \alpha\beta^2 = \mu + \frac{\mu^2}{\alpha}$
  - α^{-1} is the "dispersion" parameter and must be estimated
- $\lim_{\alpha\to\infty} \text{Var}[Y] = \mu$
  - For large enough α, the NB model approaches the Poisson model

Some Notes on the Negative Binomial Model

If the data are overdispersed, but we don't believe $\text{Var}(Y) = \phi\mu$, we can use Negative Binomial regression as a more flexible approach.

- α is usually treated as unknown, and is simultaneously estimated with the regression coefficients.
- For known α, the canonical link is the log link. If you use this link you still have the relative risk interpretation.
- If α is unknown, then the Negative Binomial distribution cannot be represented as an exponential family.
  - If you intend to compare multiple NB model fits using deviance, you need to ensure α is the same between each of them.
  - There is no formal GOF test for NB models.
- The likelihood ratio test is still fair game for nested models.

Test for Overdispersion via Negative Binomial Model

A Poisson model can be viewed as a special case of the corresponding NB model where $\alpha^{-1} = 0$. This can be used to evaluate evidence of overdispersion.

Suppose we fit both a Poisson model M_P and NB model M_{NB} with the same predictors. We can conduct a formal test for overdispersion.

$$H_0 : \alpha^{-1} = 0 \text{ (no overdispersion)} \quad H_1 : \alpha^{-1} > 0 \text{ (overdispersion)}$$

$$X^2 = -2\left\{l_P(\hat{\boldsymbol{\mu}}_{M_P}) - l_{NB}(\hat{\boldsymbol{\mu}}_{M_{NB}})\right\} \sim \chi^2_1$$

If $2 \times \Pr(\chi^2_1 > X^2) < 0.05$, then there is evidence of overdispersion. Note this is a one-sided test, so the usual p-value is doubled.

MLS Example: Checking for Overdispersion

A quick and easy way to check for overdispersion is to compare $\hat{\phi}$ to 1, since this is its expected value under the Poisson assumptions. A general rule of thumb is that overdispersion exists if $\hat{\phi} > 1.5$.

In our case, $\frac{X^2_{M_3}}{452} = 2.15 > 1.5$, so there is some evidence of overdispersion.

Another helpful check is to split the data into multiple groups (usually defined using the explanatory variables), and then plot the means and variances of the outcome. In our case, suppose we split on player position and five equally-sized bins of the transformed wage variable.

