Ayalysis of the Ontario Public Library

Tong Su

12/10/2021

Introduction

Public libraries are important community hubs. They serve as centers of learning and professional development. More visitors visiting the public library will help develop a healthy and knowledgeable community to enhance social stability. A low number of visitors visiting the public library is a waste of social resources and adversely affects social connection.

The purpose of the study is to investigate the relationship between the number of visitors and other factors of the Ontario public library. By understanding the factors that may affect the number of visitors to the library, the government could then implement actions targeting their relevant factors to increase the number of visitors in the public library.

Method

This research question could be studied using the linear model. We used the number of visitors as the response variable and the other variables as predictors. We randomly split the dataset into two portions where 50% are the training dataset and 50% are the test dataset. We set the seed to 1 to ensure that the training and test data always stay the same. Next, we built the primary model based on the Explanatory Data Analysis (EDA), existing research, and some common judgments.

Before we used Residual Plots to check whether the first three assumptions hold, we first checked if the two conditions were satisfied. To determine whether condition 1 holds, we used a plot of the response against the fitted values to check whether there is a linear relationship between the response and the fitted values. To determine whether condition 2 holds, we made a scatter plot of all the predictors to check whether all pairwise relationships are linear or none. If either condition was unsatisfied, we could try transformation of variables using Box-Cox transformations and recheck the conditions.

After two conditions were satisfied, we used residual plots to determine whether the four assumptions were satisfied. We used residuals versus predictor plots to assess whether our first three assumptions hold. If our first three assumptions hold, we could observe that residuals are uniformly scattered around 0 for the full range of each predictor's values. We used the QQ plot to check normality. We expected to observe a straight diagonal string of points in the plot with minimal deviations at the ends if the normality assumption is satisfied. If either condition was unsatisfied, we could try transformation of variables using Box-Cox transformations and recheck the conditions.

After we corrected every violated assumption, we then tried to remove the unnecessary predictors. We used stepwise selection with both AIC and BIC to find the best model. We also used manual selection by looking at the summary of the model to observe the significance of the coefficients. We wanted to choose one of them to be our final model.

We used the test dataset to fit into these preferred models. We compared the difference in the estimated regression coefficients, the significance of the same predictors, model violations, adjusted R^2, problematic observations. The model is validated if the model looks very similar to how it performed in the training

dataset. We also compared their adjusted R^2 values, AIC and BIC values. We preferred the model with relatively large adjusted R^2 and p-values values and smaller AIC and BIC values.

After we chose the final model, we did a partial F test to check whether the reduced model was better than the full model. We also rechecked the conditions and the assumption of the reduced model by using the residual plots. We then checked for problematic observations. The point is considered influential if h_ii > 2(p+1)/n, an outlier if r_i is not between -4 and 4, and influential if the Cook's distance is greater than the 50th percentile of F(p+1, n-p-1). Lastly, we checked multicollinearity by using the Variance Inflation Factor (VIF) to each predictor.

Results

There are 269 data in our dataset. I have randomly split the dataset into half where 135 of them are the training dataset and 134 are the test dataset.

Exploratory data analysis

Table 1: Numerical summaries in training and test dataset, each of size 134.

Variables	Mean in training	Mean in test	Standard deviation in training	Standard deviation in test
Number of visitors	1064.1	1201.7	1911.006	1838.827
Number of resources	81478	87129	175534.8	172337.9
Number of cardholders	4866	5176	8755.451	8444.281
Total funds	474003	563179	909788.8	968200.7
Total operating revenues	767698	880984	1427611	1545803
Total operating expenditures	737140	855198	1381359	1536276
Total space provided	10263	11188	17043.68	17239.24
Total circulation of materials	1454	2706.31	3159.106	9161.654

From the numerical summaries of the training and test data, we could observe that they have similar mean and standard deviation. This means that there is no reason for our final model could not being validated by testing data.

