Linear Regression Model of Traffic Collisions in Toronto Neighbourhoods

Yutong Lu - 1005738356

12/17/2021

Introduction

Traffic safety has called for more attention in recent years with increasing global population and motorization (World Health Organization, 2018). Road volume comes up frequently as a factor in accidents, and studies suggest that traffic volume may be related to traffic collisions (Wier et al., 2009; Xu et al., 2018). Interestingly, both traffic and sociological characteristics may affect collision frequency, where Park and Ko (2020) incorporated neighbourhood-level features such as population proportion in their vehicle-pedestrian collision model. Thus, our research question is what road, demographic and economic characteristics in Toronto neighbourhoods may be linearly related to traffic collisions. This study aims to build a descriptive linear regression model that is easier to interpret for traffic collisions in Toronto. Therefore, it is a critical study that may guide urban and transportation planning in targeted, cost-efficient city improvement, thus reducing traffic collisions.

Methods

The response variable was the number of traffic collisions based on the research question. Training and test sets were split equally before any analysis, and model building was based entirely on the training set. All potential predictors were in the initial model. Model diagnostics were performed by first assessing two conditions using scatter plots. We expected the observed response to be a single function of the fitted values, preferably the identity function. No non-linear relationships were expected between each pair of predictor variables. Transformations would be applied to the response variable, predictors, or both, depending on the shown relationships in the plots. Transformed variables were included in a linear regression model. Linear regression assumptions were assessed using standardized residual plots and Normal quantile-quantile plots. Random patterns in the standardized residual plots and an identity function in the quantile-quantile plot would indicate satisfied assumptions. Different transformations may be tried to satisfy the conditions and assumptions to the most extent.

With satisfied model assumptions, if there were insignificant predictors in the full model summary based on the T-tests and the significance level of 0.05, we would build a reduced model with those insignificant predictors removed. Model assumptions were assessed for this reduced model, and if all assumptions were satisfied, a partial F test would be used to provide evidence for removing predictors.

We then obtained models of each possible size with the highest adjusted coefficient of determination (R^2_{adj}). Models with violated assumptions were omitted because most parts of the model selection relied heavily on the assumptions. Then, we compared the variance inflation factor (VIF), where any predictors with VIFs greater than 5 indicated severe multicollinearity. VIFs close to 1 were preferred because they indicated little to no variance inflation due to relationships between predictors, resulting in a stabler regression surface. We also performed partial F tests between the full and reduced models. If we rejected the partial F test's null hypothesis, we would not have evidence to drop those variables and would omit the reduced model. Also, models with higher R^2_{adj} were preferred because R^2_{adj} would only increase when the additional predictors improve the model significantly.