MLS Example: Checking for Overdispersion

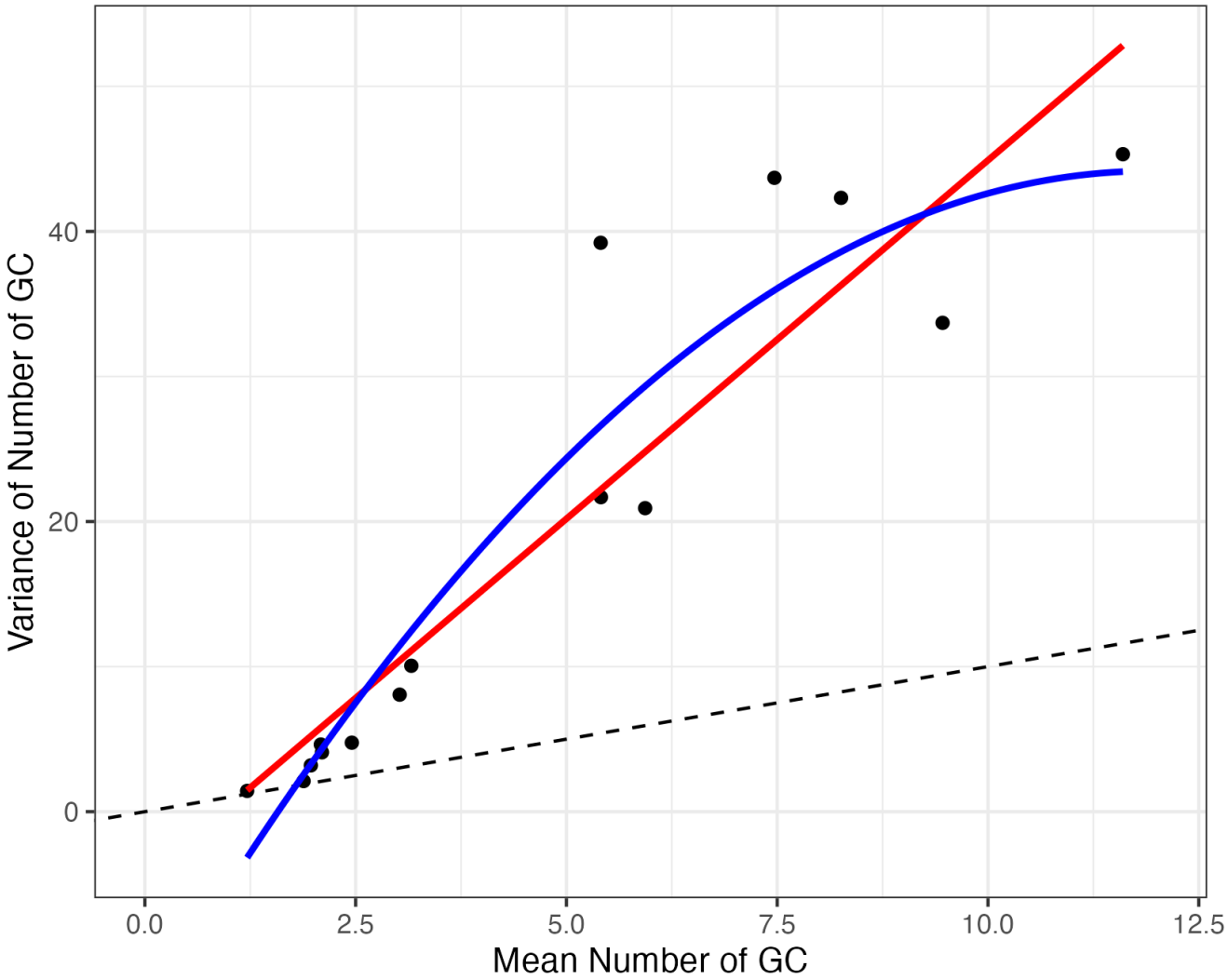

Figure: Evaluating Overdispersion in Observed Counts

MLS Example: Quasi-Poisson Regression

	Quasi-Poisson		Poisson	
	RR	95% CI	RR	95% CI
(Intercept)	0.095	(0.073, 0.125)	0.095	(0.079, 0.115)
lWage	1.093	(0.905, 1.319)	1.093	(0.961, 1.242)
MF	1.389	(0.967, 1.995)	1.389	(1.085, 1.778)
FW	4.001	(2.855, 5.609)	4.001	(3.178, 5.037)
lWage × MF	1.281	(1.030, 1.592)	1.281	(1.104, 1.486)
lWage × FW	1.056	(0.857, 1.300)	1.056	(0.916, 1.217)

Table: Parameter Estimates for Quasi-Poisson and Poisson Interaction Models

- The point estimates match, but the confidence intervals are wider for the Quasi-Poisson model.
- The midfielder (MF) main effect is no longer significant. Using the Quasi-Poisson model, we do not have enough evidence to suggest xGCp90 differs between midfielders and defenders who are making \$100,000.

MLS Example: Negative Binomial Regression

	Quasi-Poisson		Poisson		Negative Binomial	
	RR	95% CI	RR	95% CI	RR	95% CI
(Intercept)	0.10	(0.07, 0.13)	0.10	(0.08, 0.12)	0.09	(0.07, 0.12)
lWage	1.09	(0.91, 1.32)	1.09	(0.96, 1.24)	1.12	(0.95, 1.31)
MF	1.39	(0.97, 2.00)	1.39	(1.09, 1.78)	1.51	(1.10, 2.07)
FW	4.00	(2.86, 5.61)	4.00	(3.18, 5.04)	3.99	(2.89, 5.52)
lWage × MF	1.28	(1.03, 1.59)	1.28	(1.10, 1.49)	1.21	(0.99, 1.47)
lWage × FW	1.06	(0.86, 1.30)	1.06	(0.92, 1.22)	1.03	(0.85, 1.25)

Table: Parameter Estimates for All Interaction Models

- The estimates from the NB model are similar in both direction and magnitude.
- The confidence intervals are narrower than the Quasi-Poisson model but still wider than the Poisson model.

