PUAD 7130: Limited and Categorical Dependent Variables

David A. Hughes, Ph.D.

Auburn University at Montgomery

david.hughes@aum.edu

November 11, 2021

Overview

Assumptions of the Gauss-Markov Theorem

- When our data exhibit endogeneity, heteroskedasticity, autocorrelation, etc., the assumptions of Gauss-Markov are violated.
- This means that our $\hat{\beta}$s are biased or that $\hat{\sigma}_{\hat{\beta}}$s are inefficient.
- Oftentimes, we can perform a work-around by transforming variables, calculating robust standard errors, etc. But this won't always be the case.

Gauss-Markov and categorical dependent variables

- Categorical and limited dependent variables may pose grave risks to our interpretation of OLS results.
- On the one hand, we're almost certainly violating assumptions of homoskedasticity, normality, etc., which means that we're getting biased or inefficient results, and our ability to hypothesis-test may be compromised.
- On the other hand, interpreting the values of $\hat{\beta}_k$ for categorical dependent variables can be downright weird.

The linear probability model

- Suppose we code judges' votes on the US Courts of Appeals as being either liberal or conservative ("libvote=1" if yes, "0" else)
- Now suppose we want to predict the likelihood a judge casts a liberal vote solely as a function of his or her ideology.
- We'll let the ideology of a judge's appointing president stand in for their own ("potus_ideal $\in [-1, 1]$") such that increasing values represent greater conservatism.
- Imagine we estimated the following linear regression:

$$\widehat{\text{libvote}}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{potus\_ideal}_i.$$

The linear probability model: An example

```
. reg libvote potus_ideal
. predict yhat2
. predict res, res

      Source |       SS           df       MS      Number of obs   =    42,156
-------------+----------------------------------   F(1, 42154)     =    287.99
       Model |  68.2337579         1  68.2337579   Prob > F        =    0.0000
    Residual |  9987.62816    42,154  .23693192    R-squared       =    0.0068
-------------+----------------------------------   Adj R-squared   =    0.0068
       Total |  10055.8619    42,155  .238544939   Root MSE        =    .48676

------------------------------------------------------------------------------
     libvote |      Coef.   Std. Err.      t    P>|t|     [95\% Conf. Interval]
-------------+----------------------------------------------------------------
 potus_ideal |  -.0838547   .0049413   -16.97   0.000    -.0935397   -.0741696
       _cons |   .3978374   .0023882   166.59   0.000     .3931565    .4025184
------------------------------------------------------------------------------
```

Do we have homoskedasticity?

```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of potus_ideal

         chi2(1)      =     7.48
         Prob > chi2  =   0.0062
```

No.

Do we have normally distributed errors?

```
. swilk res

                   Shapiro-Wilk W test for normal data

    Variable |        Obs       W           V         z       Prob>z
-------------+------------------------------------------------------
         res |     42,156    0.70748    4745.478    23.407    0.00000

Note: The normal approximation to the sampling distribution of W'
      is valid for 4<=n<=2000
```

No.

Interpreting the LPM

- Suppose we came up with a LPM that adhered to all of the Gauss-Markov assumptions.
- We still have a problem insofar as we don't really know how to interpret model parameters like slope coefficients.
- All of these problems tell us that OLS is not the appropriate estimator.

Introduction

- Suppose our dependent variable is measured such that $Y_i \in \{0, 1\}$.
- Recall that the linear probability model violates a number of desirable assumptions of OLS.
- We'd like instead to model the actual probability of observing either a "0" or "1."

A latent variable approach

- Suppose there exists an underlying measure of Y_i that is measured on a continuous scale. Call this latent variable Y_i^*. The underlying model is therefore,

$$Y_i^* = \boldsymbol{X_i}\boldsymbol{\beta} + \epsilon_i, \tag{1}$$

where ϵ is distributed according to some normal distribution, and $\boldsymbol{X_i}\boldsymbol{\beta}$ represents a matrix of variables and their coefficients.

- We do not observe Y_i^* directly but merely its manifestations in Y_i such that:

$$\begin{aligned} Y_i &= 0 \text{ if } Y_i^* < 0 \\ Y_i &= 1 \text{ if } Y_i^* \geq 0. \end{aligned}$$

A latent approach (cont'd.)

