Crafting Customer Counts

DIY Modeling of Neighborhood Store Visits

I. Dataset

The data is a record of the number of customers who visited a large do-it-yourself store in a densely populated suburb collected over a two-week period from 110 different neighborhoods. Each observation is a nearby neighborhood, and the response variable is the amount of customers from the neighborhood who visited the store. The possible predictors variables can be found in *Table 1*.

Predictor Variable	Description
units	The number of housing units in the neighborhood
income	The median household income of the neighborhood in thousands of dollars
age	The median age of housing in the neighborhood in years
compdist	The distance from the neighborhood to the nearest competitor DIY store in miles
storedist	The distance from the neighborhood to the store in miles

Table 1. Recorded variables for each neighborhood

II. Model Steps

Since the response variable is a count, this means the random component of our model will be poisson or negative binomial. Examining the customer counts graphically in *Figure 1* leads us to preferring a negative binomial model.

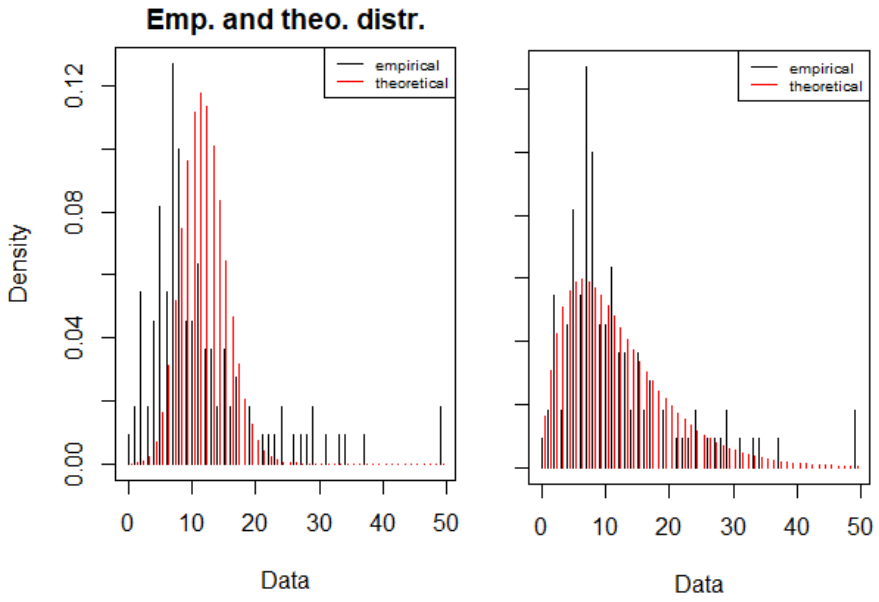

Figure 1. Observed counts in black overlaid by theoretical Poisson (left) and Negative Binomial (right) distributions in red

Performing a Likelihood Ratio Test (LRT) between the main effects models of a poisson with log link model and negative binomial with log link model resulted in a p-value ≈ 0, providing strong evidence that the negative binomial model fits the data better than the poisson model. Furthermore, testing for overdispersion in the poisson model resulted in a p-value < 0.001, signifying that overdispersion was detected, providing additional evidence for use of the negative binomial model.

Stepwise regression in both directions was performed on all predictors with two-way interactions resulting in model 1. Using a Bonferroni-adjusted p-value at a 5% level of significance (0.05 / 5 = 0.01), the interaction between compdist and storedist is not significant with a p-value of 0.0879 and is dropped (model 2).

Predictors	**Model 1 P-value**	**Model 2 P-value**
units	0.0008803	0.0014042
income	0.0008463	0.0003256
compdist	0.0005050	0.0005625
storedist	1.31e-05	1.576e-05
compdist:storedist	0.0879754	-
AIC	691.0015	691.8929

Table 2. Likelihood Ratio Test for each predictor in model 1 and model 2 and the AIC values of both models

A LRT between model 1 and model 2 results in a p-value = 0.0891, providing some, but not enough, evidence that model 1 is a better fit than model 2. Although the AIC value is slightly lower for model 1, we chose to continue with model 2 due to the outcome of the LRT and since model 2 is simpler than model 1.

A LRT between model 2 and the full model with all predictors and two-way interactions resulted in a p-value = 0.8883, not providing enough evidence to conclude that any additional predictors or interactions significantly improve model 2.

III. Final Model

The final model has a negative binomial random component with a log link function. The linear predictor (predictor variables and interactions) includes the following: units, income, compdist, and storedist. All four of these variables are quantitative, and have been centered within the model to provide a meaningful intercept interpretation.

