Toronto's gas prices, employment rates, temperature, wind speed, rainfall, snowfall, visibility and relative humidity's linear relationship with weekday TTC ridership

Isaac (Hong Yee) Hua

07/12/2022

Introduction

The goal of this study is to estimate the Toronto Transit Commission's (TTC) daily weekday ridership with an interpretable linear regression model. Toronto's gas prices, employment rates, wind speed, rainfall, temperature, visibility and relative humidity effects on daily weekday public transportation ridership will be analysed. Similar to existing research by Tao Sui et al., rainfall, temperature and wind speed had a statistically significant influence on public transportation ridership in Brisbane, Australia, using a linear regression model with ARIMAX and SARIAMX. However, the researchers noted there seemed to be variations in the results across different days of the week (Sui et al., 2018). Another study discovered fare prices, income, speed and frequency had statistically significant impacts on demand of Colombian public transportation (Toro-González et al., 2020).

The effects on weekdays specifically is a unique part of this research, in addition to the focus on a denser, cosmopolitan city spanning a longer 10 year time period, which also snows for roughly a third of the year.

Methods

Data was gathered from the City of Toronto's Toronto Dashboard, WeatherStats.ca and Statistics Canada. This data was randomly split 50/50 into training and testing datasets. Histograms, boxplots and numerical summaries were created to explore any noteworthy patterns to be cautious of, such as skews, normality or fanning.

A linear regression model was fit between all the predictors against the response variable. Verifying the underlying regression assumptions, the fitted values were plotted against the corresponding response values and every predictor variable was plotted against each other. The conditional mean response should be a single function of a linear combination of the predictor variables and the conditional mean of each predictor variable should be a linear function when compared with another predictor.

If the fitted value plot showed that points did not have a random scatter around the identity function or if non-linear associations appeared between predictors, diagnostics from verifying regression assumptions hereafter could have been inaccurate and solutions to fix them may be incorrect. Thus, past literature would support subsequent conclusions about any assumption violations.

Each regression assumption was checked individually. Assuming conditions were met, residuals were plotted against each predictor variable to identify any non-linear patterns (e.g. quadratic, logarithmic), noting down the observed predictors potentially having violated the regression assumption of a linear relationship. Furthermore, the residuals that seemed to vary in height across different values of the predictors were noted down as potentially violating the constant error variance regression assumption. The residuals were also plotted against the fitted values to identify the same phenomenon.

Lastly, a QQ plot of the residuals was constructed to verify the normality assumption of linear regression - if points did not follow roughly a straight line without severe deviations, this was noted as potentially affecting

the variability of the regression coefficients, possibly disrupting the statistical significance of the hypothesis tests. Correlated errors shall be elaborated further in the discussion section.

Transformations to variables using the Box-Cox method would be conducted commensurate to the violations in assumptions, favoring interpretable transformations. Partial F tests were conducted with models subtracting predictors with t-test p-values larger than 0.05. For any new models being fit from this point until the final model, all conditions and assumptions were checked.

Multicollinearity was investigated by calculating the variance inflation factors (VIF) of each predictor. Predictors with values larger than 5 were removed. The goodness of these models were then assessed based on adjusted R^2, Akaike's information criterion (AIC), corrected AIC and Bayesian information criterion (BIC). The best model to be picked would have one of the largest adjusted R^2 values, while having some of the lowest values for AIC, corrected AIC and BIC.

Problematic observations were identified through calculating the leverage, the standardized residuals, the Cooks' distance and the difference in fitted values of each observation, in addition to the difference in betas for each regression coefficient and observation; cutoffs used were: $2(\frac{p+1}{n})$, both $[-2, 2]$ and $[-4, 4]$, the 50th percentile of $F(p+1, n-p-1)$ distribution, $2\sqrt{\frac{p+1}{n}}$, and $\frac{2}{\sqrt{n}}$ respectively, where p was the number of predictors in the model and n was the number of observations in the dataset. Observations that exceeded cutoffs would be noted down as corresponding outliers, leverage and/or influential points which may present model limitations.

Lastly, the final model was validated by going through the same process but with the testing dataset. Assuming that the split datasets shared similar characteristics, the full linear regression model was applied to the test dataset to ensure that the same conditions and assumptions were met, with the same variable transformations being necessary. ANOVA F Test should show a linear relationship exists overall as well as the same partial F tests being necessary. For every model from the start to the end, the assumptions and conditions were checked.

Multicollinearity was checked to show that the same variables were correlated, with identical decisions to remove the same ones. Problematic points should reflect similar behaviour in the testing dataset as in the training dataset. Every single model that was fit should follow the same model violations, fulfilling the same conditions and assumptions as applied to the training dataset. Adjusted R^2 and estimated regression coefficients should be similar to the training model, otherwise there was overfitting.

