Model selection for Zero Inflated Negative Binomial Regression Models

Mia Brasil and Payton Burks

Under the Direction of Dr. Abdulla Al Mamun
Department of Mathematics
Gonzaga University

April 23, 2022

Outline of Presentation

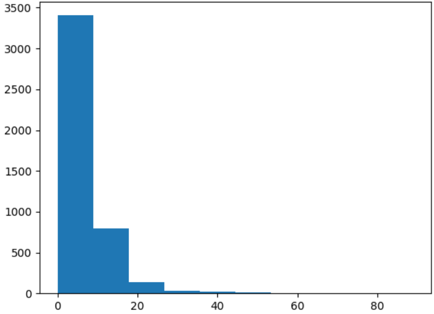
- Introduction: What is Zero-Inflated Negative Binomial and where does it apply in real life?
- Reducing Covariates
- Wald Test and Likelihood Ratio Test
- Simulations and Applications

Introduction and Motivation

Zero Inflated Negative Binomial is a negative binomial distribution which counts the number of failed trials before a certain number of successes where the data has a large number of zeros (ie zero inflated).

Zero Inflated Negative Binomial

Here is a histogram that shows the distribution of the data for the number of times each patient has visited a physician's office.



$\mu = 5.774$
$\sigma^2 = 45.6871$

The histogram demonstrates zero inflated data.
Normally for count data, we would use the Poisson distribution, but since $\mu \neq \sigma^2$ Poisson does not apply. This is why we are applying the **Zero Inflated Negative Binomial** distribution.

Reducing Covariates

When a data set has a lot of covariates, we need to reduce them in order to reduce the overhead cost of analyzing the data set. We do this using tests that determine the statistical significance of each covariate. The tests tell us whether the covariate affects the data in a meaningful way. If it does not, then that covariate can be removed.

Today, we will analyze which tests work best for our data, the **Likelihood Ratio Test** or the **Wald Test**

Wald Test vs Likelihood Ratio Test

- **The Wald Test** shows whether the parameters for the exploratory variables are zero. If they are, they can be removed from the model. If they aren't, then they are sufficiently significant and should stay in the model.
- **The Likelihood Ratio Test** runs on two nested models. This means that one model has a subset of variables of a larger model. It tests whether the larger model produces significantly better than the smaller one. If it does, then all variables should be included. If it does not, then the variables that are not in the smaller model can be removed.

Model

We are using the following model:

$$\ln(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

where

$$\beta_0 = 1, \beta_1 = 1$$

and x_1, x_2 follow the normal distribution.

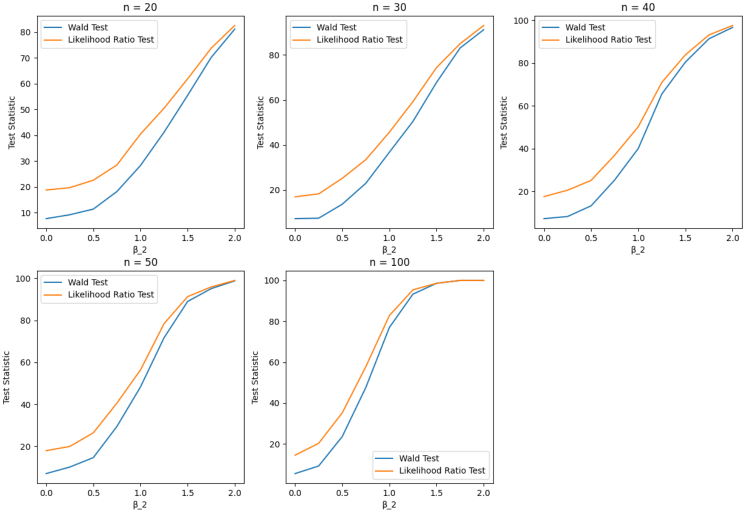
β_2 values are described in the following table.

