Ratgression: A Study of NYC's Rat Population

Clarity Kummer (ckk2129), Gethin Wade (gw2508), Jane Zhang (jz4024)

December 19th, 2025

I. Introduction

This study examines how human population, as well as environmental and infrastructure factors, may impact rat inspection frequency across the five boroughs of New York City. Our motivation for studying these factors stems from the city's longstanding and persistent rat problem. New York has consistently ranked among the most rat-infested urban environments in the United States, and despite decades of mitigation strategies such as trash reforms and poison control, it has struggled to achieve substantial population declines. It is our expectation that rat inspections scale with population, driven largely by increased human activity and waste generation, making population a key explanatory variable in the analysis. In addition, we assess whether features such as quantity of litter baskets throughout the city, food scrap dropoff locations, DSNY garage presence, and park acreage help to explain rat inspection frequency beyond population-related variation. A better understanding of what drives New York City's rat population could lead to the development of more effective pest control measures.

Exploratory data analysis reveals a positive association between population, litter basket quantity and rat inspections, alongside the more moderate relationships between garage presence and food scrap dropoff presence. (**Fig. 1 - Fig. 4** below). These patterns motivate a multiple linear regression framework to jointly assess the contributions of population and environmental covariates to rat inspection frequency.

Fig. 1: Population vs. Inspections by Zip

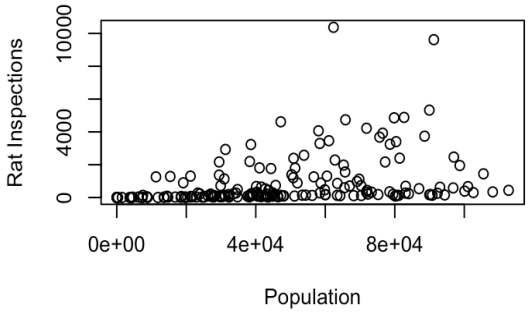


Fig. 2: Litter Baskets vs. Inspections by Zip

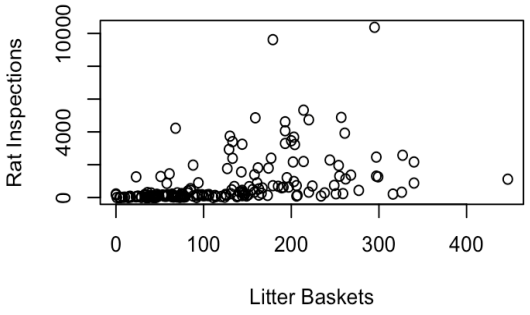


Fig. 3: Inspections by Garage Presence

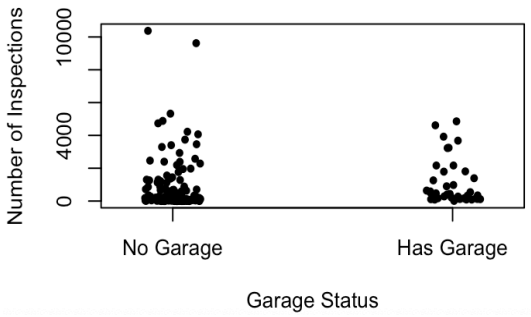
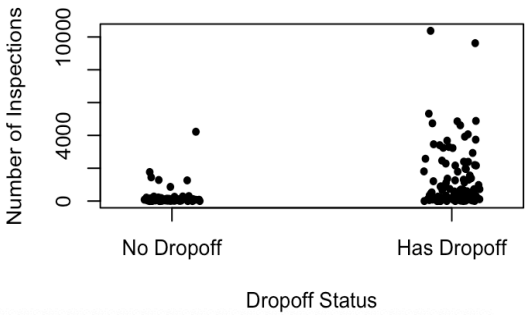
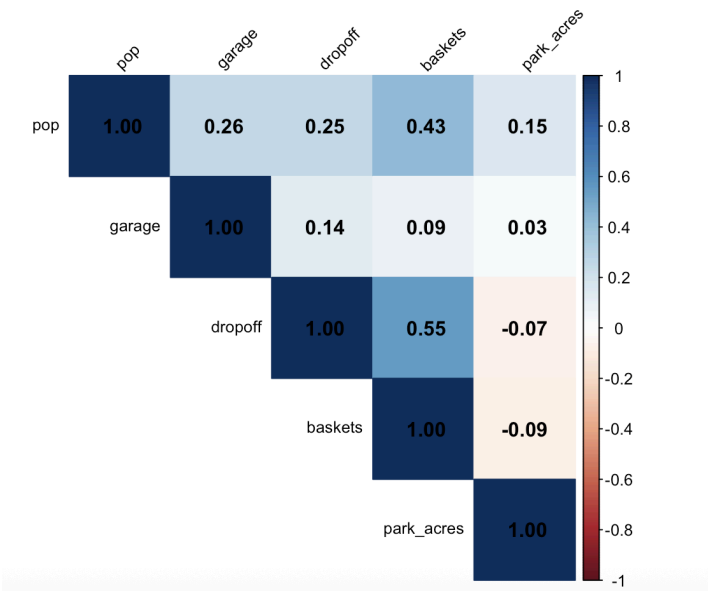


Fig. 4: Inspections by Dropoff Presence



Examining the correlation matrix below (**Fig. 5**), relatively weak correlations can be observed between total park acreage and the other covariates. However, the correlation of 0.15 between population and park acreage is worth noting: it indicates that park availability does not necessarily decrease with population density, implying that rat inspections are not simply driven by the absence of green space, but rather by human