MSA 8200 Predictive Analytics
Week 12: Regression Models for Count Data

Yichen Cheng

Spring 2020

Outline

- From continuous variables to discrete variables
- From lm (linear models) to glm (generalized linear model)
- Poisson Regression
- Overdispersion and Negative Binomial Distribution
- Zero-Inflation and Hurdle Model
- Applications and rootgram for model diagonostic

From continuous variables to discrete variables – Different types of discrete variables

Categorical

- Brand choice (Pepsi, Sevenup, Coke)
- different class labels (cat, dog, tiger)
- Logistic Regression; Discrete Choice Model

Ordinal – order of the values is meaningful

- Small, Medium, Large
- Count Data – Non-negative integers
- Ordered Choice Model; Models for Count data

Examples of Count Data

Examples of count data:

- Number of insurance Claims a person file per year.
- Number of hospital admissions/readmissions for a given person per year.
- Number of "jumps" (higher than 2σ) in stock returns per day.
- Number of a given disaster e.g., default, per month.

From lm to glm – Introduction

A linear model

$$y_i = \underline{x}_i^T \underline{\beta} + u \tag{1}$$

- The linear predictor $\eta_i = \underline{x}_i^T \underline{\beta}$.
- $y_i | \underline{x}_i \sim \mathcal{N}(\underline{x}_i^T \underline{\beta}, \sigma^2)$.
- $E(y_i | \underline{x}_i) \triangleq \mu_i = \eta_i$ - link function

From lm to glm – Introduction

A linear model

$$y_i = \underline{x}_i^T \underline{\beta} + u \quad (1)$$

- The linear predictor $\eta_i = \underline{x}_i^T \underline{\beta}$.
- $y_i | \underline{x}_i \sim \mathcal{N}(\underline{x}_i^T \underline{\beta}, \sigma^2)$.
- $E(y_i | \underline{x}_i) \triangleq \mu_i = \eta_i$ - link function

Only makes sense if $y|x$ is normal and y is continuous.

From lm to glm – Introduction

A linear model

$$y_i = \underline{x}_i^T \underline{\beta} + u \tag{1}$$

- The linear predictor $\eta_i = \underline{x}_i^T \underline{\beta}$.
- $y_i | \underline{x}_i \sim \mathcal{N}(\underline{x}_i^T \underline{\beta}, \sigma^2)$.
- $E(y_i | \underline{x}_i) \triangleq \mu_i = \eta_i$ - link function

Only makes sense if $y|x$ is normal and y is continuous.

The generalized linear models (GLM) extend it to a more general situation:

- $y_i | \underline{x}_i \sim$ a more general distribution group.
- The expected response and the linear predictor are linked by a monotonic transformation, $g(\mu_i) = \eta_i$.

From lm to glm – Example of GLM: logistic regression

$$y_i|\underline{x}_i \sim Bernoulli(p_i) \tag{2}$$

- $E(y_i|\underline{x}_i) = p_i = \mu_i$
- $log(\mu_i/(1-\mu_i)) = \eta_i = \underline{x}_i^T\underline{\beta}$

From lm to glm

Table 5.1. Selected GLM families and their canonical (default) links.

Family	Canonical link	Name
`binomial`	$\log\{\mu/(1-\mu)\}$	`logit`
`gaussian`	μ	`identity`
`poisson`	$\log\mu$	`log`

From lm to glm – Poisson distribution

Poisson distribution often used to:

- model the number of arrivals

Eg: an individual keeps track of the amount of mail they receive each day

- an average number of 4 letters per day
- the number of letters follow a Poisson distribution

Other examples:

- number of phone calls per hour by a call center

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{3}$$

Ref: https://en.wikipedia.org/wiki/Poisson_distribution

Motivating examples - RecreationDeamd

- Cross-section data of 2,000 registered leisure boat owners in 23 counties in eastern Texas.
- Dependent variable: number of recreational boating trips.
- Covariates:
  - quality ranking of the facility (quality),
  - whether the individual engaged in water-skiing at the lake (ski),
  - household income (income),
  - whether the individual paid a user's fee at the lake (userfee),
  - three cost variables (costC, costS, costH)

First Count Data Model – Poisson Regression

- $y_i | \underline{x}_i \sim Poisson(\lambda_i)$
- $E(y_i | \underline{x}_i) = \lambda_i = \mu_i = \exp(\eta_i) = \exp(\underline{x}_i^T \underline{\beta})$

Fitting the model in R:

```
R> data("RecreationDemand")
R> rd_pois <- glm(trips ~ ., data = RecreationDemand, family = poisson)
R> coeftest(rd_pois)
```

First Count Data Model – Poisson Regression

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.26499	0.09372	2.83	0.0047
quality	0.47173	0.01709	27.60	< 2e-16
skiyes	0.41821	0.05719	7.31	2.6e-13
income	-0.11132	0.01959	-5.68	1.3e-08
userfeeyes	0.89817	0.07899	11.37	< 2e-16
costC	-0.00343	0.00312	-1.10	0.2713
costS	-0.04254	0.00167	-25.47	< 2e-16
costH	0.03613	0.00271	13.34	< 2e-16

How does the model fit?

