Unit 5: Generalized linear models: Special cases

Eugene Katsevich

November 29, 2021

Unit 4 developed a general theory for GLMs. In Unit 5, we specialize this theory to several important cases, including logistic regression and Poisson regression.

1 Logistic regression

1.1 Model definition and interpretation

Model definition. Recall from Unit 4 that the logistic regression model is

$$m_i y_i \overset{\text{ind}}{\sim} \text{Bin}(m_i, \pi_i); \quad \text{logit}(\pi_i) = \log \frac{\pi_i}{1-\pi_i} = \boldsymbol{x}_{i*}^T \boldsymbol{\beta}. \tag{1}$$

Here we use the canonical logit link function, although other link functions are possible. The interpretation of the parameter β_j is that a unit increase in x_j—other predictors held constant—is associated with an (additive) increase of β_j on the log-odds scale or a multiplicative increase of e^{β_j} on the odds scale. Note that logistic regression data come in two formats: *ungrouped* and *grouped*. For ungrouped data, we have $m_1 = \cdots = m_n = 1$, so $y_i \in \{0, 1\}$ are Bernoulli random variables. For grouped data, we can have several independent Bernoulli observations per predictor $\boldsymbol{x}_{i*}$, which give rise to binomial proportions $y_i \in [0, 1]$. This happens most often when all the predictors are discrete. You can always convert grouped data into ungrouped data, but not necessarily vice versa. We'll discuss below that the grouped and ungrouped formulations of logistic regression have the same MLE and standard errors but different deviances.

Generative model equivalent. Consider the following generative model for $(\boldsymbol{x}, y) \in \mathbb{R}^{p-1} \times \{0, 1\}$:

$$y \sim \text{Ber}(\pi); \quad \boldsymbol{x}|y \sim \begin{cases} N(\boldsymbol{\mu}_0, \boldsymbol{V}) & \text{if } y = 0 \\ N(\boldsymbol{\mu}_1, \boldsymbol{V}) & \text{if } y = 1 \end{cases}. \tag{2}$$

Then, we can derive that $y|\boldsymbol{x}$ follows a logistic regression model (called a *discriminative* model because it conditions on $\boldsymbol{x}$). Indeed,

$$\begin{aligned} \text{logit}(p(y=1|\boldsymbol{x})) &= \log \frac{p(y=1)p(\boldsymbol{x}|y=1)}{p(y=0)p(\boldsymbol{x}|y=0)} \\ &= \log \frac{\pi \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T \boldsymbol{V}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)\right)}{(1-\pi)\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_0)^T \boldsymbol{V}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_0)\right)} \\ &= \beta_0 + \boldsymbol{x}^T \boldsymbol{V}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \\ &\equiv \beta_0 + \boldsymbol{x}^T \boldsymbol{\beta}_{\text{-}0}. \end{aligned} \tag{3}$$

This is another natural route to motivating the logistic regression model.

Special case: 2×2 contingency table. Suppose that $x \in \{0, 1\}$, and consider the logistic regression model $\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i$. For example, suppose that $x \in \{0, 1\}$ encodes treatment (1) and control (0) in a clinical trial, and $y_i \in \{0, 1\}$ encodes success (1) and failure (0). We make n observations of (x_i, y_i) in this ungrouped setup. The parameter e^{β_1} can be interpreted as the *odds ratio*:

$$e^{\beta_1} = \frac{\mathbb{P}[y = 1|x = 1]/\mathbb{P}[y = 0|x = 1]}{\mathbb{P}[y = 1|x = 0]/\mathbb{P}[y = 0|x = 0]}. \tag{4}$$

This parameter is the multiple by which the odds of success increase when going from control to treatment. We can summarize such data via the 2×2 *contingency table* (Table 1). A grouped version of this data would be $\{(x_1, y_1) = (0, 7/24), (x_2, y_2) = (1, 9/21)\}$. The null hypothesis $H_0 : \beta_1 = 0 \iff H_0 : e^{\beta_1} = 1$ states that the success probability in both rows of the table is the same.

	Success	Failure	Total
Treatment	9	12	21
Control	7	17	24
Total	16	29	45

Table 1: An example of a 2×2 contingency table.

Logistic regression with case-control studies. In a prospective study (e.g. a clinical trial), we assign treatment or control (i.e., x) to individuals, and then observe a binary outcome (i.e., y). Sometimes, the outcome y takes a long time to measure or has highly imbalanced distribution in the population (e.g. the development of lung cancer). In this case, an appealing study design is the *retrospective study*, where individuals are sampled based on their *response values* (e.g. presence of lung cancer) rather than their treatment/exposure status (e.g. smoking). It turns out that a logistic regression model is appropriate for such retrospective study designs as well. Indeed, suppose that $y|\boldsymbol{x}$ follows a logistic regression model. Let's try to figure out the distribution of $y|\boldsymbol{x}$ in the retrospectively gathered sample. Letting $z \in \{0, 1\}$ denote the indicator that an observation is sampled, define $\rho_1 \equiv \mathbb{P}[z = 1|y = 1]$ and $\rho_0 \equiv \mathbb{P}[z = 1|y = 0]$, and assume that $\mathbb{P}[z = 1, y, \boldsymbol{x}] = \mathbb{P}[z = 1|y]$. The latter assumption states that the predictors $\boldsymbol{x}$ were not used in the retrospective sampling process. Then,

$$\text{logit}(\mathbb{P}[y = 1|z = 1, \boldsymbol{x}]) = \log \frac{\rho_1 \mathbb{P}[y = 1|\boldsymbol{x}]}{\rho_0 \mathbb{P}[y = 0|\boldsymbol{x}]} = \log \frac{\rho_1}{\rho_0} + \text{logit}(\mathbb{P}[y = 1|\boldsymbol{x}]) = \left(\log \frac{\rho_1}{\rho_0} + \beta_0\right) + \boldsymbol{x}^T \boldsymbol{\beta}_{\text{-}0}.$$

Thus, conditioning on retrospective sampling changes only the intercept term, but preserves the coefficients of $\boldsymbol{x}$. Therefore, we can carry out inference for $\boldsymbol{\beta}_{\text{-}0}$ in the same way regardless of whether the study design is prospective or retrospective.