- Taking Equation (1), we can model the probability of observing $Y_i = 1$:

$$\begin{aligned} Pr(Y_i = 1) &= Pr(Y_i^* \geq 0) \\ & \quad Pr(\boldsymbol{X_i \beta} + \epsilon_i \geq 0) \\ & \quad Pr(\epsilon_i \geq -\boldsymbol{X_i \beta}) \\ & \quad Pr(\epsilon_i \leq \boldsymbol{X_i \beta}), \end{aligned} \tag{2}$$

where the last inequality holds due to the symmetry of the distribution of ϵ.

- Because ϵ is assumed normally distributed, we can integrate over it to find $\hat{\boldsymbol{\beta}}$.

Logit

- If we assume that ϵ is distributed according to a logistic probability density function, we get the logit model:

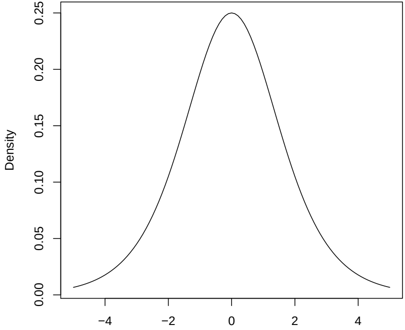
$$Pr(\epsilon) \equiv \lambda(\epsilon) = \frac{\exp(\epsilon)}{[1 + \exp(\epsilon)]^2}. \tag{3}$$

  Equation (3) gives us the probability density function (pdf) of the logistic distribution.

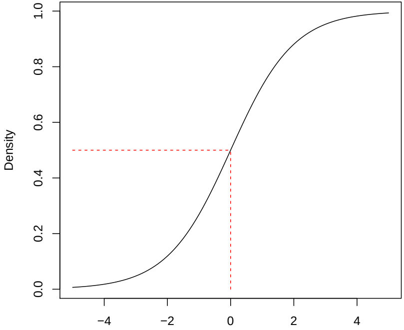
- If we want to calculate the cumulative probability that a variable distributed according to the logistic distribution is less than some value, ϵ, then we use the cumulative density function (cdf):

$$\Lambda(\epsilon) = \int_{-\infty}^{\epsilon} \lambda(\epsilon) d\epsilon = \frac{\exp(\epsilon)}{1 + \exp(\epsilon)} \tag{4}$$

The logistic pdf and cdf



Logistic pdf



Logistic cdf

Specifying the logit model

- Assuming ϵ is distributed according to the standard logistic distribution, we can rewrite Equation (2):

$$Pr(Y_i = 1) \equiv \Lambda(\boldsymbol{X_i\beta}) = \frac{\exp(\boldsymbol{X_i\beta})}{1 + \exp(\boldsymbol{X_i\beta})}. \tag{5}$$

- To extract a probabilistic statement from Equation (5), we are going to make use of a concept known as maximum likelihood estimation (MLE), the derivation of which is beyond the scope of this course.

Probit

- If we assume that ϵ_i is distributed standard normally (i.e., $\epsilon_i \sim N(0,1)$), then we estimate a probit rather than a logit.
- The pdf of a standard normal distribution is:

$$Pr(\epsilon) \equiv \phi(\epsilon) = \frac{1}{\sqrt{2\pi}} \text{exp}\left(\frac{-\epsilon^2}{2}\right) \tag{6}$$

- The cdf for the standard normal is given by:

$$\Phi(\epsilon) = \int_{-\infty}^{\epsilon} \frac{1}{\sqrt{2\pi}} \text{exp}\left(\frac{-\epsilon^2}{2}\right) d\epsilon. \tag{7}$$

