# Exploring How Neighborhood Characteristics Influence Customer Visits at a DIY Store

Rachel Roggenkemper: rroggenk@calpoly.edu
Dr. John Walker
STAT 550: Generalized Linear Models, Winter 2025

## I. Dataset Description

A large do-it-yourself (DIY) store located in a densely populated suburb collected data over a two-week period to understand the factors influencing customer visits. The store gathered information from 110 nearby neighborhoods to assess how neighborhood characteristics relate to the number of customers visiting the store. Each observation, or row in the dataset, represents a neighborhood near the store. The response variable is the number of customers from the neighborhood who visited the store. The possible predictor variables are below in *Table 1*.

*Table 1: Predictor variables which were recorded for each neighborhood*

| **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 |

The overall project goal is to:

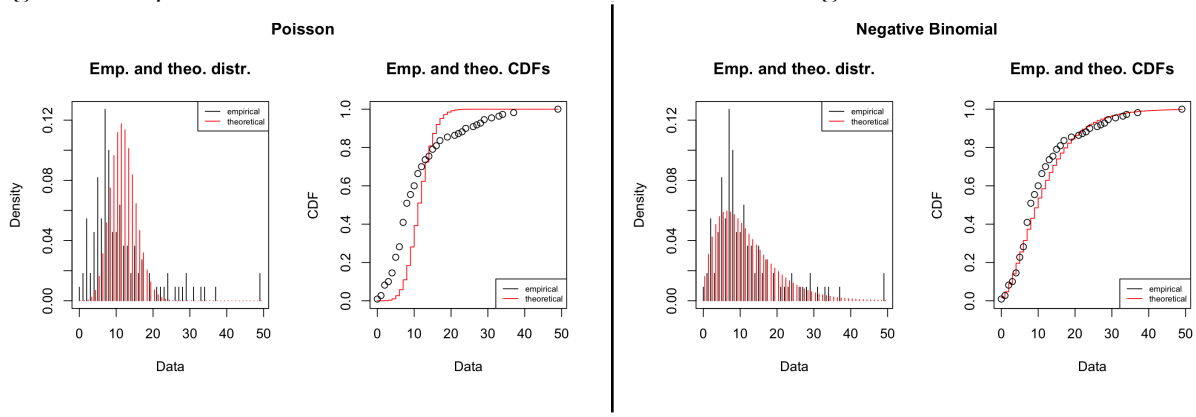
1. Develop a model that "best" explains the relationship between the response variable (the number of customers) and the predictor variables. This involves selecting an appropriate generalized linear model, assessing its fit, and interpreting how each predictor affects the customer count.
2. Use the final model to make predictions (in this case, counts) for various combinations of the predictor variables. These predictions can help understand how changes in neighborhood characteristics might influence store traffic.

## II. Methodology

Due to the response variable being a count, the random component of the model will either be Poisson or Negative Binomial. I began my analysis by fitting a Poisson GLM with a log link. However, diagnostic tests revealed severe overdispersion and poor model fit. To address the

overdispersion, I refitted the model using a Negative Binomial random component with a log link, which fixed the initial issue of overdispersion. As is shown in *Figure 1* below, which compares the observed (empirical) distribution of store customer counts with the fitted Poisson (left) and Negative Binomial (right) distributions, it illustrates that the Negative Binomial model more accurately captures the heavier right tail and overall variability in the data.

*Figure 1: Empirical vs. Theoretical Distributions: Poisson vs. Negative Binomial Fits*




After I decided on the Negative Binomial random component with a log link, I then had to address which predictors to include, and whether to include any interactions, if I should, which interaction to include. Note that all predictors (which are all quantitative) were centered to reduce multicollinearity—especially important when testing for interactions. A stepwise regression (both directions) based on BIC starting with the model with all main effects and all two-way interactions led me to a candidate model that included only the main effects of units, income, compdist, and storedist. After performing final checks of this proposed candidate model (steps outlined in depth in my *Project Log*), I confirmed that the proposed candidate model is indeed my final model.