1.2 Estimation and inference

Score and Fisher information. We recall from Unit 4 that the score is

$$\frac{\partial}{\partial \boldsymbol{\beta}} \log \mathcal{L}(\boldsymbol{\beta}) = \boldsymbol{X}^T \boldsymbol{D} \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) = \boldsymbol{X}^T \text{diag}\left(\frac{\partial \mu_i / \partial \eta_i}{\text{Var}[y_i]}\right) (\boldsymbol{y} - \boldsymbol{\mu}). \tag{5}$$

Note that

$$\frac{\partial \mu_i / \partial \eta_i}{\text{Var}[y_i]} = \frac{\partial \mu_i / \partial \theta_i}{\text{Var}[y_i]} = \frac{\ddot{\psi}(\theta_i)}{\text{Var}[y_i]} = m_i. \tag{6}$$

Therefore, the score equations are

$$0 = \boldsymbol{X}^T \text{diag}\,(m_i)\,(\boldsymbol{y} - \widehat{\boldsymbol{\mu}}) \quad \Longleftrightarrow \quad \sum_{i=1}^{n} m_i x_{ij}(y_i - \widehat{\pi}_i) = 0, \quad j = 0, \ldots, p-1. \tag{7}$$

We can solve these equations using IRLS. The Fisher information is

$$\boldsymbol{I}(\boldsymbol{\beta}) = \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}, \quad W_{ii} = \frac{(\partial \mu_i / \partial \eta_i)^2}{\text{Var}[y_i]} = \frac{\ddot{\psi}(\theta_i)^2}{\text{Var}[y_i]} = m_i^2 \text{Var}[y_i] = m_i \pi_i (1 - \pi_i). \tag{8}$$

Wald inference. Using the results in the previous paragraph, we can carry out Wald inference based on the normal approximation

$$\widehat{\boldsymbol{\beta}} \,\dot{\sim}\, N\left(\boldsymbol{\beta}, \left(\boldsymbol{X}^T \text{diag}(m_i \widehat{\pi}_i (1 - \widehat{\pi}_i)) \boldsymbol{X}\right)^{-1}\right). \tag{9}$$

This approximation holds for $\sum_{i=1}^{n} m_i \to \infty$. Unfortunately, Wald inference in finite samples does not always perform very well. The Wald test above is known to be conservative due to the *Hauck-Donner effect*. As an example, consider testing $H_0 : \beta_0 = 0.5$ in the intercept-only model

$$ny \sim \text{Bin}(n, \pi); \quad \text{logit}(\pi) = \beta_0. \tag{10}$$

The Wald test statistic is $z \equiv \widehat{\beta}/\text{SE} = \text{logit}(y)\sqrt{ny(1-y)}$. This test statistic actually tends to *decrease* as $y \to 1$, since the standard error grows faster than the estimate itself. For example, take $n = 25$. Then, $z = 3.3$ for $n = 23/25$ but $z = 3.1$ for $n = 24/25$. So the test statistic becomes less significant as we go further away from the null!

Perfect separability. If we have a situation where a hyperplane in covariate space separates observations with $y_i = 0$ from those with $y_i = 1$, we have *perfect separability*. It turns out that some of the maximum likelihood estimates are infinite in this case. The Wald test completely fails in this case, since it uses the parameter estimates as test statistics.

Likelihood ratio inference. Let's first compute the deviance of a logistic regression model. We have

$$L(\boldsymbol{y}; \boldsymbol{\pi}) = \sum_{i=1}^{n} m_i y_i \log \pi_i + m_i (1 - y_i) \log(1 - \pi_i), \tag{11}$$

so

$$D(\boldsymbol{y}; \widehat{\boldsymbol{\pi}}) = 2(L(\boldsymbol{y}; \boldsymbol{y}) - L(\boldsymbol{y}; \widehat{\boldsymbol{\pi}})) = 2 \sum_{i=1}^{n} \left(m_i y_i \log \frac{y_i}{\widehat{\pi}_i} + m_i (1 - y_i) \log \frac{1 - y_i}{1 - \widehat{\pi}_i} \right). \tag{12}$$

Letting $\widehat{\boldsymbol{\pi}}_0$ and $\widehat{\boldsymbol{\pi}}_1$ be the MLEs from two nested models, we can then express the likelihood ratio statistic as

$$T^{\text{LRT}} = 2(L(\boldsymbol{y}; \widehat{\boldsymbol{\pi}}_1) - L(\boldsymbol{y}; \widehat{\boldsymbol{\pi}}_0)) = 2 \sum_{i=1}^{n} \left(m_i y_i \log \frac{\widehat{\pi}_{i1}}{\widehat{\pi}_{i0}} + m_i (1 - y_i) \log \frac{1 - \widehat{\pi}_{i1}}{1 - \widehat{\pi}_{i0}} \right). \tag{13}$$

We can then construct a likelihood ratio test in the usual way. Likelihood ratio inference can give nontrivial conclusions in cases when Wald inference cannot, e.g. in the case of perfect separability. Indeed, suppose that

$$m_i y_i \sim \text{Bin}(m_i, \pi_i), \quad \text{logit}(\pi_i) = \beta_0 + \beta_1 x_i, \quad i = 1, 2. \tag{14}$$

We would like to test $H_0 : \beta_1 = 0$. Suppose that we observe $(x_1, y_1) = (0, 0)$, $(x_2, y_2) = (1, 1)$, giving us complete separability. Can we still get a meaningful test of H_0? We can write out the likelihood ratio test statistic, which is

$$D(\boldsymbol{y}; \widehat{\boldsymbol{\pi}}) = 2\left(m_1 \log \frac{1}{1 - \frac{m_2}{m_1+m_2}} + m_2 \log \frac{1}{\frac{m_2}{m_1+m_2}}\right) = 2\left(m_1 \log \frac{m_1+m_2}{m_1} + m_2 \log \frac{m_1+m_2}{m_2}\right).$$

This is a number that we can compare to the χ^2_1 distribution to get a p-value, as usual.