Fig 1. Histogram of the total number of visitors

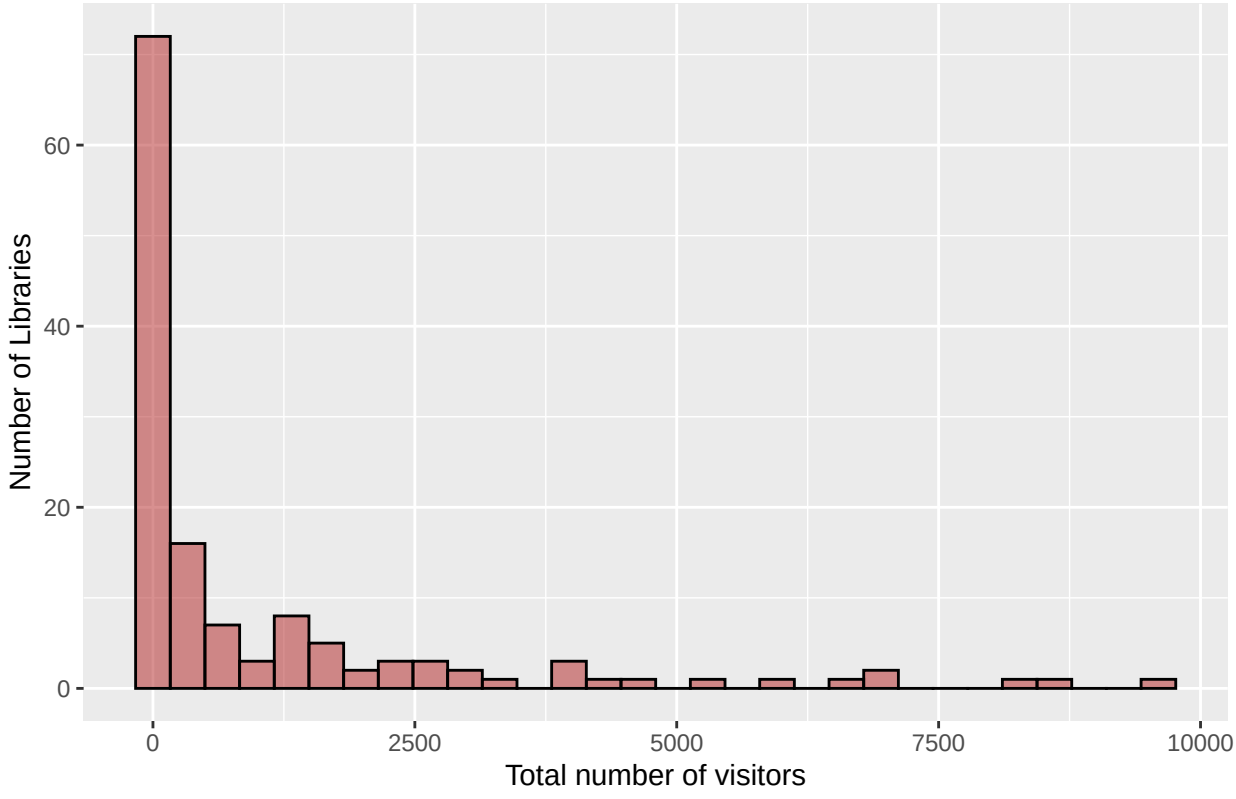
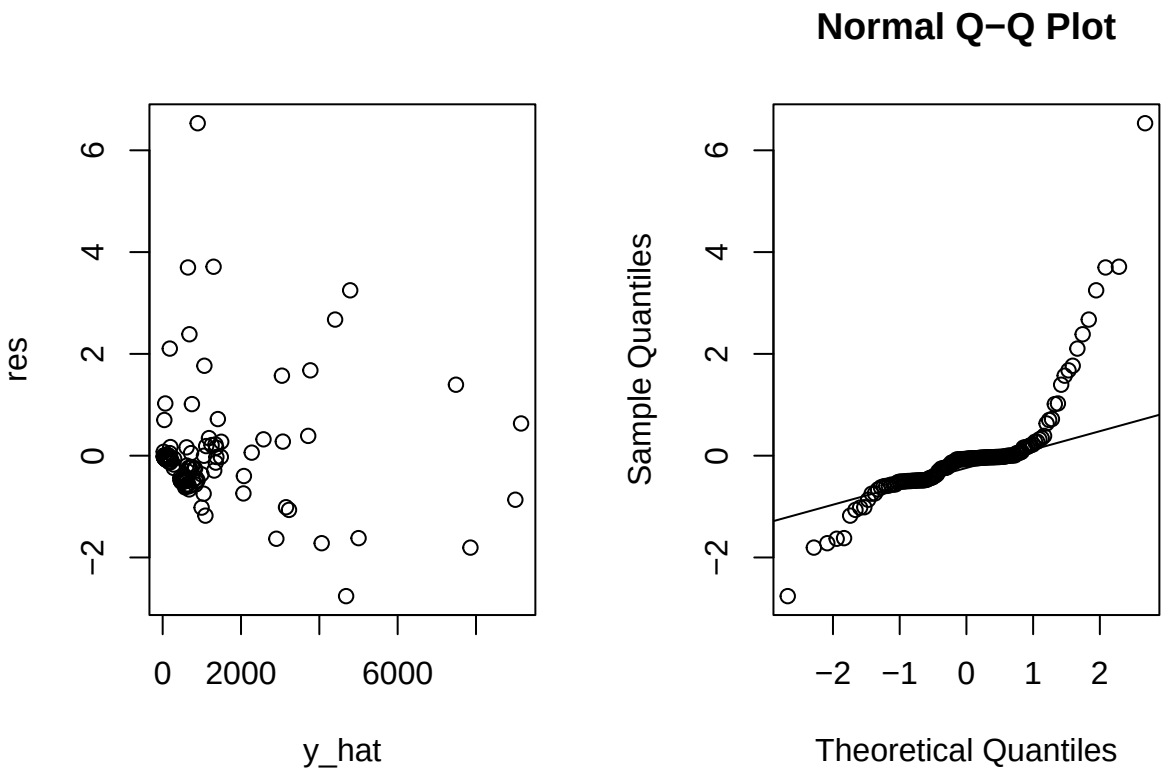