MLS Example: Negative Binomial Regression

	Quasi-Poisson		Poisson		Negative Binomial	
	RR	95% CI	RR	95% CI	RR	95% CI
(Intercept)	0.10	(0.07, 0.13)	0.10	(0.08, 0.12)	0.09	(0.07, 0.12)
lWage	1.09	(0.91, 1.32)	1.09	(0.96, 1.24)	1.12	(0.95, 1.31)
MF	1.39	(0.97, 2.00)	1.39	(1.09, 1.78)	1.51	(1.10, 2.07)
FW	4.00	(2.86, 5.61)	4.00	(3.18, 5.04)	3.99	(2.89, 5.52)
lWage × MF	1.28	(1.03, 1.59)	1.28	(1.10, 1.49)	1.21	(0.99, 1.47)
lWage × FW	1.06	(0.86, 1.30)	1.06	(0.92, 1.22)	1.03	(0.85, 1.25)

Table: Parameter Estimates for All Interaction Models

- While it is very close, neither interaction term is on its own significant in the NB model.
- A likelihood ratio test comparing the NB model with interaction to the NB model without interaction suggests interaction is not significant ($X_2^2 = 5.51, p = 0.0635$).

MLS Example: Test for Overdispersion via NB

Let us compare the Poisson model M_P and NB model M_{NB}, both with interaction. We can conduct a formal test for overdispersion.

$$H_0 : \alpha^{-1} = 0 \text{ (no overdispersion)} \quad H_1 : \alpha^{-1} > 0 \text{ (overdispersion)}$$

$$\begin{aligned} X^2 &= -2\left\{l_P(\hat{\boldsymbol{\mu}}_{M_P}) - l_{NB}(\hat{\boldsymbol{\mu}}_{M_{NB}})\right\} \\ &= 161.93 \\ &\Rightarrow 2 \times \Pr(\chi^2_1 > 161.93) < 0.0001 \end{aligned}$$

Thus, there is evidence of overdispersion.

MLS Example: Model Comparison

One way to compare unnested likelihood-based models is using Akaike's Information Criterion (AIC). The AIC is calculated as:

$$2p - 2l(\hat{\boldsymbol{\mu}})$$

where p is the number of parameters in the model.

Here is a table comparing the likelihood-based models (linear model included):

Family	Model	AIC
Gaussian	Main Effects Only	2644.8
Gaussian	Interaction	2636.3
Poisson	Main Effects Only	2234.4
Poisson	Interaction	2219.4
Negative Binomial	Main Effects Only	2061.0
Negative Binomial	Interaction	2059.5

Table: Likelihood-Based Model Comparison

Lower values are better, so both NB models are the best by this criteria.

MLS Example: Final Model Interpretation

	RR	95% CI
(Intercept)	0.083	(0.071, 0.097)
lWage	1.221	(1.136, 1.311)
MF	1.983	(1.665, 2.362)
FW	3.984	(3.336, 4.758)

Table: Parameter Estimates for Negative Binomial Model (without interaction)

- A defender earning \$100,000 has an xGCp90 of 0.083 (95% CI: 0.071, 0.097).
- For a 10% increase in annual wage, xGCp90 increases by a factor of $\exp(\log(1.221) \times \log(1.1)) = 1.019$, holding position fixed.
- Midfielders have 98.3% more xGCp90 compared to defenders who earn the same wage (95% CI: 66.5%, 136.2%).
- Forwards have an estimated 3.984 times the xGCp90 compared to defenders who earn the same wage (95% CI: 3.336, 4.758).

Outline

Residuals

- Pearson Residuals

$$e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{\text{Var}}\left(Y_i\right)}}$$

Note that the Pearson residual for observation i is simply the square root of the i^{th} component of the Pearson X^2 statistic.