Specifying the probit

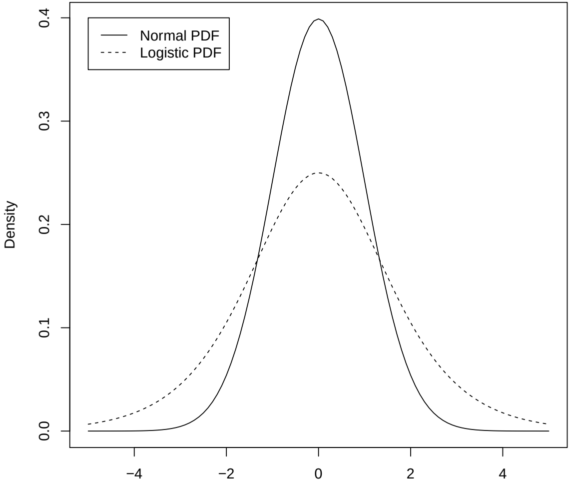
$$
\begin{aligned}
Pr(Y_i = 1) &= \Phi(\mathbf{X_i}\beta) \\
&= \int_{-\infty}^{\mathbf{X_i}\beta} \frac{1}{\sqrt{2\pi}} \text{exp}\left(\frac{-\mathbf{X_i}\beta^2}{2}\right) d\mathbf{X_i}\beta \qquad (8)
\end{aligned}
$$

- The standard normal may be a better specification for ϵ.
- But unlike the standard logistic cdf, we can't calculate the integral via a closed-form solution.
- Hence, we must use approximation methods.
- Also, we can't extract probabilities so easily as we did in logit.

Comparing logit and probit

- Each is single-peaked and symmetric.
- But logit has fatter tails than does probit.
- Logit coefficients are about 1.7 times larger than probit coefficients.
- But this turns out not to really matter.

Comparing logit and probit pdfs



Predicting election day winners

- Suppose we're interested in modeling why some candidates for office win and some lose.
- We therefore estimate the following logistic regression:

$$\begin{aligned} Pr(\text{Winner}_i = 1 \mid \boldsymbol{X_i}) &= \Lambda(\beta_0 + \beta_1 \text{Money}_i + \beta_2 \text{Incumbent}_i + \\ &\quad \beta_3 \text{Nonwhite}_i + \beta_4 \text{Female}_i), \end{aligned}$$

where Money_i measures a candidates campaign fundraising in millions, Incumbent_i is a dummy variable for whether the candidate is an incumbent, and Nonwhite_i and Female_i are dummy variables indicating nonwhite and female candidates, respectively.

What do we make of our logit/probit results?

```
. logit winner cm_justice_million incumbent nonwhite female

Iteration 0:   log likelihood = -444.63745
Iteration 1:   log likelihood = -282.03634
Iteration 2:   log likelihood = -274.42329
Iteration 3:   log likelihood = -274.29387
Iteration 4:   log likelihood = -274.29368
Iteration 5:   log likelihood = -274.29368

Logistic regression                      Number of obs   =        668
                                         LR chi2(4)      =     340.69
                                         Prob > chi2     =     0.0000
Log likelihood = -274.29368              Pseudo R2       =     0.3831
```

winner	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
cm_justice_million	.4805724	.1854172	2.59	0.010	.1171613	.8439834
incumbent	3.682391	.2611491	14.10	0.000	3.170549	4.194234
nonwhite	-1.072303	.3432889	-3.12	0.002	-1.745137	-.3994688
female	.5514401	.2426576	2.27	0.023	.0758401	1.02704
_cons	-1.159442	.1667965	-6.95	0.000	-1.486357	-.8325266

Interpreting probit and logit results

- "Signs and significance" (not great but better than nothing)
- Marginal effects (e.g., standardize the IVs or $\hat{\beta}$s)
- Predicted probabilities (but over what range?)

X_k's nonlinear effect on Y_i

- The estimated effect of some X_k on the DV ($\hat{\beta}_k$) is linear only with respect to the latent variable, Y_i^*.
- Hence, we cannot interpret $\hat{\beta}_k$ as a linear effect on $\hat{Y}_i$.
- The real net effect of X_k is also a function of the other variables, their coefficient estimates, and the constant:

$$\frac{\partial Pr(\hat{Y}_i = 1)}{\partial X_k} \equiv \lambda(X) = \frac{\exp(X_i\hat{\beta})}{[1+\exp(X_i\hat{\beta})]^2}\hat{\beta}_k. \tag{9}$$

- Unlike in OLS, then, the first derivative of the function with respect to $\hat{\beta}_k$ is non-constant.