activity and waste dynamics within these areas. There are stronger correlations between population and the other three covariates, most notably litter baskets (r = 0.43). This makes sense intuitively: more densely populated neighborhoods are likely to have a higher number of litter baskets. The same line of reasoning applies to food scrap dropoff locations (r = 0.25).

Fig. 5: Correlation Matrix of Predictor Variables



Initial exploratory data analysis suggests increasing variability in rat inspection counts as population size and litter basket density increase (**Fig. 1** and **Fig. 2**). In addition, moderate correlations among several predictors foreshadow potential concerns when considering interaction effects.

II. Data Collection and Data Description

The data used in this analysis were obtained from multiple publicly available New York City Open Data sources. The outcome variable of interest is the count of initial rat inspections by ZIP code during the 2023 calendar year, collected from the NYC Department of Health and Mental Hygiene. Population estimates by ZIP code were obtained from the U.S. Census Bureau's 2023 American Community Survey.

To examine the role of environmental and waste infrastructure, additional datasets from the NYC Department of Sanitation were incorporated, including the number of DSNY litter baskets, DSNY garage presence, and food scrap drop-off locations. Park acreage by ZIP code was also included to capture access to green space. All datasets were spatially aggregated to the ZIP code level and merged to form a single analytic dataset.

All API queries and data wrangling were done with Python. For some of the data being used, no column for zip code was available, and so a function was created to convert latitudinal and longitudinal data into the appropriate format. From there, a .csv file was created with columns for zip code, Inspections, Population, DSNY Garage Presence, Dropoff Location Presence, and Park Acreage. All subsequent data analysis was performed with R (in the code file attached).

The rat inspection dataset contains approximately 157,000 inspection records across the five New York City boroughs in 2023. The litter basket, food scrap drop-off, garage, and park datasets contain approximately 25,000, 600, 80, and 2,054 records, respectively. After aggregation and merging, the final working dataset consists of 182 ZIP codes. Summary statistics for the continuous variables are presented in **Fig. 6**.