## III. Final Model

The final model uses a negative binomial distribution as the random component with a log link function.

The linear predictor of the final model includes the following (note that all the predictors, since quantitative, have been centered to provide a meaningful interpretation of the intercept):

$$\log(\mu) = \beta_o + \beta_1(units) + \beta_2(income) + \beta_3(compdist) + \beta_4(storedist)$$

Below are the interpretations in the context of the data:

- Intercept ($\hat{\beta}_0 = 2.3361873,\ e^{\hat{\beta}_0} = 10.34173$)
    - For a neighborhood with an average number of housing units, an average median household salary, located an average distance from the store, and located an average distance from the nearest competitor DIY store, the estimated mean number of customers from the neighborhood who visited the store is 10.34173.
- units ($\hat{\beta}_1 = 0.0008511,\ e^{\hat{\beta}_1} = 1.000851$)

    - After adjusting for the median household income of the neighborhood, the distance from the neighborhood to the nearest competitor DIY store, and the distance from the neighborhood to the store, each increase of one housing unit in the neighborhood is associated with an increase of 0.0851% in the mean number of customers from the neighborhood who visited the store.

- income ($\hat{\beta}_2 = -0.0139509,\ e^{\hat{\beta}_2} = 0.986146$)
    - After adjusting for the number of housing units in the neighborhood, the distance from the neighborhood to the nearest competitor DIY store, and the distance from the neighborhood to the store, each increase of $1,000 in the median household income of the neighborhood is associated with a decrease of 1.3854% in the mean number of customers from the neighborhood who visited the store.
- compdist ($\hat{\beta}_3 = 0.1563735,\ e^{\hat{\beta}_3} = 1.169263$)
    - After adjusting for the number of housing units in the neighborhood, the median household income of the neighborhood, and the distance from the neighborhood to the store, each increase of one mile in the distance from the neighborhood to the nearest competitor DIY store is associated with an increase of 16.9263% in the mean number of customers from the neighborhood who visited the store.
- storedist ($\hat{\beta}_4 = -0.1293811,\ e^{\hat{\beta}_4} = 0.8786391$)
    - After adjusting for the number of housing units in the neighborhood, the median household income of the neighborhood, and the distance from the neighborhood to the nearest competitor DIY store, each increase of one mile in the distance from the neighborhood to the store is associated with a decrease of 12.13609% in the mean number of customers from the neighborhood who visited the store.

The only potential issue in the final model that still exists is there are three high leverage points present in the data (Observations: 15, 30, and 94). Since these observations do not have a high influence, I elected not to remove them. Additionally, since each row represents a unique neighborhood, dropping them would mean losing valuable data about store visits from those areas. To be thorough, I did compare the final model fitted on the entire dataset to the final model fitted on the dataset with the three high leverage points removed. The coefficients themselves do not change much, and the p-values and significance of the predictors do not change. Thus the interpretations above are still accurate regardless. However, the AIC does decrease slightly when the three high leverage points are removed. Ultimately, since the overall results do not change significantly, there isn't much to do except acknowledge that these high leverage observations exist.

*Disclaimer: Since this project is an exploratory observational study rather than a controlled experiment, we cannot infer cause-and-effect relationships between the predictors and the response variable. Instead, the final model best explains the observed association between the number of store customers and neighborhood characteristics.*
*Note: The predicted counts for key combinations of the predictor variables (minimum, mean. maximum) appear below in the Appendix.*

# Appendix

Rachel Roggenkemper

## Table of Minimum, Mean, and Maximum Values for Predictors

**Note: I am using the uncentered values to get more meaningful insights as it does not change the predictions**

| | units | income | compdist | storedist |
|---|---|---|---|---|
| Minimum | 19.0000 | 44.67300 | 0.340 | 0.870000 |
| Mean | 647.7636 | 73.83678 | 3.068 | 6.831727 |
| Maximum | 1289.0000 | 145.06500 | 6.610 | 9.900000 |