Predicted Probabilities

- Generically, we can estimate the change in predicting a "1" across two values of X_k:

$$\Delta Pr(Y_i = 1)_{X_A \rightarrow X_B} = \frac{\exp(\boldsymbol{X_B}\hat{\boldsymbol{\beta}})}{1+\exp(\boldsymbol{X_B}\hat{\boldsymbol{\beta}})} - \frac{\exp(\boldsymbol{X_A}\hat{\boldsymbol{\beta}})}{1+\exp(\boldsymbol{X_A}\hat{\boldsymbol{\beta}})}, \quad (10)$$

for logits, and

$$\Delta Pr(Y_i = 1)_{X_A \rightarrow X_B} = \Phi(\boldsymbol{X_B}\hat{\boldsymbol{\beta}}) - \Phi(\boldsymbol{X_A}\hat{\boldsymbol{\beta}}), \quad (11)$$

for probits.

Predicted probabilities: Example

- Suppose I want to know the change in the predicted probability a candidate wins if they raise no money, versus if they raise \$1 million, versus if they raise \$2 million.
- To isolate this effect, I'll hold the other IVs equal to zero (hence, this means a non-incumbent who is a white man).

Predicted probabilities: Example (cont'd.)

$$Pr(Y_i = 1 \mid \mathbf{X_i}) \quad = \quad \Lambda(-1.16 + 0.48\text{Money}_i)$$

Someone raising no money will win with probability:

$$Pr(Y_i = 1) = \Lambda(-1.16) = \frac{\exp(-1.16)}{1 + \exp(-1.16)} = 0.24.$$

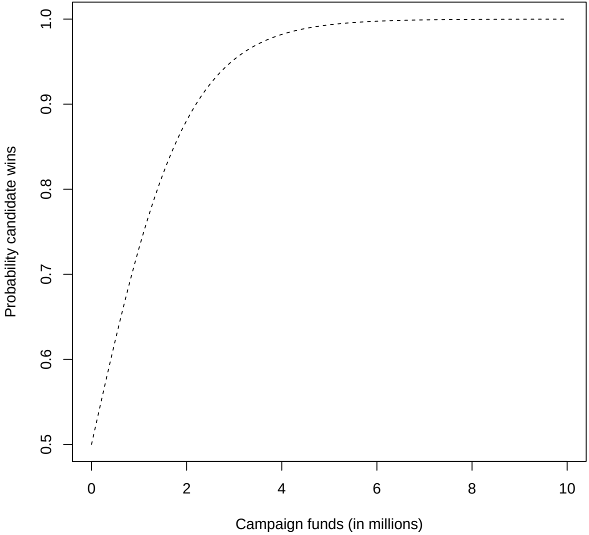
The same person who raises $1 million is predicted to win with probability:

$$Pr(Y_i = 1) = \Lambda(-1.16 + .48) = \frac{\exp(-.68)}{1 + \exp(-.68)} = 0.34.$$

And for $2 million:

$$Pr(Y_i = 1) = \Lambda(-1.16 + .96) = \frac{\exp(-.2)}{1 + \exp(-.2)} = 0.45.$$

Graphing the probability function



Measuring goodness of fit

- Pseudo-R^2s
- Wald/LR χ^2
- PREs
- Information criteria

Pseudo-R^2

- There are a few types out there, but none of them can be interpreted in the same manner as R^2 in OLS. In fact, these are rarely reported in the literture.
- Perhaps one of the more common measures is McFadden's R^2:

$$R^2_{\text{McFadden}} = 1 - \frac{LL_{m1}}{LL_{null}},$$

where LL_{m1} represents the log-likelihood from the model you estimated, and LL_{null} represents the log-likelihood from an intercept-only model.

Wald/LR χ^2 test

- Your Wald/LR χ^2 is kind of like the F-statistic from OLS.
- It's telling you how good of a job overall your model is doing at improving upon the null model.
- Always report this figure and its corresponding p-value.