Goodness of fit tests. We can test goodness of fit in the grouped logistic regression model by comparing the deviance statistic (12) to the asymptotic null distribution χ^2_{n-p}. We can alternatively use the score test, which gives us Pearson's X^2 statistic:

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \widehat{\pi}_i)^2}{\widehat{\pi}_i(1 - \widehat{\pi}_i)/m_i}. \tag{15}$$

Fisher's exact test. As an alternative to asymptotic tests for logistic regression, in the case of 2×2 tables there is an *exact* test of $H_0 : \beta_1 = 0$. Suppose we have

$$s_1 = m_1 y_1 \sim \text{Bin}(m_1, \pi_1) \quad \textit{and} \quad s_2 = m_2 y_2 \sim \text{Bin}(m_2, \pi_2). \tag{16}$$

The trick is to conduct inference *conditional on* $s_1 + s_2$. Note that under $H_0 : \pi_1 = \pi_2$, we have

$$\begin{aligned}
\mathbb{P}[s_1 = t | s_1 + s_2 = v] &= \mathbb{P}[s_1 = t | s_1 + s_2 = v] \\
&= \frac{\mathbb{P}[s_1 = t, s_2 = v - t]}{\mathbb{P}[s_1 + s_2 = v]} \\
&= \frac{\binom{m_1}{t}\pi^t(1-\pi)^{m_1-t}\binom{m_2}{v-t}\pi^{v-t}(1-\pi)^{m_2-(v-t)}}{\binom{m_1+m_2}{v}\pi^v(1-\pi)^{m_1+m_2-v}} \\
&= \frac{\binom{m_1}{t}\binom{m_2}{v-t}}{\binom{m_1+m_2}{v}}.
\end{aligned} \tag{17}$$

Therefore, a finite-sample p-value to test $H_0 : \pi_1 = \pi_2$ versus $H_1 : \pi_1 > \pi_2$ is $\mathbb{P}[s_1 \geq t | s_1 + s_2]$, which can be computed exactly based on the formula above.

2 Poisson regression

The Poisson regression model (with offsets) is

$$y_i \overset{\text{ind}}{\sim} \text{Poi}(\mu_i); \quad \log \mu_i = o_i + \boldsymbol{x}_{i*}^T \boldsymbol{\beta}. \tag{18}$$

Because the log of the mean is linear in the predictors, Poisson regression models are often called *loglinear models*. We have seen in Unit 4 how to carry out inference for this model based on the Wald, likelihood ratio, and score tests. Recall, for example, that the deviance of this model is

$$D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} y_i \log \frac{y_i}{\widehat{\mu}_i}. \tag{19}$$

2.1 Modeling rates

One cool feature of the Poisson model is that rates can be easily modeled with the help of offsets. Let's say that the count y_i is collected over the course of a time interval of length t_i, or a spatial region with area t_i, or a population of size t_i, etc. Then, it is meaningful to model

$$y_i \overset{\text{ind}}{\sim} \text{Poi}(\pi_i t_i), \quad \log \pi_i = \boldsymbol{x}_{i*}^T \boldsymbol{\beta}, \tag{20}$$

where π_i represents the rate of events per day / per square mile / per capita, etc. In other words,

$$y_i \overset{\text{ind}}{\sim} \text{Poi}(\mu_i), \quad \log \mu_i = \log t_i + \boldsymbol{x}_{i*}^T \boldsymbol{\beta}, \tag{21}$$

which is exactly equation (18) with offsets $o_i = \log t_i$. For example, in single cell RNA-sequencing, y_i is the number of reads aligning to a gene in cell i and t_i is the total number of reads measured in the cell, a quantity called the *sequencing depth*. We might use a Poisson regression model to carry out a *differential expression analysis* between two cell types.

2.2 Relationship between Poisson and multinomial distributions

Suppose that $y_i \overset{\text{ind}}{\sim} \text{Poi}(\mu_i)$ for $i = 1, \ldots, n$. Then,

$$\begin{aligned} \mathbb{P}\left[y_1 = m_1, \ldots, y_n = m_n \,\middle|\, \sum_i y_i = m \right] &= \frac{\mathbb{P}[y_1 = m_1, \ldots, y_n = m_n]}{\mathbb{P}[\sum_i y_i = m]} \\ &= \frac{\prod_{i=1}^n e^{-\mu_i} \frac{\mu_i^{y_i}}{y_i!}}{e^{-\sum_i \mu_i} \frac{(\sum_i \mu_i)^m}{m!}} \\ &= \binom{m}{m_1, \ldots, m_n} \prod_{i=1}^n \pi_i^{y_i}; \quad \pi_i \equiv \frac{\mu_i}{\sum_{i'=1}^n \mu_{i'}}. \end{aligned} \tag{22}$$

We recognize the last expression as the probability mass function of the multinomial distribution with parameters $(\pi_1, \ldots, \pi_n)$ summing to one. In words, the joint distribution of a set of independent Poisson distributions conditional on their sum is a multinomial distribution.

2.3 Poisson model for 2×2 contingency tables

Let's say that we have two binary random variables $x_1, x_2 \in \{0, 1\}$ with joint distribution $\mathbb{P}(x_1 = j, x_2 = k) = \pi_{jk}$ for $j, k \in \{0, 1\}$. We collect a total of n samples from this joint distribution and summarize the counts in a 2×2 table, where y_{jk} is the number of times we observed $(x_1, x_2) = (j, k)$, so that

$$(y_{00}, y_{01}, y_{10}, y_{11}) | n \sim \text{Mult}(n, (\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})). \tag{23}$$

Our primary question is whether these two random variables are independent, i.e.

$$\pi_{jk} = \pi_{j+}\pi_{+k}, \quad \text{where} \quad \pi_{j+} \equiv \mathbb{P}[x_1 = j] = \pi_{j1} + \pi_{j2}; \quad \pi_{+k} \equiv \mathbb{P}[x_2 = k] = \pi_{1k} + \pi_{2k}. \tag{24}$$

