Bicycle Collision Counts Modeling in Downtown Seattle

Group 13

Bowen, Ezgi, Yiran

Outline for Section 1

Introduction

The aim of our project is to select a proper model

for bicycle collisions in downtown Seattle (01/2004-06/2017).

We are going to explore the relationship

between *accident counts (y)* and covariates:

light, *road*, *weather and location (X)*.

There are **4812** records in the raw dataset.

And **176** rows with missing data are removed.

Covariate Decoding

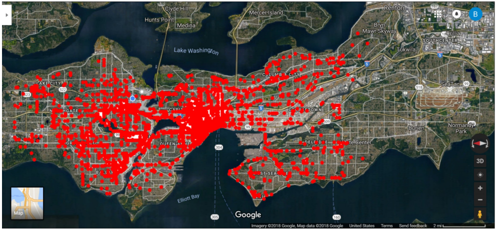
We are considering four categorical predictors.

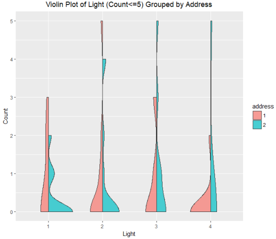
Table 1: Variables used in our model

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

Quick Look at the Data

Visualization



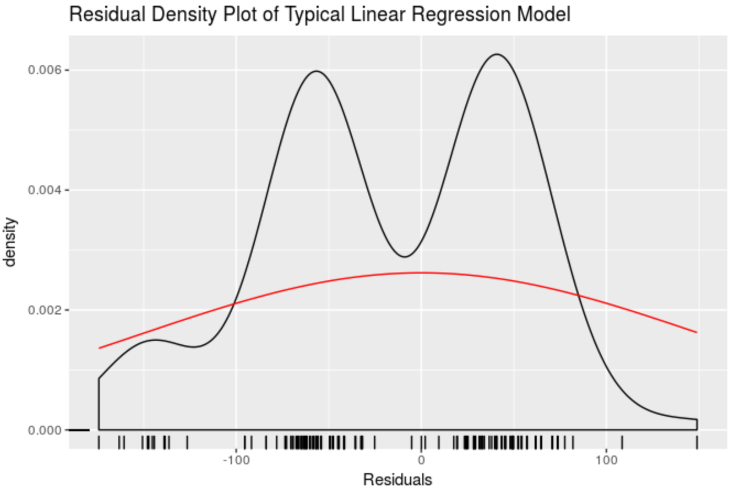


Quantile

0%	25%	50%	75%	100%
0	0	0	5	1445

Linear Regression

$Typical <- lm(accounts \sim ., data = df)$



Outline for Section 2

Poisson Regression

Count variables share certain properties and linear regression always fails here. The most common alternative model is Poisson regression model:

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

Or

$$\lambda \equiv E[y|x] = e^{\beta^T x},$$

$$p(y|x, \theta) = \frac{\lambda^y}{y!} e^{-\lambda} = \frac{e^{y\beta^T x} e^{-e^{\beta^T x}}}{y!}$$

Maximum Likelihood Estimator

A likelihood function in terms of β could be written as,

$$L(\beta|\mathbf{X},\mathbf{Y}) = \prod_{i=1}^{n} \frac{e^{y_i\beta^T x_i} e^{-e^{\beta^T x_i}}}{y_i!}$$

$$l(\beta|\mathbf{X},\mathbf{Y}) = logL(\beta|\mathbf{X},\mathbf{Y}) = \sum_{i=1}^{n} (y_i\beta^T x_i - e^{\beta^T x_i}) + const$$

The negative log-likelihood is convex and maximum likelihood estimators could be obtained by standard convex optimization techniques.

Model Quality

- Significance of one single predictor controlling for others.
- Overall goodness of fit.
- Comparison between two models.
- Model diagnostics.
- Significance of interactions.

Hypothesis Test

Z test - the analog of the Student's t-test

$$\sqrt{n}(\hat{\beta}_{MLE} - \beta) \xrightarrow{d} N_{p+1}(0, [I(\beta)]^{-1})$$

Where $I(\beta)$ is Fisher information matrix.

Likelihood ratio test - the analog of the F test

$$\sum_{i=1}^{n} 2(log(P(y_i|\hat{\beta}_A)) - log(P(y_i|\hat{\beta}_N))) \xrightarrow{d} \chi^2_{d_A - d_N}$$

odTest

Hypothesis test between Poisson and Negative Binomial model

Motivation

When sample variance is significantly larger than sample mean, we will consider use negative binomial regression instead, in which a larger variance is allowed.