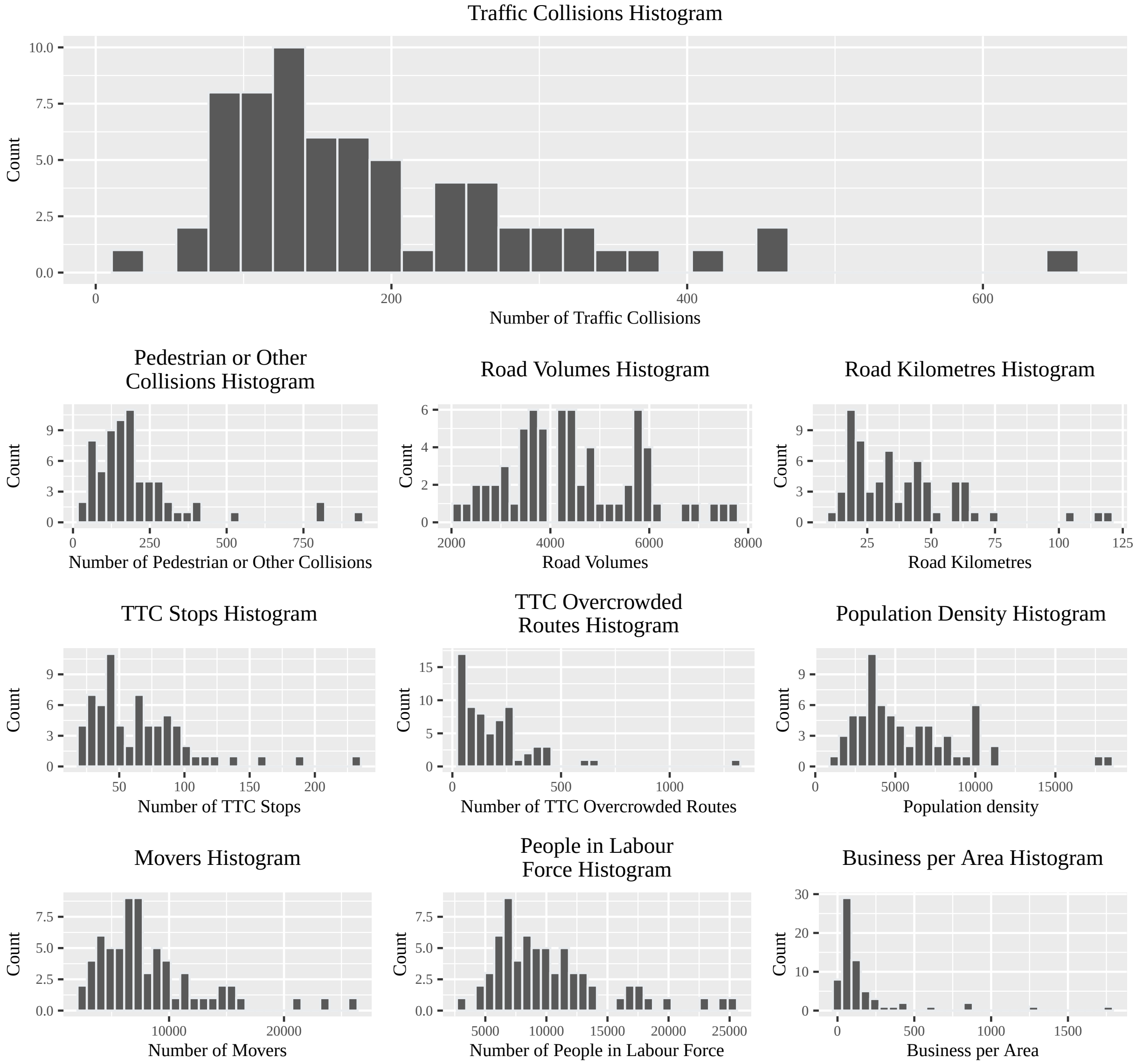
For additional justification, eight different automated forward, backward and stepwise selections based on Akaike information criterion (AIC) or Bayesian information criterion (BIC) were performed. Furthermore, we identified the leverage points, outliers, and influential observations using Cook's distance, difference in fits (DEFITS), and difference in betas (DFBETAs). Models with fewer problematic observations were preferred due to their potentials to affect the regression line.

Based on model assumptions, R^2_{adj}, multicollinearity, problematic observations, information criteria, and literature evidence, one or a few best models were selected for validation and fitted to the test set. Estimated coefficients, significant predictors, assumptions, and R^2_{adj} were compared between the training and test set models to evaluate the validation. If no discernible differences could be spotted, the model was considered validated and thus selected as the final model. However, if we could not validate the model, appropriate diagnostics would be used to address the possible reasons. Suppose there were multiple validated models, or none of the selected models could be validated. In that case, literature evidence and all aspects of the model diagnostics would be considered to choose the final model.

Results

The dataset from Toronto Open Data had six existing neighbourhoods with zero road volume and kilometres, which were considered as missing values and thus removed from the dataset (Wellbeing Toronto, 2014). The cleaned dataset had 134 observations corresponding to Toronto neighbourhoods. A training set of 67 observations was randomized and used for model building. Traffic collision appeared to be approximately linearly related to each potential predictor in the training set, according to Appendix Figure 3. However, histograms in Figure 1 show right-skewed distributions for all variables besides road volume, leading to potential linearity violations. The skewed distributions are consistent with training set statistics in Table 3, where all variable means are higher than medians. In the variable names, movers are people who did not stay at the same location over the past five years, and TTC provides public transit in Toronto.

Figure 1: Histograms of all variables in the training set with 67 observations.



Note: All histograms except for road volume are heavily right−skewed, leading to potential linearity violations in the model.

Power transformations were applied to both response and predictor variables to satisfy conditions and assumptions. There were insignificant predictors in the transformed model summary, and we had evidence to keep the reduced model without those predictors based on satisfied assumptions and partial F test. This reduced model had 4 predictors, with the only insignificant predictor being road volume backed up by literature evidence (Wier et al., 2009; Xu et al., 2018).

All models with the highest R^2_{adj} of each size had acceptable model assumptions. Table 1 shows that only models with sizes 3 to 5 had no severe multicollinearity and no rejected partial F test null hypothesis, where the 5-predictor model had the highest R^2_{adj} among them. Models with 4 and 3 predictors did not contain the variable road volume. As a result, we selected the 4-predictor model with slightly lower R^2_{adj} obtained in the previous step (4* in Table 1) and the 5-predictor model and denoted them as Model 4 and Model 5.