We can express this equivalently as

$$\pi_{00}(\pi_{00} + \pi_{01} + \pi_{10} + \pi_{11}) = \pi_{00} = \pi_{0+}\pi_{+0} = (\pi_{00} + \pi_{01})(\pi_{00} + \pi_{10}) \quad \Longleftrightarrow \quad \pi_{00}\pi_{11} = \pi_{01}\pi_{10}. \tag{25}$$

In other words, we can express the independence hypothesis concisely as

$$H_0 : \frac{\pi_{11}\pi_{00}}{\pi_{10}\pi_{01}} = 1. \tag{26}$$

Let's arbitrarily assume that, additionally, $n \sim \text{Poi}(\mu_{++})$. Then,

$$(y_{00}, y_{01}, y_{10}, y_{11}) \sim \text{Poi}(\mu_{++}\pi_{00}) \times \text{Poi}(\mu_{++}\pi_{01}) \times \text{Poi}(\mu_{++}\pi_{10}) \times \text{Poi}(\mu_{++}\pi_{11}). \tag{27}$$

Let $i \in 1, 2, 3, 4$ index the four pairs $(x_1, x_2) \in \{(0,0), (0,1), (1,0), (1,1)\}$, so that

$$y_i \overset{\text{ind}}{\sim} \text{Poi}(\mu_i); \quad \log \mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2}, \quad i = 1, \ldots, 4, \tag{28}$$

where

$$\beta_0 = \log \mu_{++} + \log \pi_{00}; \quad \beta_1 = \log \frac{\pi_{10}}{\pi_{00}}; \quad \beta_2 = \log \frac{\pi_{01}}{\pi_{00}}; \quad \beta_{12} = \log \frac{\pi_{11}\pi_{00}}{\pi_{10}\pi_{01}}. \tag{29}$$

Note that the independence hypothesis (26) reduces to the hypothesis $H_0 : \beta_{12} = 0$:

$$H_0 : \frac{\pi_{11}\pi_{00}}{\pi_{10}\pi_{01}} = 1 \quad \iff \quad H_0 : \beta_{12} = 0. \tag{30}$$

So the presence of an interaction in the Poisson regression is equivalent to a lack of independence between x_1 and x_2. We can test the latter hypothesis using our standard tools for Poisson regression. For example, we can use the Pearson X^2 goodness of fit test. To apply this test, we must find the fitted means under the null hypothesis. The normal equations state that the observed cell counts equal those that would have been expected under the null hypothesis. Using the formulation (24), we obtain

$$y_{jk} = \mathbb{E}[y_{jk}] = \widehat{\mu}_{++}\widehat{\pi}_{j+}\widehat{\pi}_{+k}, \tag{31}$$

so that

$$\widehat{\mu} = y_{++}; \quad \widehat{\mu}_{++}\widehat{\pi}_{j+} = y_{j+}; \quad \widehat{\mu}_{++}\widehat{\pi}_{+k} = y_{+k}, \tag{32}$$

from which it follows that

$$\widehat{\mu}_{jk} = \widehat{\mu}_{++}\widehat{\pi}_{j+}\widehat{\pi}_{+k} = y_{++}\frac{y_{j+}}{y_{++}}\frac{y_{+k}}{y_{++}} = \frac{y_{j+}y_{+k}}{y_{++}}. \tag{33}$$

Hence, we have

$$X^2 = \sum_{j,k=0}^{1} \frac{(y_{jk} - \widehat{\mu}_{jk})^2}{\widehat{\mu}_{jk}}. \tag{34}$$

Alternatively, we can use the likelihood ratio test, which gives

$$G^2 = \sum_{j,k=0}^{1} y_{jk} \log \frac{y_{jk}}{\widehat{\mu}_{jk}}. \tag{35}$$

2.4 Inference is the same regardless of conditioning on margins

Now, our data might actually have been collected such that $n \sim \text{Poi}(\mu)$, or maybe n was fixed in advance. Is the Poisson inference proposed above actually valid in the latter case? In fact, it is! To see this, we claim that the likelihood ratio statistic is the same for the Poisson and multinomial models. Indeed, let's write the Poisson likelihood as follows:

$$p_{\boldsymbol{\mu}}(\boldsymbol{y}) = p_{\mu_{++}}(y_{++} = n) p_{\boldsymbol{\pi}}(\boldsymbol{y} | y_{++} = n). \tag{36}$$

Note that the fitted parameter $\widehat{\mu}_{++}$ is the same under the null and alternative hypotheses: $\widehat{\mu}^0_{++} = \widehat{\mu}^1_{++}$, so we have

$$\frac{p_{\widehat{\boldsymbol{\mu}}^1}(\boldsymbol{y})}{p_{\widehat{\boldsymbol{\mu}}^0}(\boldsymbol{y})} = \frac{p_{\widehat{\mu}^1_{++}}(y_{++} = n) p_{\widehat{\boldsymbol{\pi}}^1}(\boldsymbol{y} | y_{++} = n)}{p_{\widehat{\mu}^0_{++}}(y_{++} = n) p_{\widehat{\boldsymbol{\pi}}^0}(\boldsymbol{y} | y_{++} = n)} = \frac{p_{\widehat{\boldsymbol{\pi}}^1}(\boldsymbol{y} | y_{++} = n)}{p_{\widehat{\boldsymbol{\pi}}^0}(\boldsymbol{y} | y_{++} = n)}. \tag{37}$$

The latter expression is the likelihood ratio statistic for the multinomial model. The same argument shows that conditioning on the row or column totals (as opposed to the overall total) also yields the same exact inference. Therefore, regardless of the sampling mechanism, we can always conduct an independence test in a 2×2 table via a Poisson regression.