Correction of asymptotic distribution

$$\sum_{i=1}^{n} 2(log(P(y_i|\hat{\beta}_A)) - log(P(y_i|\hat{\beta}_N))) \xrightarrow{d} 0.5 * 0 + 0.5 * \chi_1^2$$

Composite Index

View of information theory

AIC/BIC offers an estimate of the relative information lost and it suggests that we should prefer a model with a smaller AIC/BIC.

AIC & BIC

$$AIC = 2(p+1) - 2log(ML)$$
$$BIC = log(n) * (p+1) - 2log(ML)$$

Outline for Section 3

Poisson Model

Test of covariates

Selected Variable	Deviance	P-value
ADDtype(block or intersection)	71.252	$< 2.2e\text{-}16 ***$
Weather code	5751.8	$< 2.2e\text{-}16 ***$
Road condition Code	8347	$< 2.2e\text{-}16 ***$
Light code	6078.4	$< 2.2e\text{-}16 ***$

Hence, each of the selecting independent variables is significant, controlling for other covariates.

Poisson Model

The Poisson Model built in this project is shown as:

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

Variables	Estimate	Std.Error	z-value	p-value
(Intercept)	3.82798	0.05501	69.584	$< 2e\text{-}16$ ∗∗∗
roadcode2	-1.54712	0.03867	-40.006	$< 2e\text{-}16$ ∗∗∗
roadcode3	-5.68044	0.27782	-20.446	$< 2e\text{-}16$ ∗∗∗
roadcode4	-7.55224	0.70722	-10.679	$< 2e\text{-}16$ ∗∗∗
weathercode2	-4.67470	0.44929	-10.405	$< 2e\text{-}16$ ∗∗∗
weathercode3	0.24820	0.05763	4.307	1.66e-05 ∗∗∗
weathercode4	1.84975	0.04647	39.809	$< 2e\text{-}16$ ∗∗∗
adcode2	-0.24890	0.02960	-8.409	$< 2e\text{-}16$ ∗∗∗
lightcode2	-2.24723	0.11023	-20.387	$< 2e\text{-}16$ ∗∗∗
lightcode3	1.40214	0.03804	36.857	$< 2e\text{-}16$ ∗∗∗
lightcode4	-1.53774	0.08104	-18.976	$< 2e\text{-}16$ ∗∗∗

Model Quality

Overall Goodness of Fit

pchisq(deviance(m1),df=df.residual(m1),lower=F)=0

Diagnostic

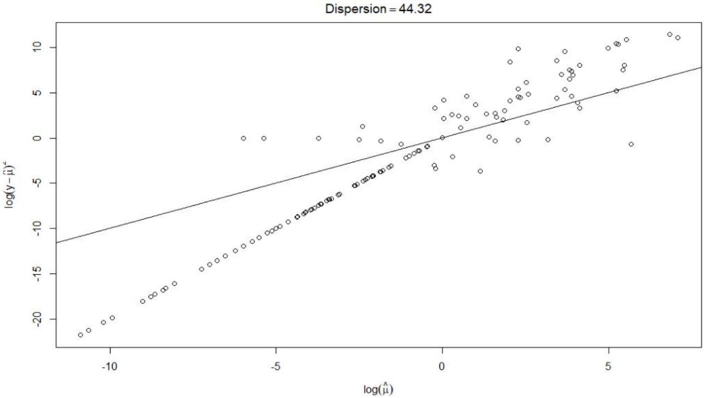
R code: $dp <- sum(residuals(m1, type = "pearson")^2)/m1\$df.res$

If dp is greater than one, the data shows signs of overdispersion.

In this Project, **dp = 44.32**, so it is overdispersion.

Overdispersion

Overdispersion plot



Solution Methods

More parameters

Interaction between weather and light condition, weather and road condition

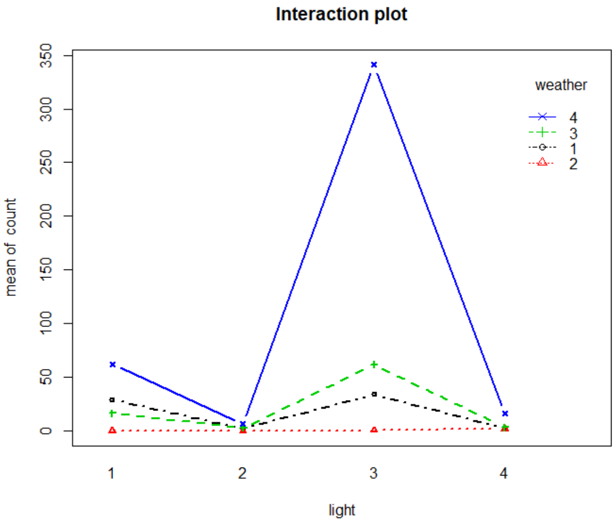
Considering the 0-value of response impact

Zero-Inflated Model

Model alternation

Negative Binomial Model

Interaction



Interaction

Hence, the model is built as:

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

dp = 1.3524, Overall Goodness of Fit: 0.4441

Hence the model has been improved.