- Deviance Residuals

$$e_i = \sqrt{d_i} \times \text{sign}(y_i - \hat{\mu}_i)$$

Where d_i is the i^{th} component of the deviance such that $D = \sum_i d_i$

Standardized Residuals

- Both the Pearson and deviance residuals can be standardized by dividing by their asymptotic standard errors:

$$e_i^* = \frac{e_i}{\sqrt{1 - \hat{h}_i}}$$

  Here $\hat{h}_i$ is the estimated leverage for the i^{th} observation.
- The standardized versions should follow a standard normal distribution.
- In R the standardized residuals can be obtained with rstandard(). In PROC GENMOD they can be obtained using STDRESCHI and STDRESDEV in the OUTPUT statement.
- DFBETAs, DFFITs, and Cook's D can all be calculated for GLMs as well.

MLS Example: Examining Model Fit

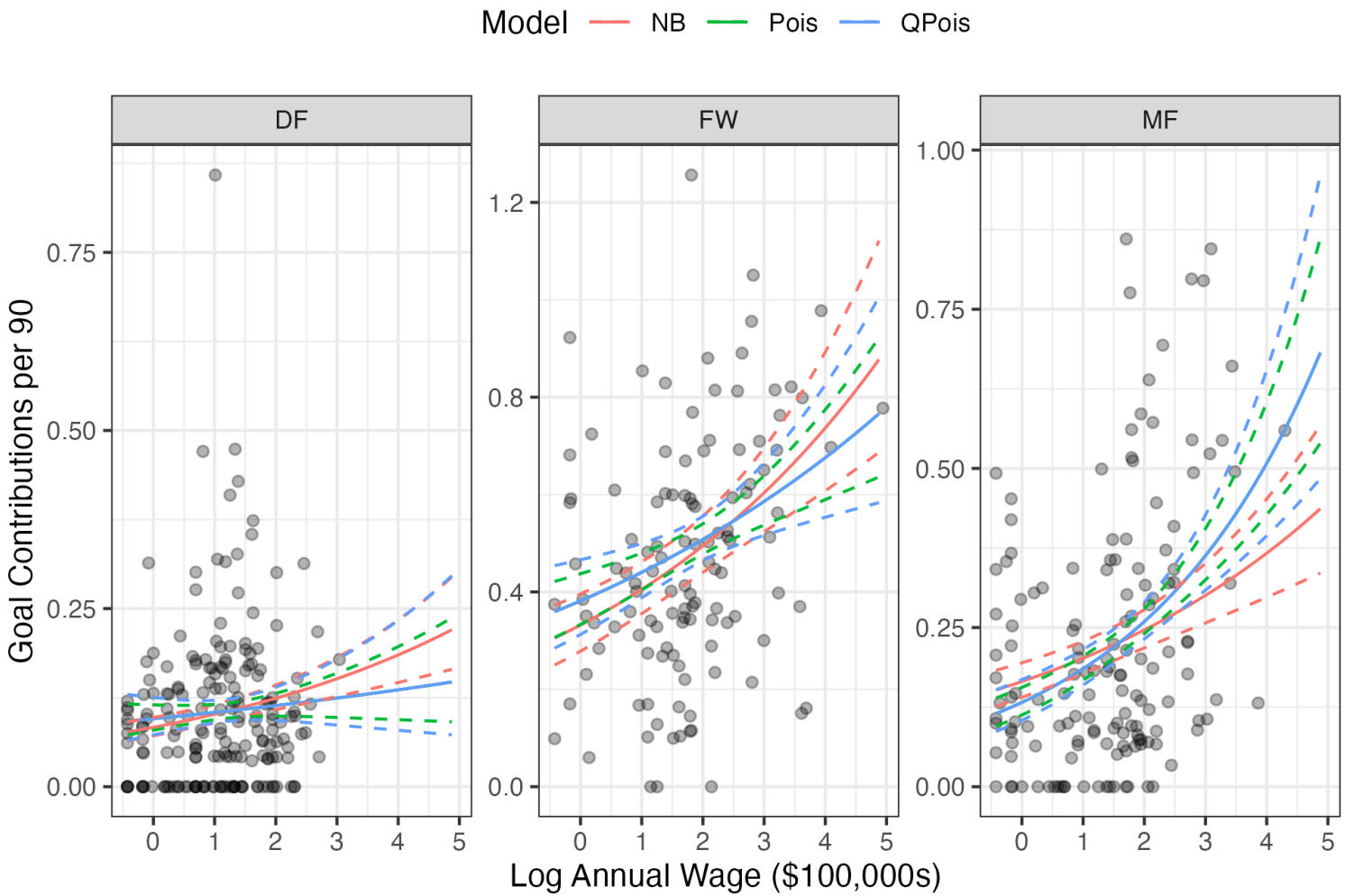

Figure: Examining Model Fit to Observed Data

MLS Example: Checking Model Assumptions

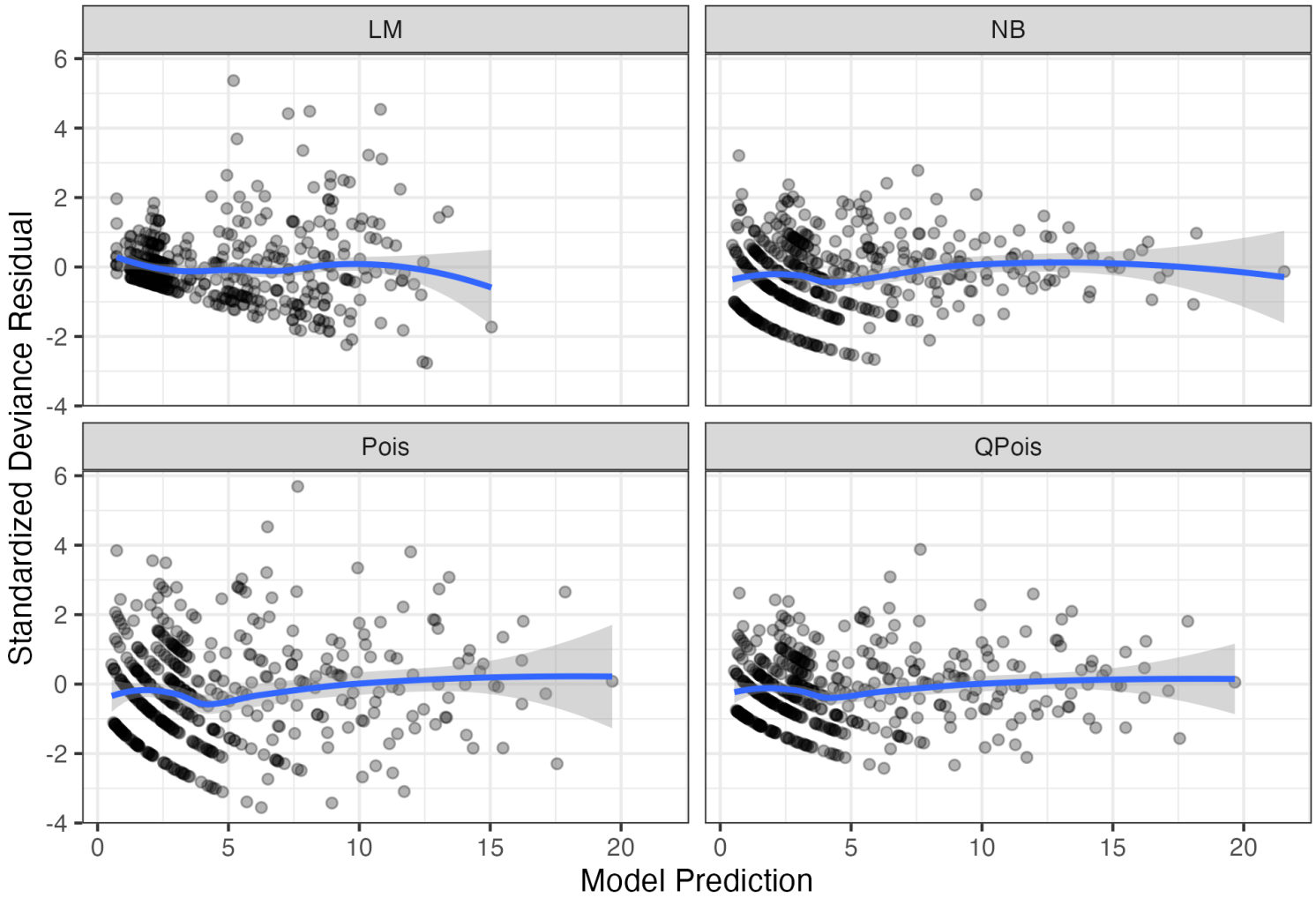

Figure: Evaluating Linearity on the Link Function Scale

MLS Example: Checking Model Assumptions

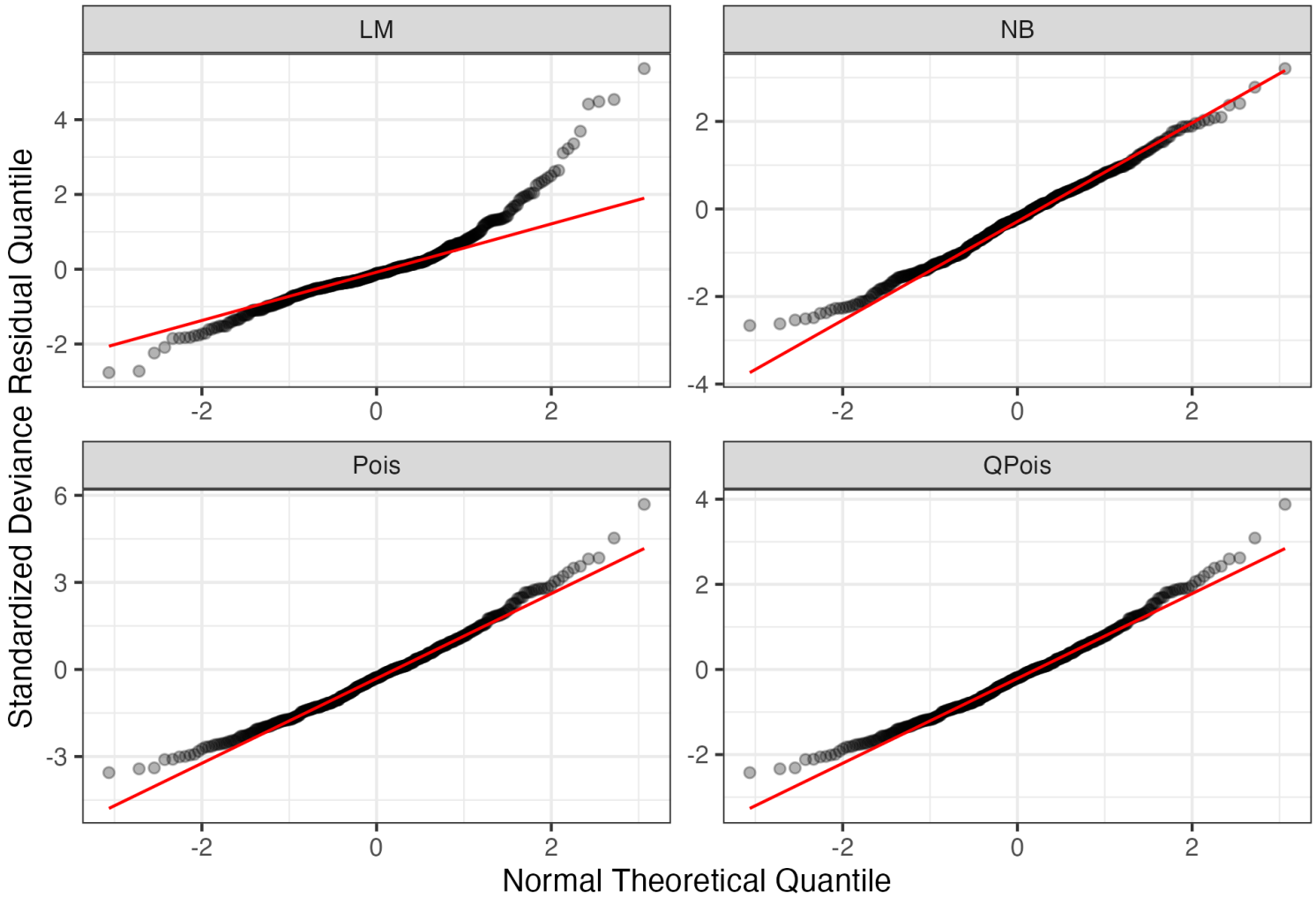

Figure: Evaluating Normality of Residuals

MLS Example: Checking Model Assumptions