2.5 Equivalence among Poisson and logistic regressions

We've talked above two ways to view a 2×2 contingency table. In the logistic regression view, we thought about one variable as a predictor and the other as a response, seeking to test whether the predictor has an impact on the response. In the Poisson regression view, we thought about the two variables symmetrically, seeking to test independence. It turns out that these two perspectives are equivalent. Note that under the Poisson model, we have

$$\text{logit } \mathbb{P}[x_2 = 1 | x_1 = 0] = \log \frac{\pi_{01}}{\pi_{00}} = \beta_2 \tag{38}$$

and

$$\text{logit } \mathbb{P}[x_2 = 1 | x_1 = 1] = \log \frac{\pi_{11}}{\pi_{10}} = \log \frac{\pi_{01}}{\pi_{00}} + \log \frac{\pi_{11}\pi_{00}}{\pi_{10}\pi_{01}} = \beta_2 + \beta_{12}. \tag{39}$$

In other words,

$$\text{logit } \mathbb{P}[x_2 = 1 | x_1] = \beta_2 + \beta_{12} x_1. \tag{40}$$

Therefore, the β_{12} parameter for the Poisson regression (28) is the same as it is for the logistic regression (40).

2.6 Poisson models for $J \times K$ contingency tables

Suppose now that $x_1 \in \{1, \dots, J\}$ and $x_2 \in \{1, \dots, K\}$. Then, we denote $\mathbb{P}[x_1 = j, x_2 = k] = \pi_{jk}$. We still are interested in testing for independence between j and k, which amounts to a goodness-of-fit test for the Poisson model

$$y_{jk} \overset{\text{ind}}{\sim} \text{Poi}(\mu_{jk}); \quad \log \mu_{jk} = \beta_0 + \beta_j^1 + \beta_k^2. \tag{41}$$

The Pearson statistic for this test is

$$\sum_{j=1}^{J} \sum_{k=1}^{K} \frac{(y_{ij} - \widehat{\mu}_{ij})^2}{\widehat{\mu}_{ij}}; \quad \widehat{\mu}_{ij} = \widehat{y}_{++} \frac{y_{i+}}{y_{++}} \frac{y_{+j}}{y_{++}}. \tag{42}$$

Like with the 2×2 case, the test is the same regardless if we condition on the row sums, column sums, total count, or if we do not condition at all. The degrees of freedom in the full model is JK, while the degrees of freedom in the partial model is $J + K - 1$, so the degrees of freedom for the goodness-of-fit test is $JK - J - K + 1 = (J - 1)(K - 1)$. Pearson erroneously concluded that the test had $JK - 1$ degrees of freedom, which when Fisher corrected created a lot of animosity between these two statisticians.

2.7 Poisson models for $J \times K \times L$ contingency tables

These ideas can be extended to multi-way tables, for example three-way tables. If we have $x_1 \in \{1, \dots, J\}, x_2 \in \{1, \dots, K\}, x_3 \in \{1, \dots, L\}$, then we might be interested in testing several kinds of null hypotheses:

- Mutual independence: $H_0 : x_1 \perp\!\!\!\perp x_2 \perp\!\!\!\perp x_3$.

- Joint independence: $H_0 : x_1 \perp\!\!\!\perp (x_2, x_3)$.
- Conditional independence: $H_0 : x_1 \perp\!\!\!\perp x_2 \mid x_3$.

These three null hypotheses can be shown to be equivalent to the Poisson regression model

$$y_{jkl} \overset{\text{ind}}{\sim} \text{Poi}(\mu_{jkl}), \tag{43}$$

where

$$\log \mu_{ijk} = \beta_0 + \beta_j^1 + \beta_k^2 + \beta_l^3 \quad \text{(mutual independence)}; \tag{44}$$

$$\log \mu_{ijk} = \beta_0 + \beta_j^1 + \beta_k^2 + \beta_l^3 + \beta_{kl}^{2,3} \quad \text{(joint independence)}; \tag{45}$$

$$\log \mu_{ijk} = \beta_0 + \beta_j^1 + \beta_k^2 + \beta_l^3 + \beta_{jk}^{1,2} + \beta_{jl}^{1,3} \quad \text{(mutual independence)}. \tag{46}$$

3 Negative binomial regression

Overdispersion. A pervasive issue with Poisson regression is *overdispersion*: that the variances of observations are greater than the corresponding means. A common cause of overdispersion is omitted variable bias. Suppose that $y \sim \text{Poi}(\mu)$, where $\log \mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. However, we omitted variable x_2 and are considering the GLM based on $\log \mu = \beta_0 + \beta_1 x_1$. If $\beta_2 \neq 0$ and x_2 is correlated with x_1, then we have a confounding issue. Let's consider the more benign situation that x_2 is independent of x_1. Then, we have

$$\mathbb{E}[y|x_1] = \mathbb{E}[\mathbb{E}[y|x_1, x_2]|x_1] = \mathbb{E}[e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}|x_1] = e^{\beta_0 + \beta_1 x_1}\mathbb{E}[e^{\beta_2 x_2}] = e^{\beta_0' + \beta_1 x_1}. \tag{47}$$

So in the model for the mean of y, the impact of omitted variable x_2 seems only to have impacted the intercept. Let's consider the variance of y:

$$\text{Var}[y|x_1] = \mathbb{E}[\text{Var}[y|x_1, x_2]|x_1] + \text{Var}[\mathbb{E}[y|x_1, x_2]|x_1] = e^{\beta_0' + \beta_1 x_1} + e^{2(\beta_0' + \beta_1 x_1)}\text{Var}[e^{\beta_2 x_2}] > e^{\beta_0' + \beta_1 x_1} = \mathbb{E}[y|x_1]. \tag{48}$$

So indeed, the variance is larger than what we would have expected under the Poisson model.