	Intercept	units	income	compdist	storedist
Estimate	2.3361873	0.0008511	-0.0139509	0.1563735	-0.1293811
Exponentiated Estimate	10.3417317	1.0008514	0.9861459	1.1692628	0.8786391

Table 3. Final model coefficients (all predictors are centered)

Interpretations in the context of the data are as follows:

Intercept: For a neighborhood with an average amount of housing units, an average median household salary, located an average distance from the store, and located an average distance from the nearest competitor store, the estimated mean count of customers visiting the store from the neighborhood is 10.34.

units: After adjusting for the median household income and the distance from the neighborhood to the store and to the nearest competitor store, each increase of 1 housing unit in the neighborhood is associated with an increase of 0.0851% in the mean count of customers visiting the store from the neighborhood.

income: After adjusting for the amount of housing units in the neighborhood and the distance from the neighborhood to the store and to the nearest competitor store, each increase of $1,000 in the median household income of the neighborhood is associated with a decrease of 1.3854% in the mean count of customers visiting the store from the neighborhood.

compdist: After adjusting for the median household income, distance from the neighborhood to the store, and for the amount of housing units in the neighborhood, each increase of 1 mile in the distance from the neighborhood to the nearest competitor store is associated with an increase of 16.9263% in the mean count of customers visiting the store from the neighborhood.

storedist: After adjusting for the median household income, distance from the neighborhood to the nearest competitor store, and for the amount of housing units in the neighborhood, each increase of 1 mile in the distance from the neighborhood to the store is associated with a decrease of 12.1361% in the mean count of customers visiting the store from the neighborhood.

The only possible issues that exist in the final model regard the leverage of observations 15, 30, and 94. However, since these observations are not influential they were kept.

IV. Appendix

A. Table of minimum, mean, and maximum values for all predictors in final model

	units	income	compdist	storedist
min	19.0000	44.67300	0.340	0.870000
mean	647.7636	73.83678	3.068	6.831727
max	1289.0000	145.06500	6.610	9.900000

B. Table of count predictions from final model

units	income	compdist	storedist	predicted count
min	min	min	min	12.8414944
mean	min	min	min	21.9287305
max	min	min	min	37.8461293
min	mean	min	min	8.5490425
mean	mean	min	min	14.5987408
max	mean	min	min	25.1955228

units	income	compdist	storedist	predicted count
min	max	min	min	3.1648895
mean	max	min	min	5.4045118
max	max	min	min	9.3274826
min	min	mean	min	19.6734208
mean	min	mean	min	33.5952442
max	min	mean	min	57.9810106
min	mean	mean	min	13.0973003
mean	mean	mean	min	22.3655565
max	mean	mean	min	38.6000340
min	max	mean	min	4.8486726
mean	max	mean	min	8.2798179
max	max	mean	min	14.2898859

units	income	compdist	storedist	predicted count
min	min	max	min	34.2314009
mean	min	max	min	58.4551252
max	min	max	min	100.8859235
min	mean	max	min	22.7890688
mean	mean	max	min	38.9156691
max	mean	max	min	67.1633702
min	max	max	min	8.4366038
mean	max	max	min	14.4067353
max	max	max	min	24.8641464
min	min	min	mean	5.9378487
mean	min	min	mean	10.1397454
max	min	min	mean	17.4998782

units	income	compdist	storedist	predicted count
min	mean	min	mean	3.9530384
mean	mean	min	mean	6.7503914
max	mean	min	mean	11.6502953
min	max	min	mean	1.4634305
mean	max	min	mean	2.4990217
max	max	min	mean	4.3129856
min	min	mean	mean	9.0969004
mean	min	mean	mean	15.5342883
max	min	mean	mean	26.8101559
min	mean	mean	mean	6.0561322
mean	mean	mean	mean	10.3417317
max	mean	mean	mean	17.8484804

units	income	compdist	storedist	predicted count
min	max	mean	mean	2.2420042
mean	max	mean	mean	3.8285502
max	max	mean	mean	6.6075783
min	min	max	mean	15.8284443
mean	min	max	mean	27.0293844
max	min	max	mean	46.6491927
min	mean	max	mean	10.5375619
mean	mean	max	mean	17.9944286
max	mean	max	mean	31.0560373
min	max	max	mean	3.9010473
mean	max	max	mean	6.6616089
max	max	max	mean	11.4970683

units	income	compdist	storedist	predicted count
min	min	min	max	3.9923116
mean	min	min	max	6.8174562
max	min	min	max	11.7660404
min	mean	min	max	2.6578247
mean	mean	min	max	4.5386246
max	mean	min	max	7.8330742
min	max	min	max	0.9839373
mean	max	min	max	1.6802168
max	max	min	max	2.8998352
min	min	mean	max	6.1162995
mean	min	mean	max	10.4444763
max	min	mean	max	18.0258041

units	income	compdist	storedist	predicted count
min	mean	mean	max	4.0718395
mean	mean	mean	max	6.9532617
max	mean	mean	max	12.0004229
min	max	mean	max	1.5074112
mean	max	mean	max	2.5741251
max	max	mean	max	4.4426042
min	min	max	max	10.6422520
mean	min	max	max	18.1732023
max	min	max	max	31.3645773
min	mean	max	max	7.0849281
mean	mean	max	max	12.0985512
max	mean	max	max	20.8805217

units	income	compdist	storedist	predicted count
min	max	max	max	2.6228685
mean	max	max	max	4.4789317
max	max	max	max	7.7300521