Table 1: Model selection table.

Number of predictors	Adjusted R squared	Severe multicollinearity (VIF > 5)	Partial F test null hypothesis rejected
1	0.5962	False	True
2	0.665	False	True
3	0.7284	False	False
4	0.7384	False	False
4*	0.7377	False	False
5	0.7509	False	False
6	0.7558	True	False
7	0.7556	True	False
8	0.752	True	False
9	0.7482	True	-

Note:
This table presents some characteristics of the models with the highest adjusted R squared for each size and one additional model from the last step of partial F test. The model with 9 predictors is the full model and thus no partial F test is performed for this model. 4* denotes the reduced model obtained after the first inspection of transformed model.

When re-examining assumptions, both models had satisfied conditions and random patterns in residual plots. Figure 2 shows that Model 5 aligned with the Q-Q line more smoothly and thus may satisfy the Normality assumption better. Similar observations appeared problematic in both models, but automated selection procedures gave the same models with 3, 5 or 6 predictors in Table 1, which further justified for Model 5. Therefore, we selected Model 5 as the final model, with the response and predictors given in Table 2.

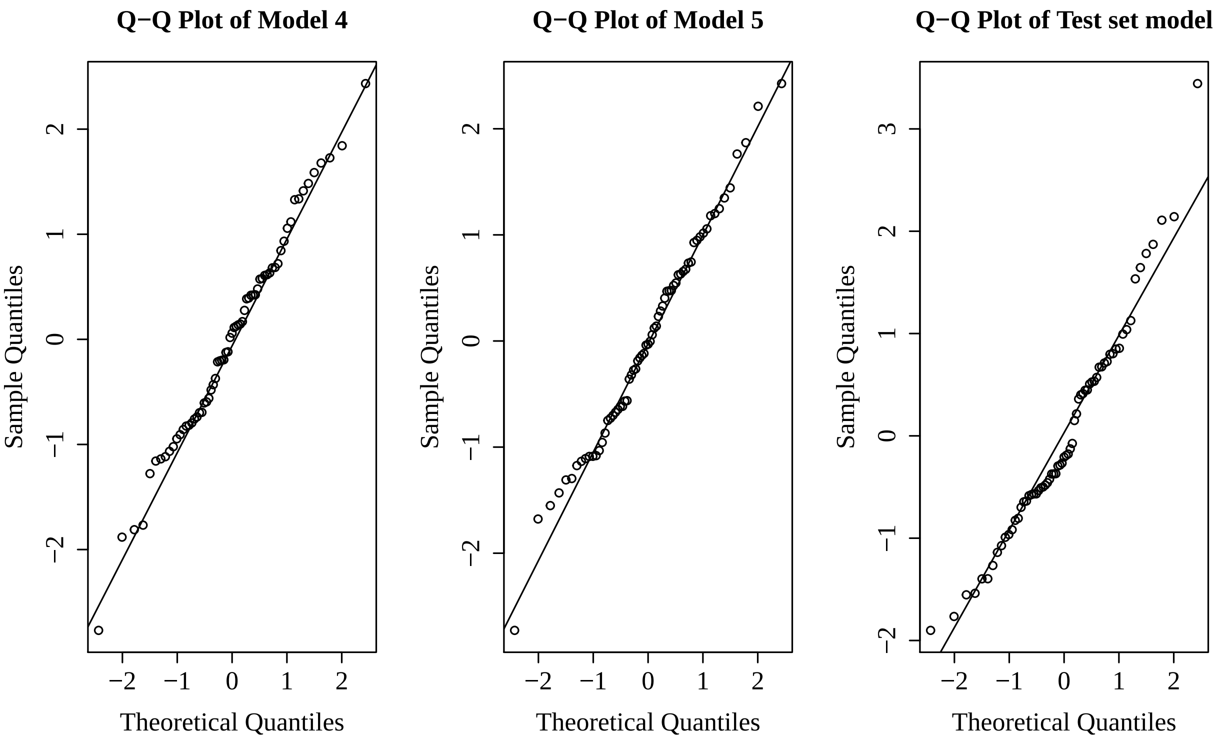
Table 2: Model summary for the final model with five predictors, fitted on the training set.

Response variable: Square root of number of traffic collisions			
	Estimated coefficient	Standard error	p-Value
Intercept	-14.4471	2.8949	0.0000
Cubic root of number of people in labour force	1.0575	0.0910	0.0000
Tenth root of business per area	9.4073	1.9643	0.0000
Square root of population density	-0.1129	0.0187	0.0000
Road Volume	0.0004	0.0002	0.0471
Cubic root of number of TTC overcrowded routes	-0.4588	0.2215	0.0426

Note:
This is the model summary for the final transformed model fitted on the training set and there is no insignificant predictors using a significance level of 0.05.

