# Bicycle Collision Counts Modelling in Downtown Seattle

**Bowen Xiao** and **Ezgi Irmak Yücel** and **Yiran Zhang**
STAT 504 Applied Linear Regression - Course Project

## Introduction

Our aim for this project is to select a proper regression model for bicycle collision data in downtown Seattle (01/2004-06/2017).
We are exploring the *accident counts (y)* observed for each combination of four categorical covariates:*light, road, weather condition and location (X).*

### Regression of Count Variables

- Linear regression fails to explain count variables as the residuals do not follow a normal distribution, there are couple of alternatives such as:

→ **Poisson Regression:**

$$log(E[y|x]) = \beta^T x$$

→ **Negative Binomial Regression:**

$$\mathrm{E}[Y_i|\mathbf{x}_i] = e^{\mathbf{x}_i^\top \boldsymbol{\beta}}$$
$$\mathrm{Var}[Y_i|\mathbf{x}_i] = \mu_i(1 + \mu_i/\theta)$$

## Data

### Visualization

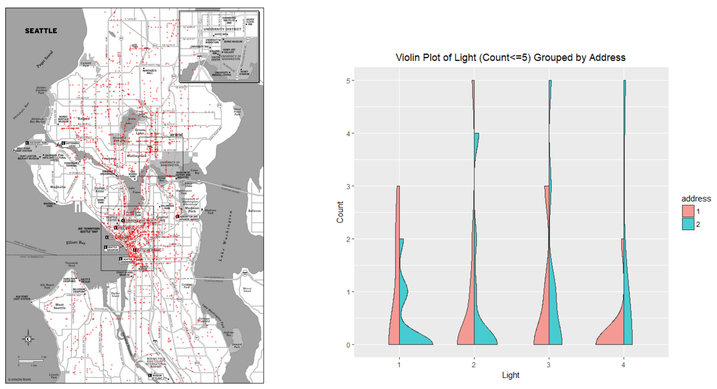



Fig. 1: Bicycle collisions across Downtown Seattle

### Factor Decoding

| Variable (Name in the SDOT Dataset) | Description | Values |
|---|---|---|
| Bicycle Collision Count (y) | Number of collisions per condition | 1: 0<br>2: 1-5<br>3: 6-10<br>4: 10+ number of collisions per condition |
| Address Type (ADDTYPE) | Whether the location is an intersection or a block | 1: Intersection<br>2: Block |
| Light Condition (LIGHTCOND) | Lighting condition of the road | 1: • Dark - No street Light<br>• Dark - Street Light On<br>• Dark - Street Light Off<br>2: Dawn<br>3: Daylight<br>4: Dusk |
| Road Condition (ROADCOND) | Whether road is dry or wet | 1: Dry<br>2: Wet |
| Weather Code (WEATHER) | Weather condition | 1: Raining<br>2: Overcast<br>3: Partly cloudy or Clear |

Tab. 1: Decoded variables used in our models

## Results

### Poisson Model

*m1=glm(accounts∼roadcode+weathercode+adcode+lightcode, data = df, family=poisson)*

| Selected Variable | P-value |
|---|---|
| ADDtype(block or intersection) | $< 2.2e\text{-}16$ * |
| Weather code | $< 2.2e\text{-}16$ * |
| Road condition Code | $< 2.2e\text{-}16$ * |
| Light code | $< 2.2e\text{-}16$ * |

Tab. 2: Significant covariates of Poisson Regression Model

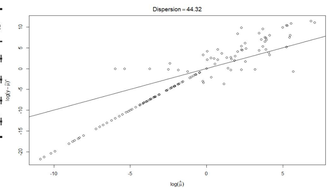



Fig. 2: Overdispersion plot for Poisson Regression

### Negative Binomial Model

Comparing with Poisson distribution, $\tau_i$ is introduced to Negative Binomial model.When $\tau_i \equiv e^{\epsilon_i}$ (unobserved heterogeneity term) is included and assumed to follow $Gamma(\theta, \theta)$, we get **Negative Binomial** distribution.

