Store Data Analysis

Consultant: Brendan Callender

Background

A large do-it-yourself (DIY) store in a densely populated suburb has collected data on the number of customers who visited their store during a two-week perdios from 110 different nearby neighborhoods. For each neighborhood in the data, the DIY store has provided the variables listed in *Table 1* below.

Table 1. Dataset Variables

Variable	Variable Description
Customers	Number of customers from the neighborhood who visited the store
Units	Number of housing units in the neighborhood
Income	Median household income of the neighborhood in thousands of dollars
Age	Median age of housing in the neighborhood in years
Distance to Competitor	Miles from the neighborhood to the nearest competitor DIY store
Distance to Store	Miles from the neighborhood to the store

The goal of the analysis is to construct a model which best models how many customers will visit the DIY store based on the predictors collected in the data. This model will then be used to provide predicted counts for the number of customers that will visit a store based on different values for the predictors in the model.

Challenges in Model Fitting

The first challenge in the model fitting process was addressing the overdispersion in the initial poisson loglinear model. To address the overdispersion, I fit a negative binomial (NB2) model and two different quasi-likelihood models with different variances. *Table 2* below shows the results for pearson goodness of fit tests conducted for each model. Looking at these results, there is very strong evidence of a poor fit for each model specification.

Table 2. Main Effect Model Goodness of Fit Test Comparisons

Model	Poisson	NB2	QL (Var = μ)	QL(Var = μ²)
X^2	652.15	264.52	652.15	265.91
P-value	< 0.0001	< 0.0001	< 0.0001	< 0.0001

Looking further into the NB2 model, which had the lowest X^2 test statistic, I uncovered a major outlier in the data that was inflating the X^2 values. The observation appears to have a potential input error of 19 for the number of housing units in the neighborhood. For all other neighborhoods, the number of units ranges from 109 and 1289 housing units. *Table 3* below shows the outlier observation. According to the collected data shown below, more than half of

the residents visited the DIY store from this neighborhood. Because of the unknown nature of this observation, it was removed from the data for the remaining parts of the analysis.

Table 3. Outlier Found in the Dataset

Customers	Units	Income	Age	Competitor Distance	Store Distance
10	19	64.2	22	2.96	6.09

After removing the outlier, I refit the four different models to the data and conducted four goodness-of-fit tests again to compare the model fits. Looking at the results summarized below in *Table 4*, the NB2 model and QL(Var = μ^2) had the smallest X^2 meaning they fit the data the best. After looking further into why the QL model had such a low test statistic, I found this model had evidence of underdispersion and decided to continue the analysis with the NB2 model.

Table 4. Main Effect Goodness of Fit Test Comparisons After Outlier Removal

Model	Poisson	NB2	QL (Var = μ)	QL(Var = μ^2)
X^2	443.96	110.23	443.96	38.63
P-value	< 0.0001	0.319	< 0.0001	1.000

Final Model

Model Description

The final model used a negative binomial random component with a log link function modeling the rate at which customers will visit the DIY store from a given neighborhood during a two week period.. The predictors in the model included the medium income of the neighborhood, the distance from the neighborhood to the nearest competitor store, the distance from the neighborhood to the store, and the number of housing units in the neighborhood which was used as an offset. All predictors except for the offset were centered to make the intercept of the model easier to interpret. The output for the model can be found below in *Table 5*.

Table 5. Final Model Output

Parameter	Estimate	SE	z value	P-value
Intercept	-4.06	0.0572	-70.9	< 2e-16
Median Income (C)	-0.390	0.0648	-6.02	1.79e-09
Distance to Competitor (C)	0.255	0.0482	3.49	4.76e-04
Distance to Store (C)	-0.295	0.0731	-4.03	5.57e-05
Dispersion Parameter	4.043	0.751		

Looking at the model output in *Table 5* above, we have significant evidence that median income, distance to competitor, and distance to store are all individually significant predictors, at the 5% overall significance level, of the rate at which individuals from a given neighborhood will visit the DIY store during a two-week period after adjusting for the other predictors in the model.