The final model was fitted to test data for model validation. Although the estimated coefficients and R^2_{adj} values were similar based on Appendix Table 4, two significant predictors in Model 5, road volume and cubic root of TTC overcrowded routes, were insignificant in the test set model. Conditions and residual plots showed no unexpected patterns, but the Normality assumption was violated for the test set model in Figure 2. Thus, we failed to validate the final model on the test set.

Figure 2: Normal Q-Q plots comparison for Model 4, Model 5 and test set model.



Note: Model 4 and 5 were fitted on the training set and the test set model was fitted on the test set. Model 5 aligned with the Q-Q line more smoothly than Model 4. Severe deviation from the Q-Q line could be spotted in the Q-Q plot for test set model.

Table 3: Training and test sets summary statistics and influential points using DFBETAs.

	Mean		Standard Deviation		Median		Number of influential points for corresponding coefficient in the models	
	training	test	training	test	training	test	training	test
Traffic collisions (n)	191	176.9552	110.9874	125.2349	161	139	-	-
Pedestrian or other collisions (n)	208.1045	229.6418	169.5587	242.3805	169	113	-	-
Road kilometres	38.806	37.9403	22.471	25.2274	33	30	-	-
Road volume	4462.3284	4673.9552	1319.0129	2264.2573	4301	4116	4	3
TTC stops (n)	66.2985	67.9403	39.0986	53.1943	60	49	-	-
TTC overcrowded routes (n)	198.7015	193	191.9246	272.4739	148	119	8	4
Population density	5702.7511	5622.177	3379.5878	5230.039	4701.6129	4631.25	5	2
Business per area	179.3045	165.7974	290.6123	290.7999	86.3014	94.2157	4	4
Movers (n)	8309.8507	7044.403	4750.4646	3880.3113	7060	6275		
People in labour force (n)	10314.403	9198.2836	4728.4248	4233.3965	9340	7950	5	3

Note:
Summary statistics of both training and test set are similar, but there are more influential observations on the esitmate coefficients in the training set. (n) denotes 'counts'. Each training and test set has 67 observations.

Discussion

Due to the monotonicity of power transformations, we could interpret that the increase in labour force, business per area and road volume may lead to more traffic collisions in the presence of other predictors. It makes sense since more commuters resulting from a greater employed population may lead to more collisions on the road, which is also consistent with the reported relationship between employed population and vehicle-pedestrian collisions (Wier et al., 2009). Conversely, greater population density and more overcrowded TTC routes may result in fewer traffic collisions when conditioning on other variables, supported by the literature relationship between population density and crashes (Park & Ko, 2020). Simultaneously, more overcrowded TTC routes may infer public transformation being a popular choice in the neighbourhood, potentially reducing traffic collisions. Therefore, the final model may provide insights into the factors related to traffic collisions. The results are important because they suggest that more convenient public transportations should be provided for commuters, and roads with higher volumes should require more attention or potential diversion to reduce the number of traffic collisions.

Limitations of this study include the removed observations with missing values, which may lead to biases in our resulting model. Also, multicollinearity was still present, although all VIFs were below 3 for both models using training and test sets, leading to a slightly unstable regression surface. Furthermore, the final model failed to be validated on the test set. In Table 3, the training set had more influential observations on estimated coefficients despite the similar summary statistics in both sets. Randomization before the analysis led to different problematic observations in training and test sets, which may result in disproportionate effects on the regression lines and the failed validation. We also saw more significant predictors in Model 5, so another reason may be the overfitting of the training set caused by overly specific transformations. Therefore, although the final model meets the goal of being interpretable and descriptive, it may not perform well on unseen data and thus have a weaker predictive ability.

Reference List

Park, S. & Ko, D. (2020). A multilevel model approach for investigating individual accident characteristics and neighborhood environment characteristics affecting pedestrian-vehicle crashes. *International Journal of Environmental Research and Public Health, 17*(9), 3107. https://doi.org/10.3390/ijerph17093107

Wellbeing Toronto (2017). *Wellbeing Toronto - Demographics* [dataset]. City of Toronto Open Data. https://open.toronto.ca/dataset/wellbeing-toronto-demographics/

Wellbeing Toronto (2014). *Wellbeing Toronto - Economics* [dataset]. City of Toronto Open Data. https://open.toronto.ca/dataset/wellbeing-toronto-economics/

Wellbeing Toronto (2014). *Wellbeing Toronto - Transportation* [dataset]. City of Toronto Open Data. https://open.toronto.ca/dataset/wellbeing-toronto-transportation/