$$f(y_i|\mathbf{x}_i) = \frac{\Gamma(y_i + \theta)}{y_i!\Gamma(\theta)} \left(\frac{\theta}{\theta + \mu_i}\right)^{\theta} \left(\frac{\mu_i}{\theta + \mu_i}\right)^{y_i}.$$

Now the conditional mean and variance in NB model are:

$$\mathrm{E}[Y_i|\mathbf{x}_i] = \mathrm{E}[e^{\mathbf{x}_i^\top \boldsymbol{\beta} + \mathrm{ffl}_i}|\mathbf{x}_i] = e^{\mathbf{x}_i^\top \boldsymbol{\beta}}$$
$$\mathrm{Var}[Y_i|\mathbf{x}_i] = \mu_i(1 + \mu_i/\theta)$$

The model by using R is built as:

*nb=glm.nb(accounts∼roadcode+weathercode+adcode+lightcode,data=df)*
P-value of odTest is less than 2.2e-16, $\theta = 0.59$

| Model type | GOF | AIC | BIC |
|---|---|---|---|
| Poisson | 0 | 3245.30 | 3276.67 |
| Negative Binomial | 0.92 | 530.98 | 565.20 |

Table1. Comparison between Poisson and NB model

### LOWESS Lines

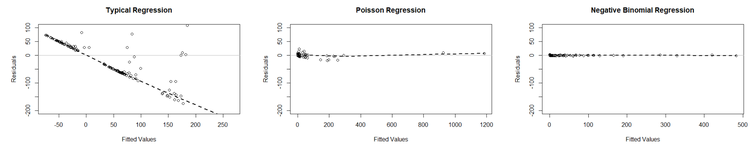
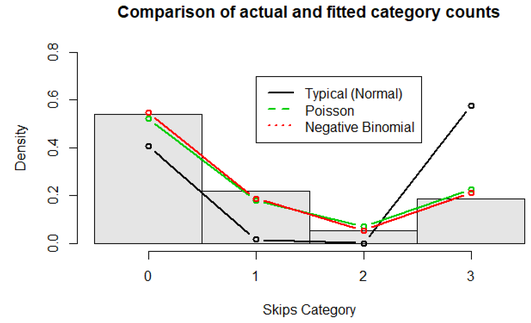



Fig. 4: There should be no relationship between these two values, so the LOWESS line should be horizontal and close to zero. The typical regression shows a obvious pattern in the results as the LOWESS line is slated downward. The other models all have horizontal LOWESS lines, with the negative-binomial model having the lowest range of residual values.

## Model Performances




| | | | Fitted Values | | |
|---|---|---|---|---|---|
| Outcome | Category | Actual | Typical | Poisson | Non Binomial |
| 0 | 0 | 67 | 52 | 67 | 70 |
| 1-5 | 1 | 28 | 2 | 23 | 24 |
| 6-10 | 2 | 7 | 0 | 9 | 7 |
| >10 | 3 | 24 | 74 | 29 | 27 |

Fig. 5: Comparison of Regression Models, when overlayed on the actual count data, linear regression overestimates collision counts larger than 10, where the Poisson and Negative Binomial Models can explain the data successfully.

## Discussion

### Findings:

We have found that:

- linear regression model fails to explain count data -our outcome, number of bike collisions-.
- Poisson regression model is overdispersed but can explain the data better than linear regression model.
- Negative Binomial model has the best performance of explaining the bicycle collision dataset.
- on average, controlling for other covariates, more bicycle collisions happen:
  - → at interactions than in blocks,
  - → on dry roads than other road conditions,
  - → on cloudy days than other conditions,
  - → during daylight than other conditions.

### Future Directions:

We can further our current model using Rate model - $log(E[\frac{y}{t}|x])$ or Spatial-Temporal model where we take temporal changes into account as well.

## References

- Cramér, Harald (1946). Mathematical Methods of Statistics. Princeton, NJ: Princeton Univ. Press. ISBN0-691-08004-6. OCLC185436716
- S.S. Wilks (1938). The large-sample distribution of the likelihood ratio for testing composite hypothesis. Annals of Mathematical Statistics 9. pp. 60-62.
- Moran, P.A.P (1971). Maximum likelihood estimation in non-standard conditions. Proc. Cambridge Philos. Soc., 70,441 -450.