Wald Test vs Likelihood Ratio Test

TABLE: Empirical level (EL) and power (in %) of the Wald and Likelihood ratio test statistics; $\alpha = 0.05$

Size (n)	Test	EL	Empirical Power							
			β_2							
		0.00	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00
20	Wald	7.7	9.20	11.40	18.20	28.30	41.20	55.50	70.20	81.20
	LR	18.8	19.70	22.60	28.50	40.50	50.50	61.90	73.80	82.60
30	Wald	7.30	7.50	13.70	23.00	36.80	50.50	67.80	83.10	91.20
	LR	17.00	18.30	25.20	33.50	45.70	59.30	74.40	85.00	93.10
40	Wald	7.30	8.30	13.30	25.40	40.10	65.60	80.50	91.40	96.70
	LR	17.60	20.60	25.20	37.00	50.30	71.00	83.90	93.20	97.60
50	Wald	7.10	10.20	14.70	29.50	48.30	71.70	89.00	95.10	98.80
	LR	18.00	20.00	26.50	40.60	56.40	78.40	91.30	95.90	99.00
100	Wald	5.50	9.20	23.60	47.60	77.00	93.30	98.60	100	100
	LR	14.50	20.30	35.20	57.80	82.80	95.40	98.60	100	100

Wald Test vs Likelihood Ratio Test



Forward Selection Procedure

The forward selection process is used to reduce covariates in the data to provide a more accurate model. We remove some of the more insignificant variables to get a better model, leading to better predictions.

We take the scores of the model by adding covariates in one at a time. After finding the minimum score for each test, we add said covariate and continues adding variables until the scores increase rather than decreasing.

Forward Selection Procedure Model

We are using the following model:

$$\ln(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

where

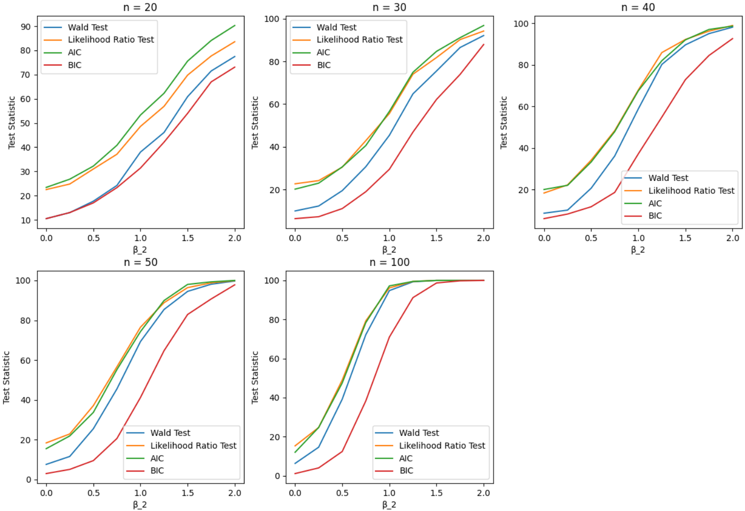
$$\beta_0 = 1, \beta_2 = -1, \beta_3 = 2, \beta_4 = -2$$

and x_1, x_2, x_3, x_4 follow the normal distribution.
β_1 values are described in the following table.

TABLE: Empirical level (EL) and power (in %) of model selection in zero-inflated Negative Binomial Regression Models