Information criteria

- In OLS, “adjusted” R^2 is a parameter that measures goodness-of-fit, scaled by the number of covariates included in the model.
- We can take similar measures in MLE.
- AIC and BIC are the two both popular approaches.
- Smaller information criteria are preferred.
- Stata reports these after a regression:
  estat ic

Proportional Reduction of Error

- Let the observed dependent variable ($\boldsymbol{y}$) equal 0 or 1 (this is generalizable beyond two categories)
- Let π represent the predicted probability that $(\mathbf{y}_i) = 1$
- And $\pi_i = \Pr(\mathbf{y} = 1|\boldsymbol{X}_i) = f(\boldsymbol{X_i\beta})$
- where $f =$ the cdf for the Normal distribution in probit and the cdf for the Logistic distribution in logit

Proportional Reduction of Error

- Define the expected value for $\hat{\mathbf{y}}$ as:

$$\hat{\mathbf{y}} = \begin{cases} 0 & \text{if } \hat{\pi}_i \leq 0.5 \\ 1 & \text{if } \hat{\pi}_i > 0.5 \end{cases}$$

- A table is helpful for comparison purposes and helps visualize the intuition behind this approach

	Observed Values	
	0	1
Predicted 0	$+$	-
Predicted 1	-	$+$

Proportional Reduction of Error

- The Proportional Reduction of Error (PRE) statistic calculates the proportion of + versus the proportion of - to determine the predictive accuracy, using this formula:

$$\text{PRE} = \frac{\%\text{ correctly predicted} - \%\text{ in modal category}}{100 - \%\text{ in modal category}}$$

How do we compare multiple models?

- In OLS, we could perform an F-test to determine whether certain indicators statistically improved the overall fit of the model.
- We do the same thing in MLE largely via the likelihood ratio test.
- We start with a fully specified model (unconstrained) and compare it to a nested version of that model (constrained).
- We then use the ratio of these two models' log-likelihoods and conduct a χ^2 test. The null hypothesis is that the two models have equal explanatory power (i.e., $LL_{m1} = LL_{m2}$).

Likelihood ratio testing

- Compares β estimates from a constrained and unconstrained model
- Assesses the imposed constraint by comparing the log-likelihoods of the constrained model to the unconstrained one
- H_0: $\beta_u = \beta_c$

Ordered Logit/Probit

- Start with a latent variable such that:

$$Y^* \;\; = \;\; \mu + u_i.$$

- Similar to how we motivated the binary response model, suppose that:

$$Y_i = j \text{ if } \tau_{j-1} \leq Y_i^* < \tau_j,\ j \in \{1, \ldots, J\}.$$

- Therefore, Y has J ordered outcome ctaegories and $J - 1$ "cutpoints" (τ).

Estimating the ordered logit/probit

- We can express the probability of any particular discrete outcome on Y as:

$$\begin{aligned} Pr(Y_i = j) &= Pr(\tau_{j-1} \leq Y^* < \tau_j) \\ &= Pr(\tau_{j-1} \leq \mu + u_i < \tau_j). \end{aligned}$$

- With minimal assumptions, we can substitute $\boldsymbol{X_i \beta}$ for μ:

$$\mu_i = \boldsymbol{X_i \beta}.$$

Estimating the ordered logit/probit (cont'd.)

- We can rewrite the above equations as:

$$
\begin{aligned}
Pr(Y_i = j \mid \mathbf{X}, \beta) &= Pr(\tau_j - 1 \leq Y_i^* < \tau_j \mid \mathbf{X}) \\
&= Pr(\tau_{j-1} \leq \mathbf{X_i}\beta + u_i < \tau_j) \\
&= Pr(\tau_{j-1} - \mathbf{X_i}\beta \leq u_i < \tau_j - \mathbf{X_i}\beta) \\
&= \int_{-\infty}^{\tau_j - \mathbf{X_i}\beta} f(u_i)du - \int_{-\infty}^{\tau_{j-1} - \mathbf{X_i}\beta} f(u_i)du) \\
&= F(\tau_j - \mathbf{X_i}\beta) - F(\tau_{j-1} - \mathbf{X_i}\beta),
\end{aligned}
$$

where f is the density for u, and F is the corresponding cdf.

- The intuition is that we "cut" the density at different points, and the probability of a given observation receiving the the of Y associated with this interval is simply the area under the density curve between those points.

Estimating the ordered logit/probit (cont'd.)