Hierarchical Poisson regression. Let's say that $y|\boldsymbol{x} \sim \text{Poi}(\lambda)$, where $\lambda|\boldsymbol{x}$ is random due to the fluctuations of the omitted variables. A common distribution used to model nonnegative random variables is the *gamma* distribution $\Gamma(\mu, k)$, parameterized by a mean $\mu > 0$ and a *shape* $k > 0$. This distribution has probability density function

$$f(\lambda; k, \mu) = \frac{(k/\mu)^k}{\Gamma(k)} e^{-k\lambda/\mu} \lambda^{k-1}, \tag{49}$$

with mean and variance given by

$$\mathbb{E}[\lambda] = \mu; \quad \text{Var}[\lambda] = \mu^2/k. \tag{50}$$

Therefore, it makes sense to augment the Poisson regression model as follows:

$$\lambda|\boldsymbol{x} \sim \Gamma(\mu, k), \quad \log \mu = \boldsymbol{x}^T \boldsymbol{\beta}, \quad y|\lambda \sim \text{Poi}(\lambda). \tag{51}$$

Negative binomial distribution. A simpler way to write the hierarchical model (51) would be to marginalize out λ. Doing so leaves us with a count distribution called the *negative binomial distribution*:

$$\lambda \sim \Gamma(\mu, k),\ y|\lambda \sim \text{Poi}(\lambda) \quad \Longrightarrow \quad y \sim \text{NegBin}(\mu, k). \tag{52}$$

The negative binomial probability mass function is

$$p(y; \mu, k) = \int_0^\infty \frac{(k/\mu)^k}{\Gamma(k)} e^{-k\lambda/\mu} \lambda^{k-1} e^{-\lambda} \frac{\lambda^y}{y!} d\lambda = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)} \left(\frac{\mu}{\mu+k}\right)^y \left(\frac{k}{\mu+k}\right)^k. \tag{53}$$

This random variable has mean and variance given by

$$\mathbb{E}[y] = \mathbb{E}[\lambda] = \mu \quad \text{and} \quad \text{Var}[y] = \mathbb{E}[\lambda] + \text{Var}[\lambda] = \mu + \frac{\mu^2}{k}. \tag{54}$$

Negative binomial as exponential dispersion model. If we treat k as known, then the negative binomial distribution is in the exponential family:

$$p(y; \mu, k) = \exp\left(y \log \frac{\mu}{\mu+k} - k \log \frac{\mu+k}{k}\right) \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)}. \tag{55}$$

We can read off that

$$\theta = \log \frac{\mu}{\mu+k}, \quad \psi(\theta) = k \log \frac{\mu+k}{k} = -k \log(1 - e^\theta). \tag{56}$$

This is a regular exponential family model, and not an exponential dispersion model. Given the extra parameter k controlling the variance, we may have been expecting to see an EDM. We can arrive at the EDM form by putting $1/k$ in the denominator:

$$p(y; \mu, k) = \exp\left(\frac{\frac{y}{k} \log \frac{\mu}{\mu+k} - \log \frac{\mu+k}{k}}{1/k}\right) \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)}. \tag{57}$$

Note that the "normalized" variable y/k has the EDM distribution rather than the count variable y; this parallels our modeling of the binomial *proportion* (rather than the binomial count) as an EDM. We then see that y/k has the dispersion parameter $\phi = 1/k$. An alternate parameterization of the negative binomial model is via $\gamma = \phi = 1/k$. Here, γ is called the negative binomial *dispersion*.

Negative binomial regression. Let's revisit the hierarchical model (51), writing it more succinctly in terms of the negative binomial distribution:

$$y_i \overset{\text{ind}}{\sim} \text{NegBin}(\mu_i, \gamma), \quad \log \mu_i = \boldsymbol{x}^T \boldsymbol{\beta}. \tag{58}$$

Notice that we typically assume that all observations share the same dispersion parameter γ. Reading off from equation (56), we see that the canonical link function for the negative binomial distribution is $\mu \mapsto \log \frac{\mu}{\mu+k}$. However, typically for negative binomial regression we use the log link $g(\mu) = \log \mu$ instead. This is our first example of a non-canonical link!

Estimation in negative binomial regression. Negative binomial regression is an EDM when γ is known, but typically the dispersion parameter is unknown. Note that there is a dependency in ψ on k (i.e. on γ), which complicates things. It means that the estimate $\widehat{\boldsymbol{\beta}}$ depends on the parameter γ (this does not happen, for example, in the normal linear model case).[1] Therefore, estimation in negative binomial regression is typically an iterative procedure, where at each step $\boldsymbol{\beta}$ is estimated for the current value of γ and then γ is estimated based on the updated value of $\boldsymbol{\beta}$. Let's discuss each of these tasks in turn. Given a value of γ, we have the normal equations

$$0 = \boldsymbol{X}^T \text{diag}\left(\frac{\partial\mu_i/\partial\eta_i}{\text{Var}[y_i]}\right)(\boldsymbol{y}-\boldsymbol{\mu}) = \boldsymbol{X}^T \text{diag}\left(\frac{\mu_i}{\mu_i+\gamma\mu_i^2}\right)(\boldsymbol{y}-\boldsymbol{\mu}) = \boldsymbol{X}^T \text{diag}\left(\frac{1}{1+\gamma\mu_i}\right)(\boldsymbol{y}-\boldsymbol{\mu}). \quad (59)$$

This reduces to the Poisson normal equations when $\gamma = 0$. Solving these equations for a fixed value of γ can be done via IRLS, as usual. Estimating γ for a fixed value of $\boldsymbol{\beta}$ can be done in several ways, including setting to zero the derivative of the likelihood with respect to γ. This results in a nonlinear equation (not given here) that must be solved iteratively.

Wald inference. Note that

$$\boldsymbol{W}_{ii} = \frac{(\partial\mu_i/\partial\eta_i)^2}{\text{Var}[y_i]} = \frac{\mu_i^2}{\mu_i+\gamma\mu_i^2} = \frac{\mu_i}{1+\gamma\mu_i}. \quad (60)$$

Hence, Wald inference is based on

$$\widehat{\text{Var}}[\widehat{\boldsymbol{\beta}}] = (\boldsymbol{X}^T\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}, \quad \text{where} \quad \widehat{\boldsymbol{W}} = \text{diag}\left(\frac{\widehat{\mu}_i}{1+\widehat{\gamma}\widehat{\mu}_i}\right). \quad (61)$$

Likelihood ratio test inference. The negative binomial deviance is

$$D(\boldsymbol{y};\widehat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n}\left(y_i \log\frac{y_i}{\widehat{\mu}_i} - \left(y_i+\frac{1}{\widehat{\gamma}}\right)\log\frac{1+\widehat{\gamma}y_i}{1+\widehat{\gamma}\widehat{\mu}_i}\right). \quad (62)$$

We can use this for comparing nested models and for goodness of fit testing, as usual.