If the process or patterns observed were different in any way, e.g. estimated coefficients differed by more than 2 standard errors or different R^2, the model cannot be validated and each difference would have implications discussed further in the results and discussion section.

Results

Table 1: Summary statistics in training and testing datasets, each of size 60. Most statistics are very similar, however response has moderately different standard deviation and mean temperature differs by 1 degree across the two datasets.

Variable	Mean (SD) in Training	Mean (s.d.) in Testing
Daily Weekday TTC Ridership	1.6515167×10^6 (7.990089×10^4)	1.6465×10^6 (9.618688×10^4)
Gas Prices in cents	117.883 (12.582)	116.737 (12.45)
Employment Rate %	59.39 (1.161)	59.62 (1.153)
Temperature in Celsius	8.432 (9.787)	9.535 (10.307)
Relative Humidity %	68.747 (5.515)	68.72 (5.351)
Wind Speed in km/h	17.3 (2.252)	17.097 (2.107)
Visibility in meters	2.0152545×10^4 (1743.701)	2.0178668×10^4 (1848.623)
Rainfall in mm	2.074 (1.276)	1.862 (1.228)
Snowfall in mm	2.793 (4.913)	2.804 (4.665)

The split datasets share very similar characteristics except for negligible disparities. The EDA showed that every variable appeared to follow a normal distribution except for rainfall and snowfall which skewed to lower values. Gas prices, temperature and visibility seemed bimodally distributed. All seemed to have constant error variance. Most predictors showed a weak linear relationship with the response variable, with snowfall and relative humidity showing almost no relationship.

Upon applying the full linear regression model fit against daily weekday TTC ridership, the residual and QQ plots showed that the conditions and assumptions were met, except where rainfall and snowfall violated linearity assumption. Through the Box-Cox method, the suggested λ values were 0.5 and 0.1 respectively.

Transforming the variables, the model fulfilled conditions and assumptions. An ANOVA F-Test with p-value cutoff 0.05 led to the null hypothesis of no linear relationship being rejected as the model had a p-value of 0.02494 and an R^2 of 0.2779. Partial F Tests were deemed unfit due to none being statistically significant at any level past 50%, since removing any predictors at all affected the overall fitting ability of the model.

Evaluating the multicollinearity of the model, the VIFs for the full model showed that temperature and snowfall as predictors indicated severe collinearity (VIF >5). Since temperature was the most statistically significant predictor, snowfall was removed but multicollinearity remained with temperature having a VIF of 8.17. Wind speed was removed, having the second highest VIF and the second lowest $|t|$ statistic in the reduced model. Thus, this final model had no severe multicollinearity.

Table 2: Summary of goodness measures for models fit against weekly TTC ridership. Visibility was the next high VIF predictor. All 4 measures indicate the best fitting model is the one that removed snowfall and wind speed.

Model	Adjusted R^2	AIC	Corrected AIC	BIC
Full Model	0.1647	1350.08	1354.39	1375.02
(-) Snowfall	0.1706	1348.82	1352.28	1371.67
(-) Snowfall and Visibility	0.1242	1351.23	1353.94	1371.98
(-) Snowfall and Wind Speed	0.1788	1347.36	1350.08	1368.12

All goodness of fit measures agree that this final model is indeed the best model, with an overall R^2 of 0.262.

1 possible leverage point was identified, with 3 outliers outside the cutoff for $[-2, 2]$ while there were no

outliers for $[-4, 4]$. No observations were influential overall, but 5 were influential on their own fitted values, including the leverage point and potential outliers. Between 3-7 points were influential on a single estimated coefficient, including the aforementioned observations.

The model was validated by applying the full original model to the testing dataset. The same characteristics were found, fulfilling conditions but the same assumptions were not being met for rainfall and snowfall.

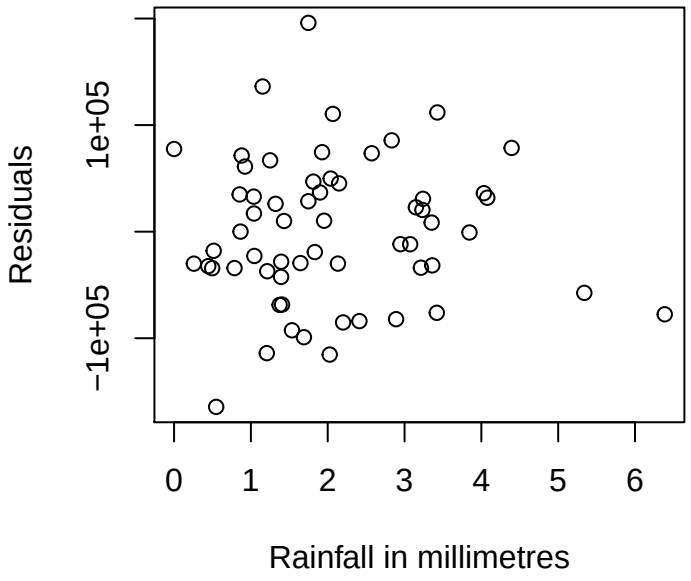

Figure 1: Example of Model Violations in Testing Dataset, Mirroring Training Dataset