First Count Data Model – Poisson Regression

	predicted											
actual	0	1	2	3	4	5	6	7	8	9	10	>10
0	230	162	11	6	2	3	1	2	0	0	0	0
1	6	16	16	10	8	4	4	0	1	0	1	2
2	1	8	7	4	8	3	4	0	2	0	0	1
3	0	3	5	6	10	1	3	3	0	0	0	3
4	2	2	4	1	1	2	2	2	0	0	1	0
5	0	3	2	0	3	1	3	0	0	1	0	0
6	0	1	4	3	1	1	0	1	0	0	0	0
7	0	0	0	1	0	0	0	0	0	0	0	1
8	0	0	2	2	1	1	1	0	0	0	0	1
9	0	0	0	0	0	0	0	0	0	0	0	1
10	0	1	0	1	2	3	2	1	2	0	0	1
>10	0	2	1	3	3	3	4	2	3	3	2	11

First Count Data Model – Poisson Regression

actual \ predicted	0	1	2	3	4	5	6	7	8	9	10	>10
0	230	162	11	6	2	3	1	2	0	0	0	0
1	6	16	16	10	8	4	4	0	1	0	1	2
2	1	8	7	4	8	3	4	0	2	0	0	1
3	0	3	5	6	10	1	3	3	0	0	0	3
4	2	2	4	1	1	2	2	2	0	0	1	0
5	0	3	2	0	3	1	3	0	0	1	0	0
6	0	1	4	3	1	1	0	1	0	0	0	0
7	0	0	0	1	0	0	0	0	0	0	0	1
8	0	0	2	2	1	1	1	0	0	0	0	1
9	0	0	0	0	0	0	0	0	0	0	0	1
10	0	1	0	1	2	3	2	1	2	0	0	1
>10	0	2	1	3	3	3	4	2	3	3	2	11

- underestimated the number of >10 counts – overdispersion.
- underestimated the number of zeros – zero-inflation.

First Count Data Model – Poisson Regression: Overdispersion

Poisson distribution:

- $z \sim Poisson(\lambda)$
- $E(z) = Var(z) = \lambda$

Poisson Regression: $y_i | \underline{x}_i \sim Poisson(\eta_i)$

- $E(y_i | \underline{x}_i) = \mu_i = \exp(\eta_i)$
- $Var(y_i | \underline{x}_i) = \mu_i = \exp(\eta_i)$

For real data, often times, the variance will be greater than the mean.

Count Data Model – Poisson Regression: Overdispersion

Overdispersion:

- $Var(y_i|\underline{x}_i) = (1+\alpha)\mu_i = dispersion \cdot \mu_i$

Overdispersion test:

- H_0: $\alpha = 0$ vs. H_1: $\alpha > 0$.

R package "AER" provides the function: dispersiontest().

Count Data Model – Poisson Regression: Overdispersion

```
R> dispersiontest(rd_pois)

        Overdispersion test

data:  rd_pois
z = 2.412, p-value = 0.007941
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
     6.566
```

Count Data Model – Negativa Binomial Regression

Negative Binomial distribution allows us to model overdispersion.

$$f(y;\mu,\theta) = \frac{\Gamma(\theta+y)}{\Gamma(\theta)y!}\frac{\mu^y\theta^\theta}{(\mu+\theta)^{y+\theta}},\ \ y = 0, 1, 2, ... \tag{4}$$

- $Var(y;\mu,\theta) = \mu + \mu^2/\theta$
- $\mu_i = \exp(\eta_i)$

Count Data Model – Negativa Binomial Regression

actual \ predicted	0	1	2	3	4	5	6	7	8	9	10	>10
0	367	32	4	6	2	2	0	0	0	0	2	2
1	11	17	7	11	6	2	3	1	2	1	1	6
2	4	8	5	5	4	1	2	3	0	1	0	5
3	0	4	5	1	7	3	3	2	2	0	0	7
4	2	3	3	2	0	1	1	0	0	2	1	2
5	0	1	2	2	1	3	1	0	0	0	0	3
6	0	3	2	0	2	2	0	1	0	0	0	1
7	0	0	1	0	0	0	0	0	0	0	0	1
8	0	1	1	2	0	1	0	0	0	1	1	1
9	0	0	0	0	0	0	0	0	0	0	0	1
10	0	1	0	2	0	0	0	3	1	0	2	4
>10	1	1	1	0	2	3	4	1	3	1	0	20

- underestimated the number of zeros – zero-inflation.

Count Data Model – Zero-inflated data

Another common problem with count data: zero-inflation:

- Number of zeros much larger than a Poisson or NB regression allow.