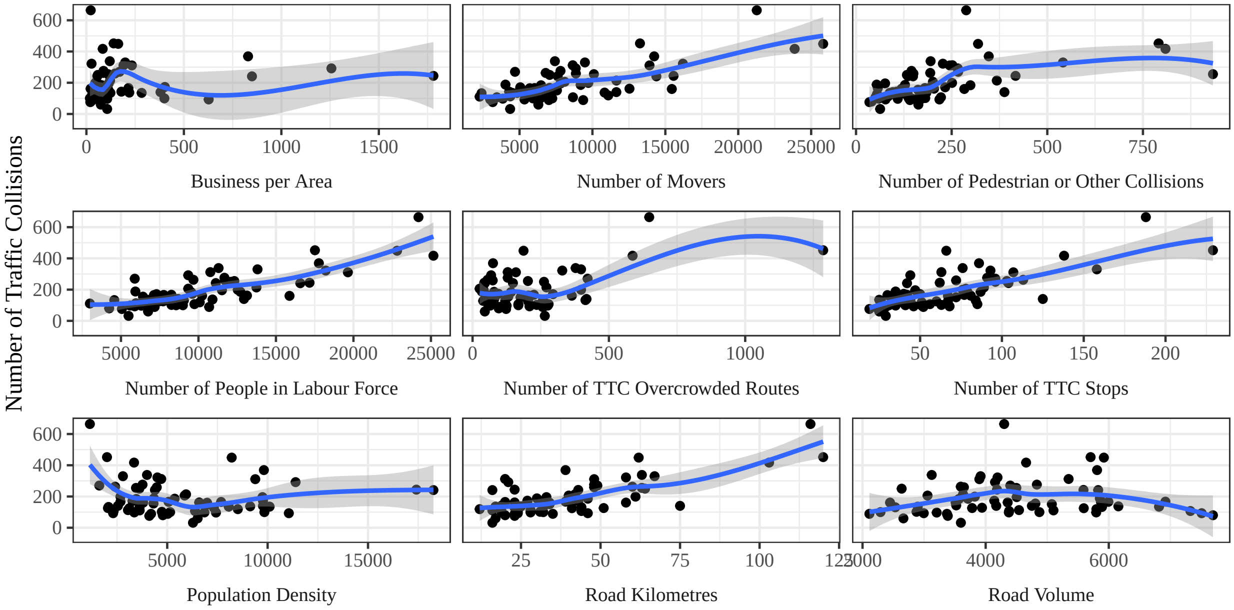
Wier, M., Weintraub, J., Humphreys, E. H., Seto, E., & Bhatia, R. (2009). An area-level model of vehicle-pedestrian injury collisions with implications for land use and transportation planning. *Accident Analysis and Prevention, 41*(1), 137–145. https://doi.org/10.1016/j.aap.2008.10.001

World Health Organization. (2018). *Global status report on road safety 2018: summary* (No. WHO/NMH/NVI/18.20). World Health Organization.

Xu, C., Wang, Y., Liu, P., Wang, W., & Bao, J. (2018). Quantitative risk assessment of freeway crash casualty using high-resolution traffic data. *Reliability Engineering & System Safety, 169*, 299–311. https://doi.org/10.1016/j.ress.2017.09.005

Appendix

Figure 3: Scatter plots between traffic collisions and each potential predictors.



Note: Each scatter plot is for the untransformed training set with 67 observations. Smooth line in blue is fitted to each plot.

Table 4: Comparison of the final model on training set and test set

Response variable: Square root of number of traffic collisions						
	Training set model (Model 5)			Test set model		
	Adjusted R squared = 0.7509			Adjusted R squared = 0.7428		
	Estimated coefficient	Standard error	p-Value	Estimated coefficient	Standard error	p-Value
Intercept	-14.4471	2.8949	0.0000	-18.5441	3.3117	0.0000
Cubic root of number of people in labour force	1.0575	0.0910	0.0000	0.9548	0.1126	0.0000
Tenth root of business per area	9.4073	1.9643	0.0000	10.8126	2.1090	0.0000
Square root of population density	-0.1129	0.0187	0.0000	-0.0884	0.0153	0.0000
Road Volume	0.0004	0.0002	0.0471	0.0002	0.0001	0.0859
Cubic root of number of TTC overcrowded routes	-0.4588	0.2215	0.0426	0.0514	0.2105	0.8080

Note:

Road volume and cubic root of number of TTC overcrowded routes are significant in the training set model, but not in the test set model. The corresponding p-values are highlighted in red.