Fig. 6: Descriptive Statistics

Variable	Median	Mean	SD	Min	Max
inspections	183.00	855.35	1,527.77	1.00	10,373.00
pop	43,165.50	47,196.59	27,434.76	0.00	112,750.00
baskets	103.50	122.38	86.90	0.00	447.00
park_acres	105.24	464.02	784.66	0.00	5,558.26

Food scrap dropoff location and DSNY garages are binary predictors, with 1 indicating a given zip code has a location and 0 indicating a given zip code does not have a location. As such, these variables have less meaningful summary statistics. 42 of the 182 zip codes have a DSNY garage, or roughly 23%. Similarly, 117 of the 182 zip codes have a food scrap dropoff location, or roughly 64%.

III. Statistical Model

After iterating through multiple regression models and conducting residual analysis throughout, we fit the multiple linear regression model below using the log transformation of rat inspection counts as the response variable. The final model is specified as follows:

$$\log(\text{Inspections}) = \beta_0 + \beta_{1,\text{pop\_c}} + \beta_{2,\text{baskets\_c}} + \beta_{3,\text{garage}} + \beta_{4,\text{dropoff}} + \beta_{5,\text{pop\_c*baskets\_c}} + \beta_{6,\text{pop\_c*garage}} + \beta_{7,\text{pop\_c*dropoff}}$$

Given the non-constant variance observed in the exploratory analysis, rat inspection counts were log-transformed prior to model fitting. The log transformation serves to stabilize variance and improve adherence to linear model assumptions while preserving interpretability of regression effects on a multiplicative scale. This transformation choice was guided by a Box–Cox transformation analysis, discussed in ***Model Selection.***

In the final model, population (pop_c) and litter basket (baskets_c) counts are treated as continuous ZIP-code-level predictors and were centered to reduce multicollinearity, particularly given the inclusion of interaction terms. Garage (garage) and food scrap drop-off location (dropoff) presence are included as binary indicators denoting whether a ZIP code contains a DSNY garage or a food scrap drop-off location, respectively. The model further includes interaction terms between population and each infrastructure variable (garage, dropoff, baskets_c) to assess whether the relationship between population and rat inspection frequency varies with waste infrastructure.

Below is the R summary output of the final model:

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.6838 -0.8749 -0.0424  0.7894  3.8343