Box-Cox evaluated λ at 0.5 and 0.08 for snowfall and rainfall respectively, very similar to the testing dataset. The model fulfilled assumptions after transformation of variables. However, under ANOVA F-Test with p-value 0.1619 meant that the null hypothesis actually failed to be rejected. Again, any partial F tests failed.

Snowfall and temperature had the highest VIF values at 7.42 and 12.2 respectively. Removing snowfall still showed severe multicollinearity - the best end result being the removal of only snowfall and wind speed. Problematic observations showed similar behaviour, e.g. 0 leverage points, 4 outliers, same DFBETAS cutoffs but 0 influential points overall.

Table 3: Summary of characteristics of the final model in the training and test datasets. Coefficients are presented as estimate $\pm$ standard error (* = significant t-test at $\alpha = 0.05$).

Characteristic	Model on Training Data	Model on Testing Data
Intercept	$-1.4157324 \times 10^5 \pm 5.9844793 \times 10^5$	$-5.2392351 \times 10^5 \pm 7.8617189 \times 10^5$
Gas Prices	764.04 ± 804.51	-335.57 ± 1126.19
Employment Rate	$1.841975 \times 10^4 \pm 8839.29$ (*)	$3.076637 \times 10^4 \pm 1.149799 \times 10^4$ (*)
Temperature	-6716.8 ± 1963.9 (*)	-5450.11 ± 2910.62
Relative Humidity	1910.28 ± 2555.97	-743.59 ± 2971.2
Visibility	23.12 ± 11.89	23.65 ± 16.15
Square Rooted Rainfall	$5.009509 \times 10^4 \pm 2.815327 \times 10^4$	$946.56 \pm 3.492586 \times 10^4$

Characteristic	Model on Training Data	Model on Testing Data
R^2 Value	0.262	0.192
Adjusted R^2	0.179	0.101
p-value	0.01041	0.06851

At each modification of the model in the model validation, the conditions and assumptions held. However, under a 0.05 significance level, the ANOVA F Test showed every single model failed to reject a non-existent linear relationship. Note that regression coefficients do not differ by less than 2 standard errors.

Discussion

The lack of normality in the EDA suggests the regression coefficients should vary widely. The linear correlation in predictor plots, high VIFs and possibly correlated errors indeed show that reduced information led to large variance, evident through the high standard errors of each coefficient. Thus, considering the low R^2 and the correlation, the usefulness of this final model is limited especially due to the low observation count contributing to high variability. Approximately 5% of 60 points were problematic, which suggest more data is needed. These limitations cannot be corrected because they are in the nature of the predictors and data is not readily available. Population and frequency of vehicles data could improve the model.

Answering the research question, the final model shows that the predictors as a whole have a very weak statistically significant linear relationship with the response, since R^2 is low and p-value differs in both datasets. Collinearity shows that snowfall and visibility may not directly have an impact. Employment rate is individually statistically significant in both datasets, while other predictors are not. Holding other factors constant, an increase in employment rate by 1% should lead to a 25,000 (based on mean from both datasets) increase in weekday TTC ridership.

References

City of Toronto. (2022). Toronto's Dashboard [Data dashboard]. https://www.toronto.ca/city-government/data-research-maps/toronto-progress-portal/

Statistics Canada. (2022). Table 18-10-0001-01 Monthly average retail prices for gasoline and fuel oil, by geography [Data table]. https://doi.org/10.25318/1810000101-eng

Canada Weather Stats. (2022). Climate Daily/Forecast/Sun based on Environment and Climate Change Canada data [Data set] https://toronto.weatherstats.ca/download.html

Tao, S., Corcoran, J., Rowe, F., & Hickman, M. (2018). To travel or not to travel: 'weather' is the question. modelling the effect of local weather conditions on bus ridership. Transportation Research Part C: Emerging Technologies, 86, 147–167. https://doi.org/10.1016/j.trc.2017.11.005

Toro-González, D., Cantillo, V., & Cantillo-García, V. (2020). Factors influencing demand for public transport in Colombia. Research in Transportation Business & Management, 36, 100514. https://doi.org/10.1016/j.rtbm.2020.100514