Fig 1. is the histogram of the library's total number of visitors. We could observe that it is right-skewed. It indicates that there are outliers in our training data response variable. Since the graph is not normally distributed, the normality of the errors assumption may be violated.

Check assumptions

From the response against the fitted values plot, we observed a linear relationship with some data clustering around 0. From the scatterplots of all pairwise predictors, all pairwise predictors appeared to be linear. This meant that our two conditions were both satisfied.

Figure 2. Residual Plots of the Full Model



Next, we used residual plots to check for assumptions. From Fig 2, we could observe that residuals were scattered around 0 for the full range of each predictor's values but there are some data that were clustering around zero. There was a severe deviation observed at the end of our QQ plot. It showed that our uncorrelated and normality of the errors assumption were violated.

Correct Assumptions

We used Box-Cox Transformation on our predictors to correct the normality assumption. We transformed all variables by a power of 0.2. We then rechecked all the conditions and assumptions. From Fig 4 shown in the appendix, no ore data clustering. The normality of the errors was also corrected with little deviations at the end of the QQplot.

Final Model Selection

We used stepwise selection with both AIC and BIC to find the best model. In both cases, the model was reduced to only two variables: total operating revenues and total circulation of materials. We then manually reduced the model. All the coefficients were not significant except for the variable total circulation of materials. We used the test dataset to fit into these two models and compared their properties with the training dataset.

Table 2: Comparison of two fitted models with test dataset

	Adjusted R square	AIC	BIC	Leverage points	Outliers	Influential points
Model1 using AIC/BIC stepwise selection	0.8063	242.1778	253.7989	10 30 36 55 63 66 69 75 108 110	None	None
Model1 with test data	0.7769	265.071	276.6624	2 16 40 45 65 70 128 130	None	None
Model2 using manual selection	0.7995	245.8672	254.583	5 32 36 55 64 66 98 108	None	None
Model2 with test data	0.706	301.0642	309.7577	16 40 130	130	40 130

	Coefficient of total circulation of materials	Significance of variable: total circulation of materials	Coefficient of total operating revenues	Significance of variable: total operating revenues
Model1 using AIC/BIC stepwise selection	0.65747	3.09e-12	0.06453	0.0186
Model1 with test data	0.276948	0.000405	0.179804	1.18e-09
Model2 using manual selection	0.84269	$<$ 2e-16	-	-
Model2 with test data	0.72089	$<$ 2e-16	-	-

According to table 2, Model 1 had a bigger adjusted R square and smaller AIC and BIC values for both training and test dataset. In terms of the coefficients, the significance was similar for both models' training and test datasets. However, the coefficients were more different in model 1 compared to model 2. In terms of the problematic observations, model 1 had similar values for both the training and test dataset. For Model 2, training data had more leverage points while test data had more outliers and influential points.