Zero-inflated model

Since many 0-value exists from the dataset, we can use a logistic regression to recognize 0 (Precision-0.84):

logi <- glm((accounts==0)∼.,family=binomial(link='logit'),data=df)

Finally, our model is built as:

zeroinfl(accounts∼adcode+weathercode+roadcode+lightcode |adcode+weathercode+roadcode+lightcode,data=df,EM=TRUE)

dp = 37.35

Negative Binomial Regression

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:

$$\mathrm{E}[Y_i|\mathbf{x}_i] = \mathrm{E}[\mathrm{e}^{\mathbf{x}_i^\top \beta + \epsilon_i}|\mathbf{x}_i] = e^{\mathbf{x}_i^\top \beta}$$

$$\mathrm{Var}[Y_i|\mathbf{x}_i] = \mu_i(1 + \mu_i/\theta)$$

NB model

The NB model is built as:

nb=glm.nb(accounts∼roadcode+weathercode+adcode+lightcode,data=df)

P-value of odTest is less than 2.2e-16, $\hat{\theta} = 0.59$

Variables	Estimate	Std.Error	z-value	p-value
(Intercept)	4.1104	0.4134	9.943	$<$ 2e-16 $***$
roadcode2	-0.3398	0.3091	-1.099	0.271648
roadcode3	-5.0492	0.4861	-10.388	$<$ 2e-16 $***$
roadcode4	-6.7232	0.7713	-8.717	$<$ 2e-16 $***$
weathercode2	-4.6370	0.6169	-7.517	5.59e-14 $***$
weathercode3	0.0393	0.3718	0.106	0.915821
weathercode4	1.2973	0.3593	3.610	0.000306 $***$
adcode2	-0.3534	0.2845	-1.242	0.214119
lightcode2	-2.1905	0.4258	-5.144	2.69e-07 $***$
lightcode3	0.8188	0.3695	2.216	0.026679 $*$
lightcode4	-1.9047	0.4194	-4.541	5.59e-06 $***$

NB model

Comparison with Poisson model

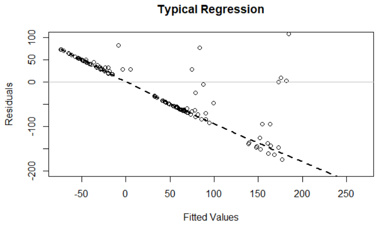
Model type	GOF	AIC	BIC
Poisson	0	3245.3	3276.672
Negative Binomial	0.9188	530.9798	565.2041

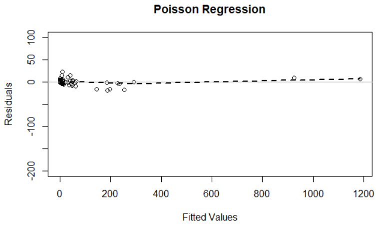
Test of covariates

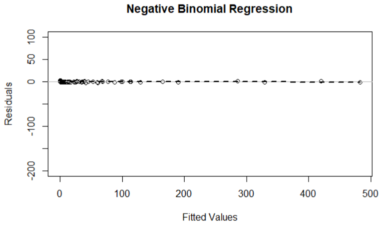
Selected Variable	Deviance	P-value
ADDtype(block or intersection)	1.4985	0.2209
Weather code	59.922	6.108e-13 $***$
Road condition Code	106.76	$<$ 2.2e-16 $***$
Light code	48.814	1.429e-10 $***$

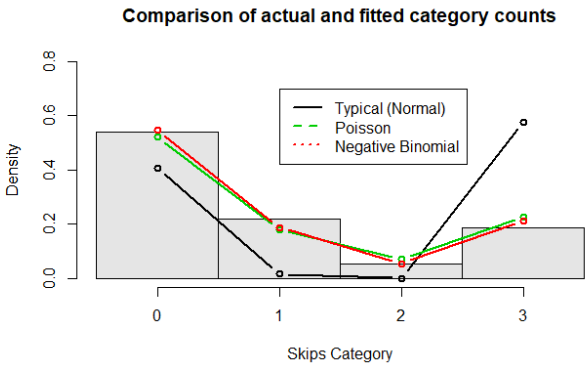
Outline for Section 4

LOWESS Lines









			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

Conclusion

- Linear regression model fails to explain count data.
- Poisson regression with interactions and Negative Binomial model can describe the data successfully.
- We also build NB model with interactions and Zero-inflated NB model, whose results are similar with the above.
- Future Directions:
  - Rate model - $log(E[\frac{y}{t}|x])$
  - Spatial-Temporal model

Reference

- 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.