- Proceeding with the standard normal, and assuming we have a three outcome DV:

$$
\begin{aligned}
Pr(Y_i = 1) &= \Phi(\tau_1 - \mathbf{X_i}\beta) - 0 \\
Pr(Y_i = 2) &= \Phi(\tau_2 - \mathbf{X_i}\beta) - \Phi(\tau_1 - \mathbf{X_i}\beta) \\
Pr(Y_i = 3) &= 1 - \Phi(\tau_2 - \mathbf{X_i}\beta).
\end{aligned}
$$

What about the intercept?

- We often think of the τs in ordered models as being a series of "intercepts."
- In the binary model, the intercept tells us the probability $Y = 1 \mid \mathbf{X_i} = 0$—that is, the probability of being in either category of Y.
- An identification problem occurs if we try to estimate the intercept *and* all $J - 1$ cutpoints.

Motivating the model

- Consider a set of N individuals, $i \in \{1, 2, \ldots, N\}$ with dependent variable Y_i that takes on J *unordered* responses.
- Let $Pr(Y_i = 1) = P_{ij}$ and note that $\sum_{j=1}^{J} P_{ij} = 1$. That is, every i is required to make at least *some* choice in J.
- Naturally, we want to allow P_{ij} to vary as a function of some k independent variable(s), $\boldsymbol{X_i}$, indexed by a $k \times 1$ vector of parameters specific to that outcome, $\boldsymbol{\beta_j}$.

Motivating the model (cont'd.)

- As before, we'll make use of the exponential function:

$$P_{ij} \quad = \quad \exp(\boldsymbol{X_i \beta_j}).$$

- However, $\sum_{j=1}^{J} \neq 1$, which it must be. Therefore, we rescale P_{ij} by dividing each by the sum of all P_{ij}s:

$$Pr(Y_i = j) \equiv P_{ij} = \frac{\exp(\boldsymbol{X_i \beta_j})}{\sum_{j=1}^{J} \exp(\boldsymbol{X_i \beta_j})}. \qquad (12)$$

Motivating the model (cont'd.)

- Equation one helps us to express what we will term the mult-inomial logit (MNL).
- Unfortunately, as specified, Equation 1 is unidentified.
- That is, there are an infinite set of $\boldsymbol{\beta_j}$s that will render identical sets of probabilities.
- This problem is similar to what we encounted with the ordinal logit/probit *vis-à-vis* the constant term.

Motivating the model (cont'd.)

- To address the identification problem, we constrain the parameters for one of the outcomes, J, to zero making that category the baseline for comparison to other outcomes.
- If we omit the first category, then Equation 1 changes as such:

$$Pr(Y_i = 1) \quad = \quad \frac{1}{1 + \sum_{j=2}^{J} \exp(\boldsymbol{X_i \beta'_j})},$$

where $\boldsymbol{\beta'_j}$ represents the rescaled influence of the various $\boldsymbol{X}s$ on a given outcome, relative to $Pr(Y_i = 1)$.

- We express the probability of the other $J - 1$ alternatives as:

$$\frac{\exp(\boldsymbol{X_i \beta'_j})}{1 + \sum_{j=2}^{J} \exp(\boldsymbol{X_i \beta'_j})}.$$

Interpreting MLE coefficients for the MNL

- MLE yields separate $\hat{\beta}$s for each of the alternatives (except the baseline, which is omitted as its parameters are set to zero).
- Coefficients on given covariates reflect the change in the probability of a given outcome, relative to the omitted category.

The Poisson process

- A good way to think about an event count outcome is to think of them as events that occur over time.
- Let λ denote the constant rate at which events occur—this could be the expected number of events in a given period of length h.
- Then the probability that an event occurs in a given interval is λh, and the probability is does not occur is $1 - \lambda h$.

The Poisson process (cont'd.)

- Let Y_t reflect the number of events that occur in the interval t of length h.
- The probability that the number of events that occurs in $(t, t+h]$ is equal to some value $y \in \{0, 1, 2, 3, \ldots\}$ is:

$$Pr(Y_t = y) = \frac{\exp(-\lambda h)\lambda h^y}{y!}. \quad (13)$$

- And if all the intervals are of equal length 1, Equation (1) becomes:

$$Pr(Y_t = y) = \frac{\exp(-\lambda)\lambda^y}{y!}. \quad (14)$$

The Poisson distribution

- Equations (2) and (3) give the Poisson distribution.
- Critically, we assume that values of Y_t arrive at a constant rate (λ) and are independent across draws from the distribution.
- The parameter λ is interpreted as a rate or the expected number of events during a given period, t. That is, $E(Y) = \lambda$.
- As λ increases:
  - The mean of the distribution gets bigger (shockingly enough)
  - The variance of the distribution gets larger too, and it turns out that $E(Y) = Var(Y) = \lambda$.
  - The distribution becomes a normal distribution (relevant for deciding between MLE and OLS).