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.181e+00  1.897e-01  27.314  < 2e-16 ***
pop_c            2.275e-05  6.835e-06   3.329  0.00106 **
baskets_c        8.746e-03  1.392e-03   6.285 2.55e-09 ***
garage           4.703e-01  2.410e-01   1.951  0.05267 .
dropoff          4.400e-01  2.304e-01   1.910  0.05782 .
pop_c:baskets_c -2.247e-07  5.345e-08  -4.204 4.18e-05 ***
pop_c:garage    -3.992e-05  9.040e-06  -4.416 1.76e-05 ***
pop_c:dropoff    2.184e-05  8.166e-06   2.675  0.00818 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.215 on 174 degrees of freedom
Multiple R-squared:  0.6445,    Adjusted R-squared:  0.6302
F-statistic: 45.06 on 7 and 174 DF,  p-value: < 2.2e-16
```

The summary output above indicates that at a standard significance level of $\alpha = 0.05$, population size and litter basket density are positively associated with rat inspection at baseline levels of other covariates. While garage presence and food scrap drop-off locations do not exhibit statistically significant main effects at the average population level, both variables demonstrate statistically significant interaction effects with population at a standard significance level of $\alpha = 0.05$, indicating their influence on inspection frequency is conditional on population size.

Specifically, the negative interaction effects between population and both garage presence and litter basket density suggest a diminishing marginal effect of population in areas with more extensive waste infrastructure. Although increases in population are generally associated with higher inspection activity, this relationship is attenuated in ZIP codes with greater sanitation infrastructure, consistent with a partially mitigating role of waste management capacity. In contrast, the positive interaction effect between population and dropoff indicates that inspection frequency increases more steeply with population in ZIP codes that contain food scrap drop-off locations than in ZIP codes without them. Thus, the presence of drop-off locations amplifies the population-inspection relationship, particularly in higher density areas where waste generation is inherently greater.

The R^2 for our model is 0.64, indicating that our predictor variables account for roughly 64% of the variation in our outcome variable.

IV. Research Question

The main focus of our research is to assess the extent to which population size explains the variation in rat inspection frequency across New York City ZIP codes, and to evaluate whether environmental or waste-management infrastructure variables provide additional explanatory power beyond population alone. More specifically, we wanted to address:

1. How much of the variation in our outcome variable does our model explain, and are these effects statistically significant?
2. Do areas with a larger number of litter baskets explain any of the variation in rodent activity?
3. Does proximity to DSNY garages or to food scrap dropoff locations explain any of the variation in rodent activity?
4. Does proximity to green space (as measured by park acreage) impact rat inspection frequency?

Addressing our first research question, our final model accounts for approximately 64% of the variation in our outcome variable, log transformed rat inspection frequency. Population and litter basket density exhibit statistically significant main effects at the $\alpha = 0.001$ level. While the main effects of garage presence and food scrap drop-off presence are not statistically significant at the 5% level, both variables were retained due to their statistically significant interaction effects with population.

To formally assess the contribution of non-population predictors, we compared the final model (as provided in ***Statistical Model***) to a reduced model including only population via ANOVA. The reduced model yielded an R^2 of 0.39 and produced an F-statistics of 20 with a p-value approximately 0, leading us to reject the null hypothesis that the reduced model is sufficient. Thus, we reject the null hypothesis that the reduced model is adequate in explaining rat inspection frequency, and therefore accept the hypothesis that infrastructure variables and their interactions with population significantly improve model fit.

These results also address our second and third questions. All retained interaction terms are statistically significant at the $\alpha = 0.001$ level. The negative interaction coefficients for Population*Baskets as well as Population*Garage indicates that, as population increases, the marginal association between these infrastructure variables and rat inspection frequency weakens. In contrast, the positive Population*Dropoff interaction suggests that the association between population and rat inspections becomes stronger in ZIP codes containing food scrap drop-off locations.

Finally to test our fourth research question, we compared our final model (as provided in ***Statistical Model***) to an augmented model including park acreage by ZIP code via ANOVA. The ANOVA output provided an F statistic of 0.35 and a p-value of 0.55, indicating no evidence that park acreage improves explanatory power. We thus failed to reject the null hypothesis that our reduced model is adequate and exclude park acreage from the final model specification.

