Bicycle counts prediction in Downtown Seattle

Group 13: Bowen Xiao, Ezgi Irmak Yucel, Yiran Zhang

2018-2-3

1.Research question

The aim of our project is to identify and relate possible covariates for the number of bicycle collisions in downtown Seattle. More specifically we want to *infer* the relationship between the outcome i.e. accident counts (y); and covariates, light condition (eg: dark, down or daylight), the road condition(eg: wet, dry or ice), weather, and the location type (eg: Intersection or Block) (X).

2. Data description

The data is obtained from SDOT (Seattle Department of Transportation)[1] which has been collected in 2017. The raw data have GPS information for each accident and the collision location can be viewed in the map based on their coordinate[2]. There are 4812 records in the raw dataset.

In this project, all the selected independent variables are categorical data and the reponse in the count data. Hence, the coveriates are decoded and the details is shown in Table 1.

Table 1: Variables used in our model

Variable (Name in the SDOT Dataset)	**Description**	**Decoding**
Address Type (ADDTYPE)	Whether the location is an intersection or a block	1: Intersection 2: Block
Light Condition (LIGHTCOND)	Lighting condition of the road	1: Dark • No street Light • Street Light On • Street Light Off 2: Dawn 3: Daylight 4: Dusk
Road Condition (ROADCOND)	Whether road is dry or wet	1: Dry 2: Wet 3: Ice/Snoe/Slush 4: Sand/Mud/Dirt
Weather Code (WEATHER)	Weather condition	1: Raining 2: Unusual • Fog/Smog/Smoke • Sleet/Hail/Freezing Rain • Snowing • Blowing Sand or Dirt or Snow 3: Overcast 4: Partly cloudy or Clear

3. Methodology

3.1 Poisson Regression

Count variables share certain properties, for example, they are always non-negative integers and they frequently appear to be positively skewed. Moreover, when using linear regression model, the residual variance turns to increase as the predicted value increases, which produces heteroscedasticity. Thus, typical regression model fails here.

The most common alternative model is Poisson regression model. More specifically, the count variable is assumed to be independently Poisson distributed with the parameter:

$$log(E[y|\mathbf{X}]) = \beta\mathbf{X} \tag{1}$$

where X is the design matrix. The maximum likelihood estimators are obtained by using iterative algorithms.

3.2 Z test

In statistics, Z value is the signed number of standard deviations by which the value of an observation or data point is above the mean. The Z value for parameters of Poisson regression are approximately normally distributed for large samples, so that they can provide an asymptotic Z test - the analog of the Student's t-test in linear regression.

3.3 Deviance test

The deviance statistic - the analog of F statistic is:

$$D(\mathbf{y}, \hat{\mu}) = \sum_{i=1}^{n} 2(log(p(y_i|\hat{\theta}_s)) - log(p(y_i|\hat{\theta}_0))) \tag{2}$$

In goodness-of-fit test, $\hat{\theta}_0$ denotes the fitted values of the parameters in the model to be tested, and $\hat{\theta}_s$ denotes a perfect model. In likelihood-ratio test, $\hat{\theta}_0$ denotes the fitted values of the parameters in the null model, and $\hat{\theta}_s$ denotes the fitted values of the parameters in the alternative model. D is asymptoticly chi-square distributed. And Large values mean that the data do not agree well with the assumed/proposed model[3].

3.4 Overdispersion

Poisson regression inherently asks for a same mean and variance for predicted variable, which could be too strict sometimes. When observed variance is larger than the assumed variance, overdispersion raises. In this case, we will consider use negative binomial regression instead, in which a larger variance is allowed.

Reference

[1] "Bicycle Collision Data | City of Seattle Open Data Portal." Seattle, data.seattle.gov/Transportation/bicycle-collision-data/4nyx-mzt8/data.

[2] Layer: Collision (ID: 51), gisrevprxy.seattle.gov/arcgis/rest/services/SDOT_EXT/DSG_datasharing/MapServer/51.

[3] S.S. Wilks, 1938. The large-sample distribution of the likelihood ratio for testing composite hypothesis. Annals of Mathematical Statistics 9. pp. 60-62.