Exposure and offsets in Poisson models

- We've been modeling the number of outcomes in a given period so far.
- But what if there never were any opportunities during a given period for a non-zero outcom to occur?
- For example, if I were to model the number of congressional acts the Supreme Court invalidated in a given year, it might be pertinent to know if no congressional acts were even reviewed.

Exposure and offsets in Poisson models (cont'd.)

- We need to account for the exposure term, and the easiest way to do this is to include M_i as an "offset" in the model:

$$\lambda_i = \exp[\boldsymbol{X_i\beta} + ln(M_i)],$$

  which constrains the effect of the offset to a coefficient of 1.
- In Stata, we use the `exposure option when estimating a Poisson.`
- We could also include M_i as a covariate and model its effect on $E(Y_i)$ directly, examining its coefficient to see how close to one it really is.

Some problems with Poisson

- We've made strong assumptions in setting up the Poisson model.
- First, we required that the probability of some event occurring is constant within a given period. And second, we required that the probability of some event occuring was independent of other events during the same period.
- But what if this assumption were violated?

Contagion and dispersion

- Suppose I count the total number of leaves on my hydrangea over four periods: winter, spring, summer, and fall.
- How likely is it that the rate of occurrence (λ) is constant across all four seasons?
- Because we find that the occurrence of observing one leaf increases the likelihood of observing another, we have a "positive contagion."
- This increases the variance of the observed counts, which is bad mojo when we have assumed that $E(Y) = Var(Y) = \lambda$ and leads to a problem known as "overdispersion."

Testing for over-dispersion

- We can test whether we have over/under-dispersion in our data.
- The easiest way to do this is by running a negative binomial regression and checking the statistical significance of the dispersion parameter, α.

Addressing overdispersion

- If we have problems with dispersion, it makes sense to just go ahead and model it directly rather than to rely upon inadequate results from a Poisson.
- Droping the assumption that λ is a constant, we can instead treat it as a random variable:

$$\begin{aligned} E(Y_i) \equiv \lambda_i &= \exp(\boldsymbol{X_i}\boldsymbol{\beta} + u_i) \\ &= \exp(\boldsymbol{X_i}\boldsymbol{\beta})\exp(u_i) \\ &= \lambda_i \nu_i. \end{aligned} \tag{15}$$

- All that's left now is to specify a distribution on u_i. We usually use the Gamma distribution.

The negative binomial distribution

- If ν_i is assumed to be randomly distributed according to a one-parameter Gamma distribution with mean $E(\nu) = 1$ and variance $Var(\nu) = \frac{1}{\alpha}$, then the marginal density of Y is said to be negative binomial:

$$Pr(Y_i = y \mid \lambda_i, \alpha) = \left(\frac{\Gamma(\alpha^{-1} + Y_i)}{\Gamma(\alpha^{-1})\Gamma(Y_i + 1)}\right)\left(\frac{\alpha^{-1}}{\alpha^{-1} + \lambda_i}\right)^{\alpha^{-1}}\left(\frac{\lambda_i}{\lambda_i + \alpha^{-1}}\right)^{Y_i},$$

where Γ is the gamma function.

- We model $\lambda_i = \exp(\boldsymbol{X_i\beta})$, which has $E(Y) = \lambda$ and $Var(Y) = \lambda(1 + \alpha\lambda)$, where $\alpha > 0$.
- Note that when $\alpha = 0$, the negative binomial reduces to the Poisson.

Conclusion

- In this section, we have considered a raft of estimators for models with limited or categorical dependent variables.
- Your choice of estimator largely comes down to the level of measurement in your dependent variable.
- That said, there's a lot we didn't cover today such as model assumptions and so forth.
- This is why your methods training should *not* end with this class.