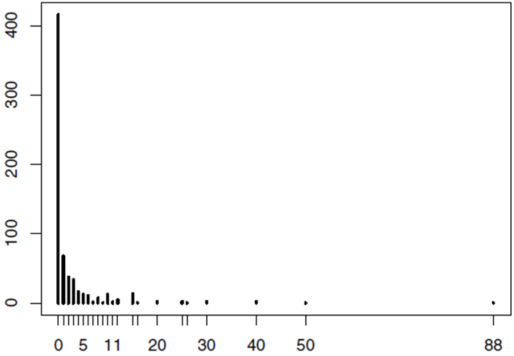

Fig. 5.3. Empirical distribution of trips.

Count Data Model – Zero-inflated data

zero-inflated models:

- A mixture of count model and point mass at zero:

$$f_{zeroinfl}(y) = p_i I_{\{0\}}(y) + (1 - p_i) f_{count}(y; \mu_i) \tag{5}$$

- $I_{\{0\}}(y)$ – indicator function
- f_{count} be a count distribution such as:
  - Poisson (ZIP)
  - NB
  - geometric, etc.

R> rd_zinb <- zeroinfl(trips ~ . | quality + income, data = RecreationDemand, dist = "negbin")

Count Data Model – Zero-inflated data

The output of the regression consists of two parts:

```
Count model coefficients (negbin with log link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.09663    0.25668    4.27  1.9e-05
quality      0.16891    0.05303    3.19  0.00145
skiyes       0.50069    0.13449    3.72  0.00020
income      -0.06927    0.04380   -1.58  0.11378
userfeeyes   0.54279    0.28280    1.92  0.05494
costC        0.04044    0.01452    2.79  0.00534
costS       -0.06621    0.00775   -8.55  < 2e-16
costH        0.02060    0.01023    2.01  0.04415
Log(theta)   0.19017    0.11299    1.68  0.09235

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    5.743      1.556    3.69  0.00022
quality       -8.307      3.682   -2.26  0.02404
income        -0.258      0.282   -0.92  0.35950
```

Hurdle model (Mullahy, 1986):

- More widely used than the zero-inflated models.
- Consists of two parts (also called “two-part model”)
    - A binary part: Is y_i 0 or positive? (Is the hurdle crossed?)
    - A count part: Given $y_i > 0$, how large is y_i?

The Resulting model:

- $P(y = 0) = f_{zero}(0; z, \gamma)$;
- $P(y|y > 0) = \{1 - f_{zero}(0; z, \gamma)\}\frac{f_{count}(y;x,\beta)}{1-f_{count}(0;x,\beta)}$

R> rd_hurdle <- hurdle(trips ~ . | quality + income, data = RecreationDemand, dist = "negbin")

Count Data Model – Zero-inflated data

The output of the hurdle model consists of two parts:

```
Count model coefficients (truncated negbin with log link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.8419     0.3828    2.20   0.0278
quality        0.1717     0.0723    2.37   0.0176
skiyes         0.6224     0.1901    3.27   0.0011
income        -0.0571     0.0645   -0.88   0.3763
userfeeyes     0.5763     0.3851    1.50   0.1345
costC          0.0571     0.0217    2.63   0.0085
costS         -0.0775     0.0115   -6.71  1.9e-11
costH          0.0124     0.0149    0.83   0.4064
Log(theta)    -0.5303     0.2611   -2.03   0.0423

Zero hurdle model coefficients (binomial with logit link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.7663     0.3623   -7.64  2.3e-14
quality        1.5029     0.1003   14.98  < 2e-16
income        -0.0447     0.0785   -0.57     0.57
```

Why the results are different from the previous ZIP model?

Count Data Model – Zero-inflated data

In summary: two methods for fitting zero-inflated data:

- zero-inflated model
- Hurdle model

Which one to choose:

- The hurdle model is somewhat easier to fit than the zero-inflation model
- because the resulting likelihood decomposes into two parts that may be maximized separately

Count Data Model – Applications

Data with excess zeroes and dispersion:

- Number of insurance Claims.
- Number of hospital admissions/readmissions.
- Number of vaccine adverse event.
- Number of âĂIJjumpsâĂİ (higher than 2σ) in stock returns per day
- Number of a given disaster e.g., default, per month.

Count Data Model – Zero-inflated data

https://data.library.virginia.edu/getting-started-with-hurdle-models/

obs	$p(y=0)$	$p(y=1)$
1	0.3	0.7
2	0.2	0.8
3	0.9	0.1
	1.4	1.6

Binary Case.

$$E(y_1) = 0.3 \times 0 + 0.7 \times 1 = 0.7$$

$$E(y_2) = 0.2 \times 0 + 0.8 \times 1 = 0.8$$

$$E(y_3) = 0.9 \times 0 + 0.1 \times 1 = 0.1$$

Count Data Model – Prediction

obs	$P(y=0)$	$P(y=1)$	$P(y=2)$	$P(y=3)$	Count Case.
1	0.3	0.2	0.1	0.4	$E(y_1)=$
2	0.1	0.1	0.1	0.7	$E(y_2)=$
3	0.5	0.2	0.1	0.2	$E(y_3)=$

Reference

https://bookdown.org/ccolonescu/RPoE4/qualitative-and-ldv-models.html#ordered-choice-models
https://www.econometrics-with-r.org/11-4-application-to-the-boston-hmda-data.html