Size (n)	Method	EL	Empirical Power							
			β_1							
		0.00	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00
20	For-W	10.50	13.00	17.70	24.20	38.00	46.10	60.90	71.50	77.48
	For-L	22.50	24.80	31.00	37.10	48.60	56.90	69.80	77.70	83.60
	AIC	23.40	26.80	32.20	40.80	53.30	62.30	75.60	84.10	90.30
	BIC	10.50	13.00	17.00	23.30	31.40	42.10	54.00	67.00	73.10
30	For-W	10.00	12.30	19.50	30.70	45.40	64.80	75.50	86.50	92.10
	For-L	22.7	24.20	30.50	42.90	55.60	74.00	81.80	90.20	94.20
	AIC	20.20	23.00	30.60	40.60	56.60	74.90	84.70	91.10	96.80
	BIC	6.40	7.30	11.10	19.00	29.50	46.90	62.20	73.90	87.90
40	For-W	8.70	10.20	20.70	36.20	58.80	80.20	89.60	95.00	98.10
	For-L	18.40	22.30	34.20	48.40	67.90	85.90	92.20	96.20	98.90
	AIC	20.10	22.10	33.30	47.90	67.50	82.00	92.00	97.00	98.60
	BIC	6.10	8.30	11.80	18.70	37.20	54.90	72.90	84.50	92.60
50	For-W	7.60	11.60	25.50	45.50	69.30	85.40	94.50	98.10	99.70
	For-L	18.40	22.90	37.10	56.60	76.50	88.80	96.40	98.80	99.90
	AIC	15.50	21.90	33.70	55.20	74.30	89.90	98.00	99.30	100
	BIC	3.00	5.10	9.50	20.60	41.20	64.70	82.90	90.70	97.80
100	For-W	6.30	14.60	39.20	72.21	94.80	99.30	100	100	100
	For-L	15.30	24.70	49.00	79.30	96.20	99.50	100	100	100
	AIC	12.00	24.70	47.50	78.40	97.20	99.50	100	100	100
	BIC	1.10	4.00	12.40	38.30	71.00	91.20	98.70	99.80	100

Forward Selection Procedure



RESULTS

For small samples, the Wald test for forward selection proved to be the best.

For larger samples, BIC for forward selection steadily proved to be a better indicator of significance.

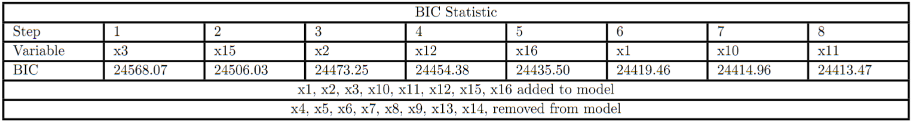
Application on Real Data

We will use forward selection to generate ZINB models from the 1988 NMES Survey.

Per the AER package documentation,

”The (US National Medical Expenditure Survey) NMES is based upon a representative, national probability sample of the civilian non-institutionalized population and individuals admitted to long-term care facilities during 1987 (through 1988). The data are a subsample of individuals ages 66 and over all of whom are covered by Medicare (a public insurance program providing substantial protection against health-care costs)”. The sample size of the dataset is $n = 4406$, which can be classified as a large dataset.

Application on Real Data



BIC Statistic								
Step	1	2	3	4	5	6	7	8
Variable	x3	x15	x2	x12	x16	x1	x10	x11
BIC	24568.07	24506.03	24473.25	24454.38	24435.50	24419.46	24414.96	24413.47
x1, x2, x3, x10, x11, x12, x15, x16 added to model								
x4, x5, x6, x7, x8, x9, x13, x14, removed from model								

Application on Real Data

After running the BIC algorithm for forward selection on the NMES data, we determined the following covariates to be the most significant for the model:

x1 = EXCLHLTH - 1 if self-perceived health is excellent
x2 = POORHLTH - 1 if self-perceived health is poor
x3 = NUMCHRON - num chronic illnesses
x10 = MALE - 1 if the person is male
x11 = MARRIED - 1 if person is married
x12 = SCHOOL - num years of education
x15 = PRIVINS - 1 if the person is covered by private healthcare insurance
x16 = MEDICAID - 1 if the person is covered by Medicaid

These variables best predict OFP - num physician office visits

Application on Real Data

We also determined the following covariates to be the least significant for the model:

x4 = ADLFDIFF - 1 if the person has a condition that limits daily activity

x5 = NOREAST - 1 if the person lives in the NE US

X6 = MIDWEST - 1 if the person lives in the MW US

x7 = WEST - 1 if the person lives in the W US

x8 = AGE - age in years, divided by 10

x9 = BLACK - 1 if the person is African American

x13 = FAMINC - family income in 10,000s

x14 = EMPLOYED - 1 if person is employed

REFERENCES

https://www.rdocumentation.org/packages/AER/versions/1.2-9/topics/NMES1988

http://users.stat.umn.edu/ sandy/courses/8053/Data/trevedi/debtrevedi.pdf

Thank You