After careful considerations, we chose Model 1 as our final model. Which is

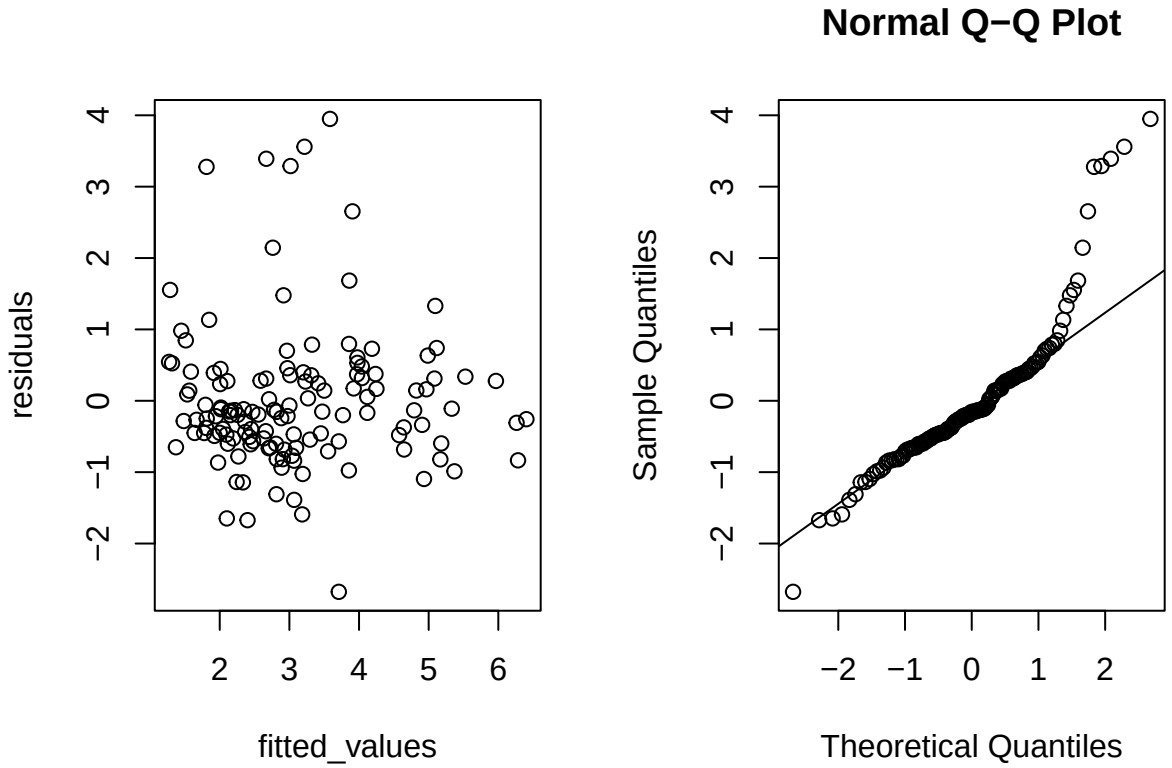
$$\hat{y} = 0.18090 + 0.06453x_1 + 0.65747x_2$$

Final Model Validation

We used a partial F-test to consider how the conditional relationship between the predictors and the response changes. According to the partial F test, the p-value was 0.8733. Since this value was much greater than 0.05, we did not reject the null hypothesis and concluded that the reduced model is better than the full model.

We then rechecked the conditions and the assumption of the reduced model. From Fig 2, we observed that the residuals were uniformly scattered around 0 for the full range of fitted values. There was a straight diagonal string of points in the QQplot with some deviations at the ends. This meant that our first three assumptions were satisfied but the fourth assumption was violated.

Figure 3. Residual Plots of the Final Model



We then checked the multicollinearity by using the Variance Inflation Factor (VIF) to each predictor. We obtained the value for both variables is 5.710922. Since the VIF value was greater than 5, we concluded that there was severe multicollinearity present.

Discussion

Interpretation of the final model

We used multiple linear regression to estimate the relationship between the number of visitors in the Ontario public library and different factors that may affect the number of visitors. The equation looks as follow:

$$\hat{y} = 0.18090 + 0.06453x_1 + 0.65747x_2$$

The estimate of the $\hat{\beta}_0$ is $\hat{\beta}_0 = 0.18090$. This means that when the total operating revenues and total circulation of materials are equal to 0, the mean number of visitors is 0.18090. Since the value of $\hat{\beta}_0$ is close to 0, this result is reasonable as when a library has no revenue and no books and material, no one will visit that library.

The estimate of $\hat{\beta}_1$ is $\hat{\beta}_1 = 0.06453$. This means that each additional dollar increase of the total operating revenues will trigger an average increase of 0.06453 number of visitors visiting that library when the total circulation of materials is held fixed.

The estimate of $\hat{\beta}_2$ is $\hat{\beta}_2 = 0.65747$. This means that each additional increase of total circulation of materials will trigger an average increase of 0.65747 number of visitors visiting that library when total operating revenues are held fixed.

Based on this result, we could infer that one dollar increase in total operating revenue will trigger less increase than the increase in total circulation of materials. In this case, instead of providing more funds to the public library, the government may choose to increase the total circulation of materials to the public library to increase the number of visitors.

Limitations

In our final model, the normality of errors assumption is violated even after transformations. Therefore we could not trust the results of hypothesis tests and confidence intervals.