For a nearby neighborhood with an average median income, average distance to the nearest competitor, and average distance to the DIY store, we expect 1.73% of the residents in the neighborhood to visit the DIY store during two-week periods similar to the period when the data were collected.

Please note that the following interpretations are only applicable for two-week periods that are similar to the period when the data were collected.

For every $1,000 increase in the median income of a nearby neighborhood, there is an associated decrease in the rate at which people from the neighborhood will visit the DIY store of 32.3% after adjusting for the neighborhood's distance to the nearest competitor and distance to the store.

For every 1 mile increase in the distance from a nearby neighborhood to the nearest competitor DIY store, there is an associated increase in the rate at which people from the neighborhood who will visit the DIY store of 29.0% after adjusting for the neighborhood's median income and distance to the store.

For every 1 mile increase in the distance from a nearby neighborhood to the DIY store, there is an associated decrease in the rate at which people from the neighborhood will visit the store of 25.5% after adjusting for the neighborhood's median income and distance to the nearest competitor.

Discussion of Model Diagnostics

The scope of this analysis and model is limited to neighborhoods with more than 100 housing units that are near the DIY store since Observation 11 was removed from the analysis. The final model residual plots show no violations with unequal variance and linearity. This indicates the loglinear, negative binomial model is an appropriate model form for these data. This is shown below in *Figure 1.* Additionally, the final model does not demonstrate strong evidence (p-value = 0.029) of fitting the data poorly when performing a pearson goodness of fit test. When looking at the diagnostic plots shown in *Figure 2* below, there is one high leverage observation in the data. This corresponds to the neighborhood with the largest median income, which is not considered influential and was thus kept in the data. Observations 31 and 81 stand out with large Cook's distances however these fall well below the cutoff of 0.88 which was used as a threshold for influential observations. This value was calculated using the 50th percentile of an F distribution with 5 numerator degrees of freedom and 104 denominator degrees of freedom.