## Table of Predictions (counts) from my Final Model for All Combinations of the Predictor Variables

**Note: I am using the uncentered values to get more meaningful insights as it does not change the predictions**

| | units_label | income_label | compdist_label | storedist_label | predicted_customers |
|---|---|---|---|---|---|
| 1 | Minimum | Minimum | Minimum | Minimum | 12.8414944 |
| 2 | Mean | Minimum | Minimum | Minimum | 21.9287305 |
| 3 | Maximum | Minimum | Minimum | Minimum | 37.8461293 |
| 4 | Minimum | Mean | Minimum | Minimum | 8.5490425 |
| 5 | Mean | Mean | Minimum | Minimum | 14.5987408 |
| 6 | Maximum | Mean | Minimum | Minimum | 25.1955228 |
| 7 | Minimum | Maximum | Minimum | Minimum | 3.1648895 |
| 8 | Mean | Maximum | Minimum | Minimum | 5.4045118 |
| 9 | Maximum | Maximum | Minimum | Minimum | 9.3274826 |
| 10 | Minimum | Minimum | Mean | Minimum | 19.6734208 |
| 11 | Mean | Minimum | Mean | Minimum | 33.5952442 |
| 12 | Maximum | Minimum | Mean | Minimum | 57.9810106 |

| | | | | | |
|---|---|---|---|---|---|
| 13 | Minimum | Mean | Mean | Minimum | 13.0973003 |
| 14 | Mean | Mean | Mean | Minimum | 22.3655565 |
| 15 | Maximum | Mean | Mean | Minimum | 38.6000340 |
| 16 | Minimum | Maximum | Mean | Minimum | 4.8486726 |
| 17 | Mean | Maximum | Mean | Minimum | 8.2798179 |
| 18 | Maximum | Maximum | Mean | Minimum | 14.2898859 |
| 19 | Minimum | Minimum | Maximum | Minimum | 34.2314009 |
| 20 | Mean | Minimum | Maximum | Minimum | 58.4551252 |
| 21 | Maximum | Minimum | Maximum | Minimum | 100.8859235 |
| 22 | Minimum | Mean | Maximum | Minimum | 22.7890688 |
| 23 | Mean | Mean | Maximum | Minimum | 38.9156691 |
| 24 | Maximum | Mean | Maximum | Minimum | 67.1633702 |
| 25 | Minimum | Maximum | Maximum | Minimum | 8.4366038 |
| 26 | Mean | Maximum | Maximum | Minimum | 14.4067353 |
| 27 | Maximum | Maximum | Maximum | Minimum | 24.8641464 |
| 28 | Minimum | Minimum | Minimum | Mean | 5.9378487 |
| 29 | Mean | Minimum | Minimum | Mean | 10.1397454 |
| 30 | Maximum | Minimum | Minimum | Mean | 17.4998782 |
| 31 | Minimum | Mean | Minimum | Mean | 3.9530384 |
| 32 | Mean | Mean | Minimum | Mean | 6.7503914 |
| 33 | Maximum | Mean | Minimum | Mean | 11.6502953 |
| 34 | Minimum | Maximum | Minimum | Mean | 1.4634305 |
| 35 | Mean | Maximum | Minimum | Mean | 2.4990217 |
| 36 | Maximum | Maximum | Minimum | Mean | 4.3129856 |
| 37 | Minimum | Minimum | Mean | Mean | 9.0969004 |
| 38 | Mean | Minimum | Mean | Mean | 15.5342883 |
| 39 | Maximum | Minimum | Mean | Mean | 26.8101559 |
| 40 | Minimum | Mean | Mean | Mean | 6.0561322 |
| 41 | Mean | Mean | Mean | Mean | 10.3417317 |
| 42 | Maximum | Mean | Mean | Mean | 17.8484804 |
| 43 | Minimum | Maximum | Mean | Mean | 2.2420042 |
| 44 | Mean | Maximum | Mean | Mean | 3.8285502 |
| 45 | Maximum | Maximum | Mean | Mean | 6.6075783 |
| 46 | Minimum | Minimum | Maximum | Mean | 15.8284443 |
| 47 | Mean | Minimum | Maximum | Mean | 27.0293844 |
| 48 | Maximum | Minimum | Maximum | Mean | 46.6491927 |
| 49 | Minimum | Mean | Maximum | Mean | 10.5375619 |
| 50 | Mean | Mean | Maximum | Mean | 17.9944286 |
| 51 | Maximum | Mean | Maximum | Mean | 31.0560373 |
| 52 | Minimum | Maximum | Maximum | Mean | 3.9010473 |
| 53 | Mean | Maximum | Maximum | Mean | 6.6616089 |
| 54 | Maximum | Maximum | Maximum | Mean | 11.4970683 |
| 55 | Minimum | Minimum | Minimum | Maximum | 3.9923116 |