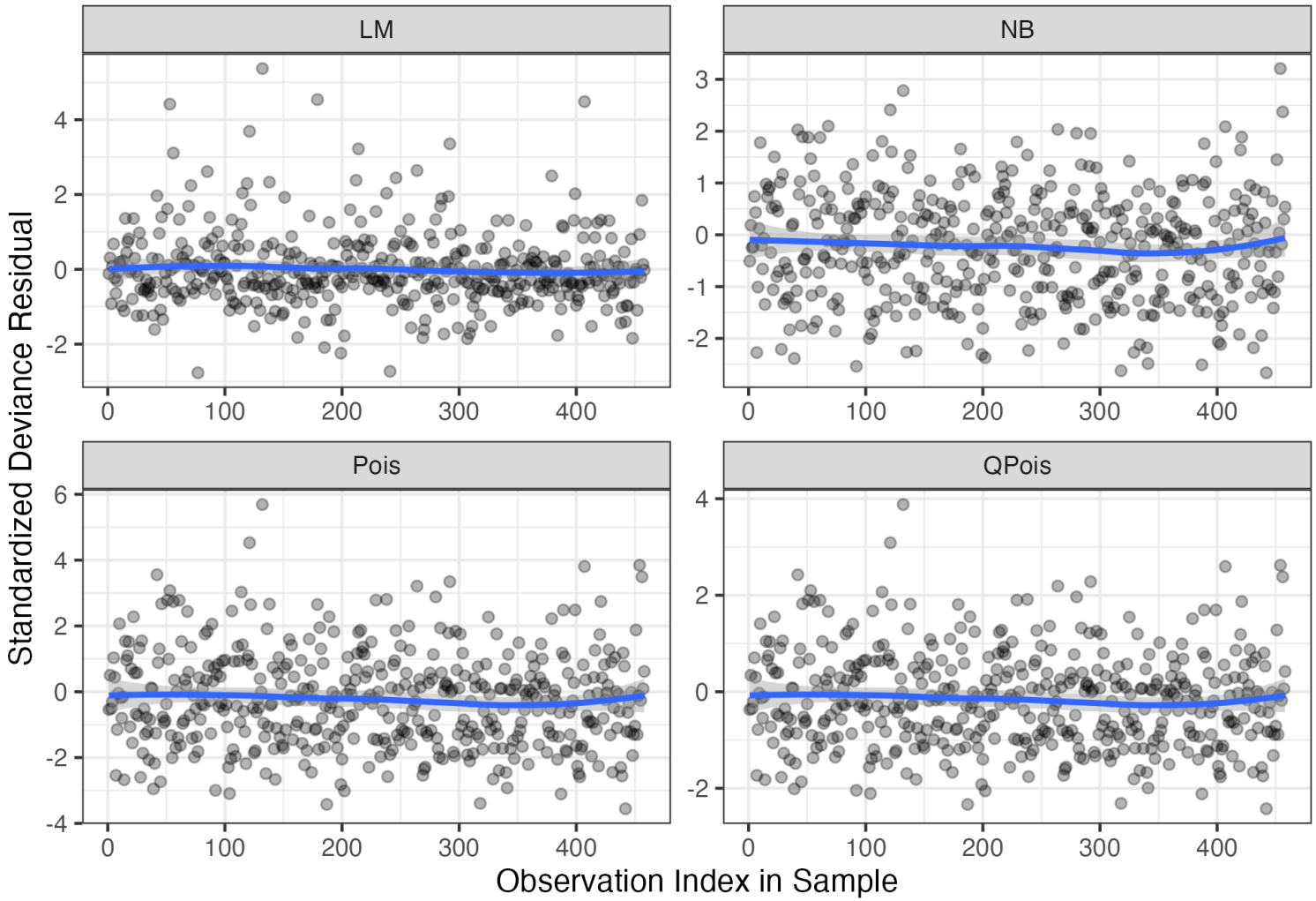

Figure: Evaluating Independence of Observations

Helpful Resources

- Generalized Linear Models with Examples in R (Dunn and Smyth 2018)
- Beyond Multiple Linear Regression (Roblack and Legler 2021)
- Generalized Linear Models: Residuals and Diagnostics (Horvath 2019)
- Notes for Predictive Modeling (García-Portugués 2023)
- Modeling Count Data (PSU)