Testing for overdispersion. It is reasonable to want to test for overdispersion, i.e. to test the null hypothesis $H_0 : \gamma = 0$. This is somewhat of a tricky task, because $\gamma = 0$ is at the edge of the parameter space. There are formal tests of this hypothesis, but they are beyond the scope of this course. Another approach is to simply fit a negative binomial model and get a confidence interval for γ. It is probably not particularly reliable for small values of γ, but if it is far away from zero then likely we have some overdispersion on our hands. Finally, if goodness of fit tests in the Poisson model are significant, this may be an indication of overdispersion. It may also be an indication of omitted variable bias (e.g. you forgot to include an interaction), so it's somewhat tricky.

Overdispersion in logistic regression. Note that overdispersion is potentially an issue not only in Poisson regression models, but in logistic regression models as well. Dealing with overdispersion in the latter case is more tricky, because the analog of the negative binomial model (the beta-binomial model) is not an exponential family. An alternate route to dealing with overdispersion is quasi-likelihood modeling, but this topic is beyond the scope of the course.

[1] Having said that, the dependency between $\widehat{\boldsymbol{\beta}}$ and $\widehat{\gamma}$ is weak, as the two are asymptotically independent parameters.

4 R demo

```
library(tidyverse)
```

Here we are again, face to face with the crime data, with one last chance to get the analysis right. Let's load and preprocess it, as before.

```
# read crime data
crime_data = read_tsv("../data/Statewide_crime.dat")

# read and transform population data
population_data = read_csv("../data/state-populations.csv")
population_data = population_data %>%
  filter(State != "Puerto Rico") %>%
  select(State, Pop) %>%
  rename(state_name = State, state_pop = Pop)

# collate state abbreviations
state_abbreviations = tibble(state_name = state.name,
                             state_abbrev = state.abb) %>%
  add_row(state_name = "District of Columbia", state_abbrev = "DC")

# add CrimeRate to crime_data
crime_data = crime_data %>%
  mutate(STATE = ifelse(STATE == "IO", "IA", STATE)) %>%
  rename(state_abbrev = STATE) %>%
  filter(state_abbrev != "DC") %>%       # remove outlier
  left_join(state_abbreviations, by = "state_abbrev") %>%
  left_join(population_data, by = "state_name") %>%
  select(state_abbrev, Violent, Metro, HighSchool, Poverty, state_pop)

crime_data
```

```
## # A tibble: 50 x 6
##    state_abbrev Violent Metro HighSchool Poverty state_pop
##    <chr>          <dbl> <dbl>      <dbl>   <dbl>     <dbl>
##  1 AK               593  65.6       90.2     8      724357
##  2 AL               430  55.4       82.4    13.7   4934193
##  3 AR               456  52.5       79.2    12.1   3033946
##  4 AZ               513  88.2       84.4    11.9   7520103
##  5 CA               579  94.4       81.3    10.5  39613493
##  6 CO               345  84.5       88.3     7.3   5893634
##  7 CT               308  87.7       88.8     6.4   3552821
##  8 DE               658  80.1       86.5     5.8    990334
##  9 FL               730  89.3       85.9     9.7  21944577
## 10 GA               454  71.6       85.2    10.8  10830007
## # ... with 40 more rows
```

Let's recall the logistic regression we ran on these data in Unit 4:

```
bin_fit = glm(Violent/state_pop ~ Metro + HighSchool + Poverty,
              weights = state_pop,
              family = "binomial",
              data = crime_data)
```

Recall that everything was significant:

```
summary(bin_fit)

##
## Call:
## glm(formula = Violent/state_pop ~ Metro + HighSchool + Poverty,
##     family = "binomial", data = crime_data, weights = state_pop)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -21.043   -9.176    0.418    9.053   47.174
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.609e+01  3.520e-01  -45.72   <2e-16 ***
## Metro       -2.586e-02  5.727e-04  -45.15   <2e-16 ***
## HighSchool   9.106e-02  3.450e-03   26.39   <2e-16 ***
## Poverty      6.077e-02  4.852e-03   12.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 15590  on 49  degrees of freedom
## Residual deviance: 11742  on 46  degrees of freedom
## AIC: 12136
##
## Number of Fisher Scoring iterations: 5
```

But there were already signs of trouble in this regression summary. The summary tells us that the residual deviance is 11742 on 46 degrees of freedom. This is a measure of the goodness of fit of the model, as the residual deviance should have a chi-square distribution with 46 degrees of freedom if the model fits well. But this distribution has a mean of 46, so having a value of 11742 seems way too large. We can confirm this suspicion with a formal deviance-based goodness of fit test:

```
pchisq(bin_fit$deviance,
       df = bin_fit$df.residual,
       lower.tail = FALSE)

## [1] 0
```

Wow, we get a p-value of zero! Let's try doing a score-based (i.e. Pearson) goodness of fit test:

```
pchisq(sum(resid(bin_fit, "pearson")^2),
       df = bin_fit$df.residual,
       lower.tail = FALSE)

## [1] 0
```

Here the code `sum(resid(bin_fit, "pearson")^2)` extracts the sum of the squares of the Pearson residuals, which we did not discuss, but which gives us Pearson's X^2 statistic. So in this case, we again get a p-value of zero! So this model definitely does not fit well. We have therefore omitted some important variables and/or we have serious overdispersion on our hands.
We haven't discussed in any detail how to deal with overdispersion in logistic regression models, so let's try a Poisson model instead. The natural way to model rates using Poisson distributions is via offsets:

```
pois_fit = glm(Violent ~ Metro + HighSchool + Poverty + offset(log(state_pop)),
               family = "poisson",
               data = crime_data)
summary(pois_fit)

##
## Call:
## glm(formula = Violent ~ Metro + HighSchool + Poverty + offset(log(state_pop)),
##     family = "poisson", data = crime_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -21.042   -9.176    0.418    9.051   47.170
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.609e+01  3.520e-01  -45.72   <2e-16 ***
## Metro       -2.585e-02  5.727e-04  -45.15   <2e-16 ***
## HighSchool   9.106e-02  3.450e-03   26.39   <2e-16 ***
## Poverty      6.077e-02  4.852e-03   12.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 15589  on 49  degrees of freedom
## Residual deviance: 11741  on 46  degrees of freedom
## AIC: 12135
##
## Number of Fisher Scoring iterations: 5
```