| | | | | | |
|---|---|---|---|---|---|
| 56 | Mean | Minimum | Minimum | Maximum | 6.8174562 |
| 57 | Maximum | Minimum | Minimum | Maximum | 11.7660404 |
| 58 | Minimum | Mean | Minimum | Maximum | 2.6578247 |
| 59 | Mean | Mean | Minimum | Maximum | 4.5386246 |
| 60 | Maximum | Mean | Minimum | Maximum | 7.8330742 |
| 61 | Minimum | Maximum | Minimum | Maximum | 0.9839373 |
| 62 | Mean | Maximum | Minimum | Maximum | 1.6802168 |
| 63 | Maximum | Maximum | Minimum | Maximum | 2.8998352 |
| 64 | Minimum | Minimum | Mean | Maximum | 6.1162995 |
| 65 | Mean | Minimum | Mean | Maximum | 10.4444763 |
| 66 | Maximum | Minimum | Mean | Maximum | 18.0258041 |
| 67 | Minimum | Mean | Mean | Maximum | 4.0718395 |
| 68 | Mean | Mean | Mean | Maximum | 6.9532617 |
| 69 | Maximum | Mean | Mean | Maximum | 12.0004229 |
| 70 | Minimum | Maximum | Mean | Maximum | 1.5074112 |
| 71 | Mean | Maximum | Mean | Maximum | 2.5741251 |
| 72 | Maximum | Maximum | Mean | Maximum | 4.4426042 |
| 73 | Minimum | Minimum | Maximum | Maximum | 10.6422520 |
| 74 | Mean | Minimum | Maximum | Maximum | 18.1732023 |
| 75 | Maximum | Minimum | Maximum | Maximum | 31.3645773 |
| 76 | Minimum | Mean | Maximum | Maximum | 7.0849281 |
| 77 | Mean | Mean | Maximum | Maximum | 12.0985512 |
| 78 | Maximum | Mean | Maximum | Maximum | 20.8805217 |
| 79 | Minimum | Maximum | Maximum | Maximum | 2.6228685 |
| 80 | Mean | Maximum | Maximum | Maximum | 4.4789317 |
| 81 | Maximum | Maximum | Maximum | Maximum | 7.7300521 |

## Final Model Output

**Note: The final model is still in terms of the centered predictors (only changes the intercept coefficient)**

```
Call:
glm.nb(formula = customers ~ units_c + income_c + compdist_c +
    storedist_c, data = Store, link = log, init.theta = 4.637931311)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.3361873  0.0541336  43.156  < 2e-16 ***
units_c      0.0008511  0.0002638   3.226 0.001256 **
```

```
income_c    -0.0139509  0.0039865  -3.500 0.000466 ***
compdist_c   0.1563735  0.0456166   3.428 0.000608 ***
storedist_c -0.1293811  0.0301078  -4.297 1.73e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(4.6379) family taken to be 1)

    Null deviance: 194.85  on 109  degrees of freedom
Residual deviance: 112.41  on 105  degrees of freedom
AIC: 691.89

Number of Fisher Scoring iterations: 1

              Theta:  4.638
          Std. Err.:  0.891

 2 x log-likelihood:  -679.893
```