The value of multicollinearity is large. We have obtained that the VIF value for both variables is 5.710922. It is not feasible for us to collect more data or determine which one we should remove since there are only two predictors. If the wrong predictor is removed, it may reduce the ability of the model to accurately predict.

The values of AIC and BIC are large compared to normal values. This indicates that the variables may be inappropriate for predicting the number of visitors visiting the library.

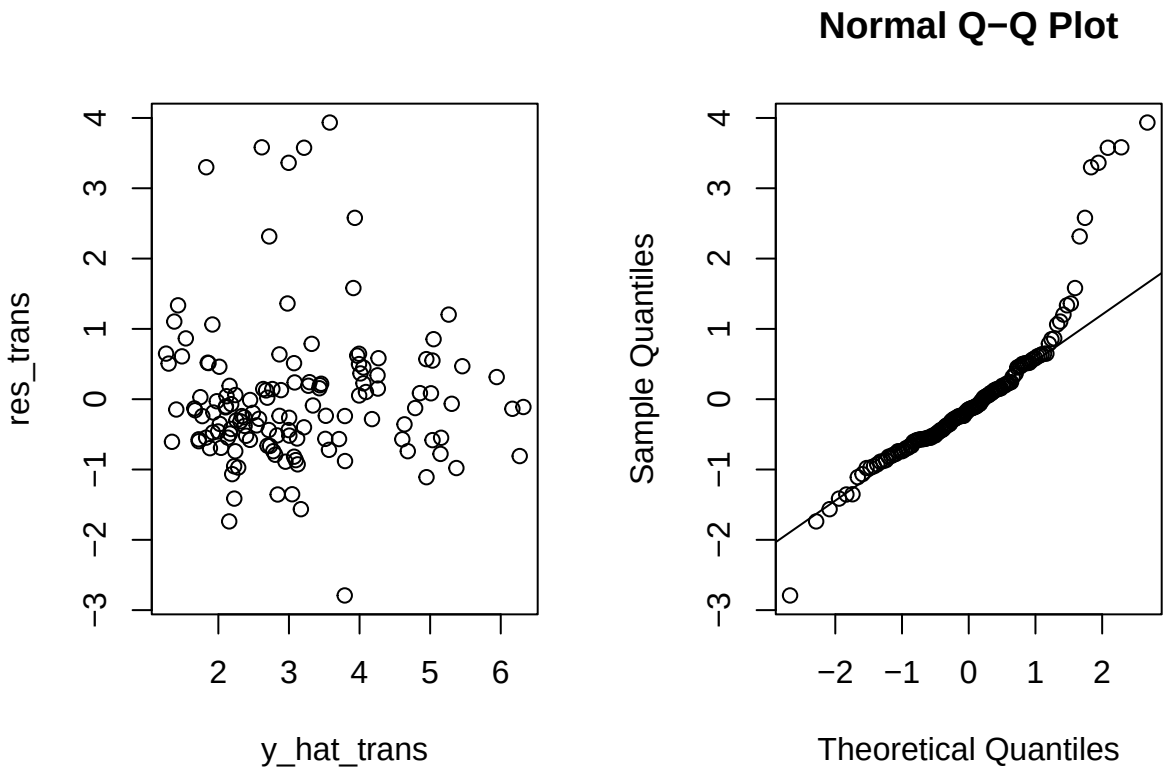
The model chosen was based on automated selection methods. This method is running based on the assumption that four assumptions are held which in reality is not true. The outputs are based solely on the goodness of fit measures. In this case, some ethical and statistical issues were risen when we chose this model.

Although there were no outliers and influential points present, there are 10 leverage points in our model. These leverage points would have a disproportionate impact on the placement of our regression line.

#Appendix

Figure 4. Residual Plots of the Full Model after Transformation

```
model_trans_full <- lm(number_of_visitors ~ . -Name, data=train_trans)
res_trans<-rstandard(model_trans_full)
y_hat_trans <- fitted(model_trans_full)
par(mfrow=c(1,2))
plot(y_hat_trans, res_trans)
qqnorm(res_trans)
qqline(res_trans)
```



Reference

1. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

2. Sam Firke (2021). janitor: Simple Tools for Examining and Cleaning Dirty Data. R package version 2.1.0. https://CRAN.R-project.org/package=janitor

3. Ontario Data Catalogue. (2021). Ontario public library statistics, 2020 Retrieved from https://data.ontario.ca/en/dataset/ontario-public-library-statistics/resource/c897cb92-d90f-4965-b2d1-c5b08569bb58