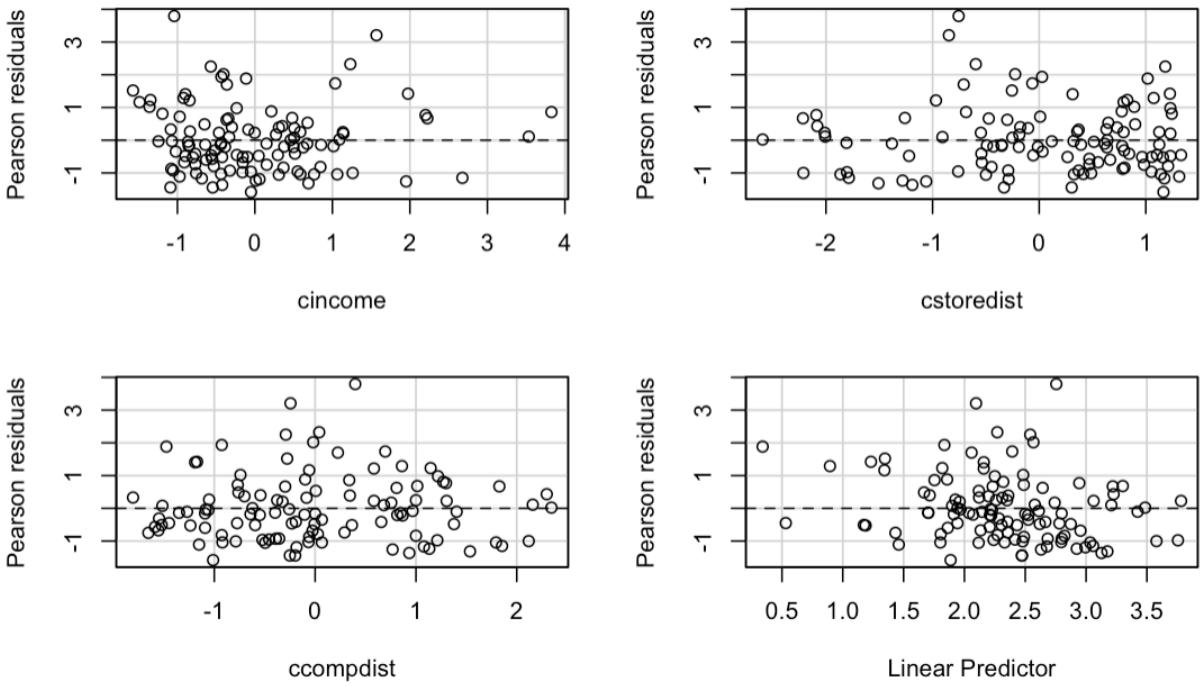

Figure 1. Residual Plots for Final Model

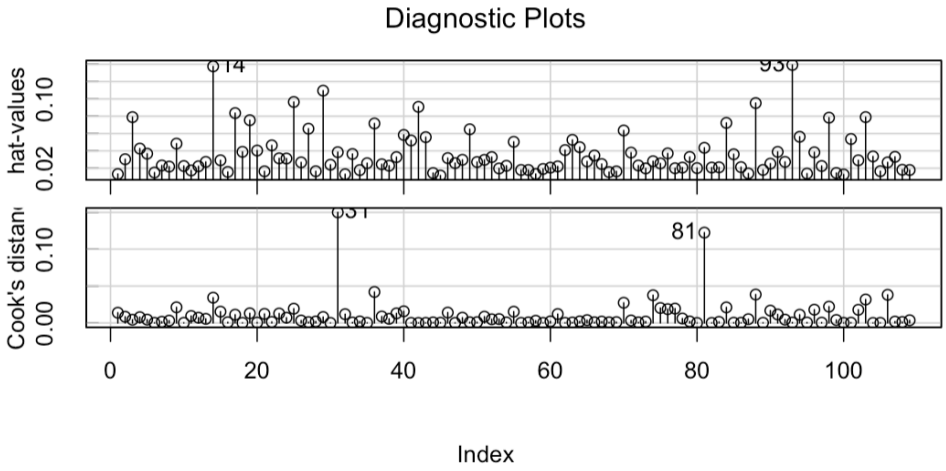

Figure 2. Plots Showing Leverages (Top) and Cook's Distances (Bottom) for Final Model

Model Predictions

Table 6. Final Model Predictions

Median Income (C)	***Distance to Competitor (C)***	***Distance to Store (C)***	***Predicted Mean Count***
-1.57	*2.34*	*-2.59*	*80.999*
-1.57	*0.00*	*-2.59*	*44.648*
0.00	*2.34*	*-2.59*	*43.941*
-1.57	*2.34*	*0.00*	*37.773*
-1.57	*-1.81*	*-2.59*	*28.165*
-1.57	*2.34*	*1.33*	*25.531*
0.00	*0.00*	*-2.59*	*24.221*
-1.57	*0.00*	*0.00*	*20.821*
0.00	*2.34*	*0.00*	*20.491*
0.00	*-1.81*	*-2.59*	*15.279*
-1.57	*0.00*	*1.33*	*14.073*
0.00	*2.34*	*1.33*	*13.850*
-1.57	*-1.81*	*0.00*	*13.135*
0.00	*0.00*	*0.00*	*11.295*
3.83	*2.34*	*-2.59*	*9.883*
-1.57	*-1.81*	*1.33*	*8.878*
0.00	*0.00*	*1.33*	*7.634*
0.00	*-1.81*	*0.00*	*7.125*
3.83	*0.00*	*-2.59*	*5.448*
0.00	*-1.81*	*1.33*	*4.816*
3.83	*2.34*	*0.00*	*4.609*
3.83	*-1.81*	*-2.59*	*3.437*
3.83	*2.34*	*1.33*	*3.115*
3.83	*0.00*	*0.00*	*2.541*
3.83	*0.00*	*1.33*	*1.717*
3.83	*-1.81*	*0.00*	*1.603*
3.83	*-1.81*	*1.33*	*1.083*

Note: The predicted counts are for a neighborhood with 654 housing units which represents an average sized neighborhood