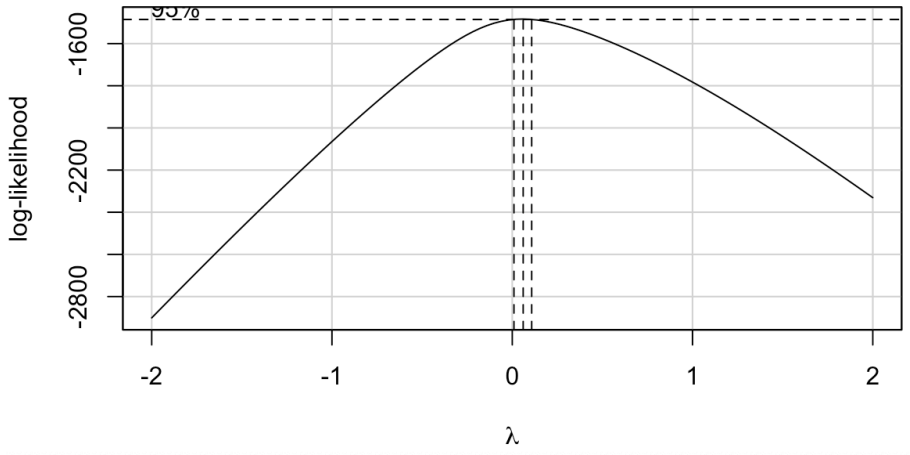
V. Appendix

Model Selection:

An initial, base model was fit to the data, which considered only the main effects of all predictor variables: population, baskets, garages, dropoff and park acreage. Such a model, as deliberated in ***Diagnostics and Model Validation,*** was a poor fit ($R^2 = 0.24$) and violated many regression assumptions. Given the heteroscedasticity, non-normality of residuals, and multicollinearity observed in the initial, base model's diagnostics, our first corrective step was to apply a log transformation to rat inspection counts. This transformation was suggested by a Box-Cox analysis (**Fig. 7**), which yielded an estimated lambda value of 0.05050505, shown below. After applying the log transformation, the model performance increased significantly, generating an $R^2 = 0.57224$. Such transformation additionally improved adherence to the normal error assumption as indicated by a Shapiro-Wilk normality test providing a p-value = 0.3099 and W=.99096. It further adheres to the constant error variance assumption as indicated by a Breush-Pagan test yielding a p-value = 0.3441, which suggests error variance is constant under the log transformation.

Following the stabilization of model assumptions, interaction terms were introduced to evaluate whether the association between population size and rat inspection frequency varied across levels of waste management infrastructure, as suggested in the correlation plot. Specifically, all two-way interactions between population and infrastructure variables were initially included. Statistical significance of interaction terms was assessed using standard t-tests, and only interactions providing evidence of effect modification were retained in the final model. Non-significant interaction terms were excluded to preserve model parsimony. The final model is provided in ***Statistical Model*** and is further justified in ***Diagnostics and Model Validation.***

Fig. 7: Box-Cox Transformation



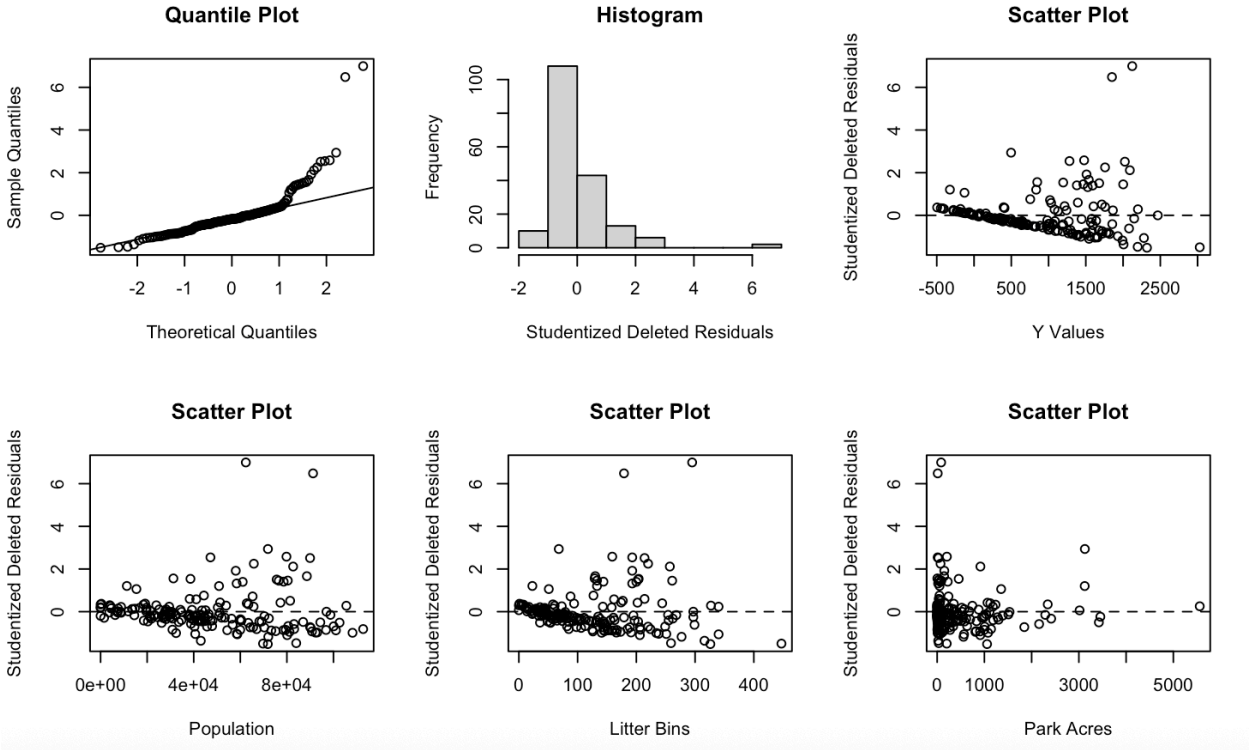
Diagnostics and Model Validation:

The initial model we fit to the data is specified below. The summary output indicated that only population and baskets were statistically significant predictors of rat population, and the R^2 was 0.24.

$$\text{Inspections} = \beta_0 + \beta_1(\text{Population}) + \beta_2(\text{Baskets}) + \beta_3(\text{Garages}) + \beta_4(\text{Dropoffs}) + \beta_5(\text{Park Acreage})$$