Again, everything is significant, and again, the regression summary shows that we have a huge residual deviance. This was to be expected, given that $\text{Bin}(m, \pi) \approx \text{Poi}(m\pi)$ for large m and small π. So, the natural thing to try is a negative binomial regression! Negative binomial regression is not implemented in the regular `glm` package, but `glm.nb()` from the `MASS` package is a dedicated

function for this task. Let's see what we get:

```
nb_fit = MASS::glm.nb(Violent ~ Metro + HighSchool + Poverty + offset(log(state_pop)),
                data = crime_data)
summary(nb_fit)

##
## Call:
## MASS::glm.nb(formula = Violent ~ Metro + HighSchool + Poverty +
##     offset(log(state_pop)), data = crime_data, init.theta = 1.467747388,
##     link = log)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.62929  -1.02800  -0.54853   0.07234   2.71356
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -10.254088   5.273418  -1.944   0.0518 .
## Metro        -0.012188   0.008518  -1.431   0.1525
## HighSchool    0.028052   0.052482   0.535   0.5930
## Poverty      -0.026852   0.068449  -0.392   0.6948
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.4677) family taken to be 1)
##
##     Null deviance: 59.516  on 49  degrees of freedom
## Residual deviance: 55.487  on 46  degrees of freedom
## AIC: 732.58
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.468
##           Std. Err.:  0.268
##
##  2 x log-likelihood:  -722.575
```

Aha! Things are not looking so significant anymore! And the residual deviance is not as huge! The estimated value of γ (confusingly called θ in the summary) is significantly different from zero, indicating overdispersion. Now it appears that this model fits. Finally! Let's do a deviance-based goodness of fit test to make sure:

```
pchisq(nb_fit$deviance,
       df = nb_fit$df.residual,
       lower.tail = FALSE)

## [1] 0.1594508
```

Ok, great. Now that we have a well-fitting model, we can do inference within this model that we can trust. For example, we can get Wald confidence intervals:

```
confint.default(nb_fit)
```

```
##                    2.5 %      97.5 %
## (Intercept) -20.58979658 0.081620714
## Metro        -0.02888413 0.004507747
## HighSchool   -0.07481066 0.130915138
## Poverty      -0.16100973 0.107305015
```

Or we can get LRT-based (i.e. profile) confidence intervals:

```
confint(nb_fit)
```

```
## Waiting for profiling to be done...
```

```
##                    2.5 %       97.5 %
## (Intercept) -19.20209590 -0.860399348
## Metro        -0.03153902  0.006365841
## HighSchool   -0.06265118  0.115318303
## Poverty      -0.13930110  0.085200541
```

Or we can get confidence intervals for the predicted means:

```
predict(nb_fit,
        newdata = crime_data %>% column_to_rownames(var = "state_abbrev"),
        type = "response",
        se.fit = TRUE)
```

```
## $fit
##        AK        AL        AR        AZ        CA        CO        CT        DE
##  116.1520  617.7064  375.4895  700.6931 3257.5300  725.1538  436.7863  127.2572
##        FL        GA        HI        ID        IL        IN        IA        KS
## 2232.2308 1301.2937  157.1416  263.8572 1379.1847  954.3366  546.5503  439.0649
##        KY        LA        MA        MD        ME        MI        MN        MO
##  541.5706  391.6745  747.7454  737.0032  274.2879 1322.9956  970.4078  871.2829
##        MS        MT        NC        ND        NE        NH        NJ        NM
##  380.6756  199.4947 1313.0904  134.8128  305.0634  261.1975  966.9940  204.3311
##        NV        NY        OH        OK        OR        PA        RI        SC
##  327.7316 1926.3861 1477.1713  495.9711  517.8397 1600.0813   96.3565  684.9102
##        SD        TN        TX        UT        VA        VT        WA        WI
##  160.9225  867.0224 2423.0647  416.6648 1244.5168  148.1635 1012.1932  892.0644
##        WV        WY
##  226.4515  100.1906
##
## $se.fit
##        AK        AL        AR        AZ        CA        CO        CT        DE
##  21.00552 143.65071 130.44272 165.08459 910.57769 121.34777  85.53768  32.15169
##        FL        GA        HI        ID        IL        IN        IA        KS
```

```
## 427.89514 173.04544  31.73873  40.28262 239.43324 147.21049 104.05752  68.82044
##        KY        LA        MA        MD        ME        MI        MN        MO
## 133.28938 129.40665 150.23524 158.93816  92.04222 171.28409 216.32477 110.88843
##        MS        MT        NC        ND        NE        NH        NJ        NM
## 138.28105  65.60335 379.90855  26.74061  69.62560  66.73731 220.88371  59.26953
##        NV        NY        OH        OK        OR        PA        RI        SC
##  64.30971 387.25204 241.24541  95.44911  81.97419 220.42078  33.97964 119.45174
##        SD        TN        TX        UT        VA        VT        WA        WI
##  41.50215 169.68896 738.95321 107.62725 209.14651  51.32810 191.75629 137.35158
##        WV        WY
##  71.55328  22.79279
##
## $residual.scale
## [1] 1
```

We can carry out some hypothesis tests as well, e.g. to test $H_0 : \beta_{\text{Metro}} = 0$:

```
nb_fit_partial = MASS::glm.nb(Violent ~ HighSchool + Poverty + offset(log(state_pop)),
                    data = crime_data)
anova_fit = anova(nb_fit_partial, nb_fit)
anova_fit
```

```
## Likelihood ratio tests of Negative Binomial Models
##
## Response: Violent
##                                                    Model    theta Resid. df
## 1         HighSchool + Poverty + offset(log(state_pop)) 1.428675        47
## 2 Metro + HighSchool + Poverty + offset(log(state_pop)) 1.467747        46
##      2 x log-lik.   Test    df LR stat.   Pr(Chi)
## 1       -724.1882
## 2       -722.5753 1 vs 2     1 1.612878 0.2040877
```