However, as stated in our exploratory data analysis, there were concerns about heteroscedasticity, multicollinearity, and non-normality of errors.

Fig. 8: Residual Plots



Looking at **Fig. 8**, it is immediately clear that many of the model assumptions are being violated. The quantile plot illustrates a serious deviation from normality (likely right skewed), as does the histogram. A formal Shapiro-Wilk normality test computed a p-value of less than 2.2e-16, therefore providing strong empirical evidence against normality of the residuals. The scatter plots against the deleted residuals suggest non-constant error variance, which is more formally verified by Breush-Pagan test yielding a p-value = 0.01098. While the full model selection process is detailed above, taking the log-transformation of rat inspections and centering the continuous predictors addressed these fundamental issues.

Fig. 9: Residual Plots

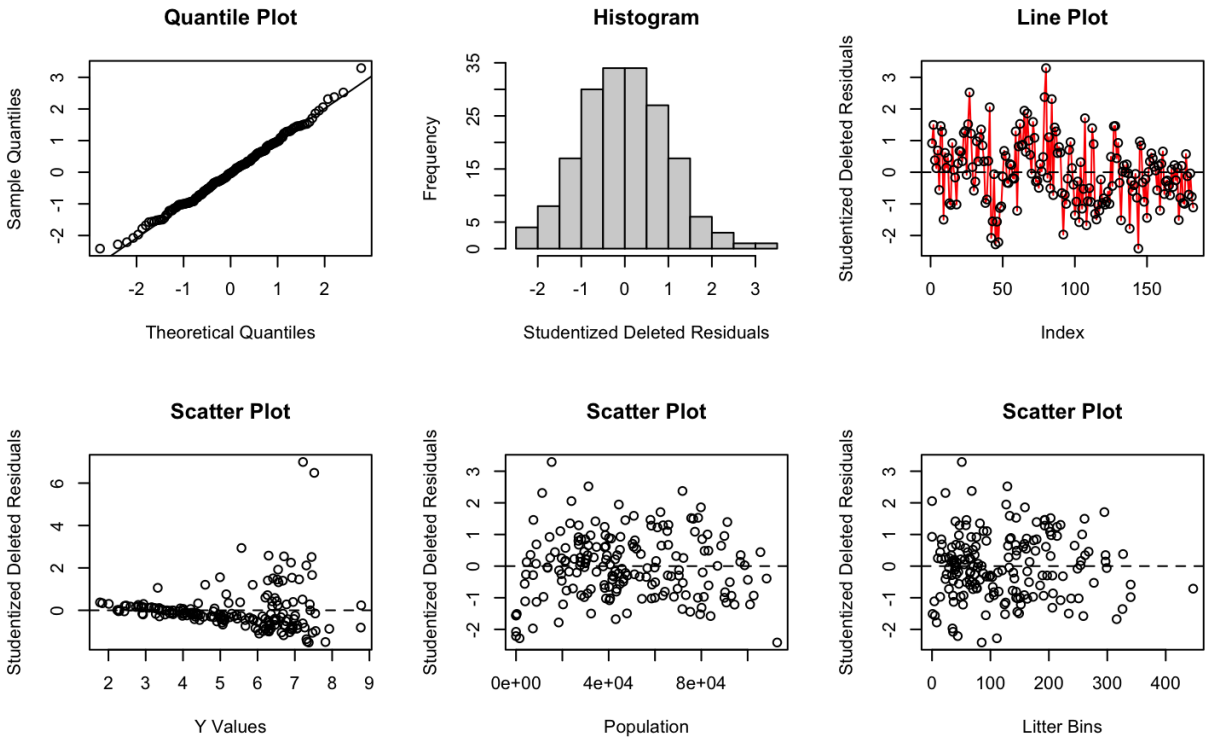


Fig. 9 shows the residual plots for our final model. There is a marked improvement in the quantile plot, histogram and scatter plots of covariates against the studentized deleted residuals. To further confirm that our model's assumptions are met, we performed two diagnostic tests. The Breusch-Pagan test for homogeneity of variance returned a p-value of 0.53. Therefore, we fail to reject the null hypothesis that our data has constant variance. The Shapiro-Wilk test for normality returned a p-value of 0.7877, and again we fail to reject the null hypothesis that our data is normally distributed. To assess multicollinearity, the Variance Inflation Factor (VIF) was calculated for each of the covariates. The average VIF across predictors was roughly 2.24, indicating that there are no major concerns about collinearity after centering the continuous variables (an average VIF much greater than 1 is indicative of multicollinearity).

To validate our model's results, we employed a k-fold cross validation and compared the RMSE (test error) to our multiple linear regression model's RMSE, the results of which are below. 10 folds were chosen, and each sample had roughly 164 of the 182 observations. We can see that the cross-validation R^2 is slightly lower than our model, by approximately 0.02. This is a very reasonable decrease, and certainly not so extreme as to suggest that our model is overfitting the data. Similarly, the cross-validation RMSE is only slightly higher than that of our model.

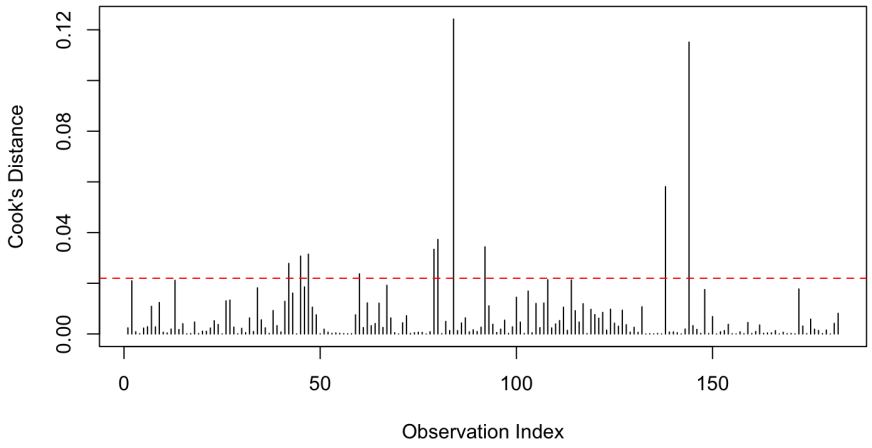
```
--- Standard Model ---          --- Cross-Validation ---

R-squared: 0.6444944            R-squared: 0.6248646

RMSE: 1.18775                   RMSE: 1.235642
```

Turning our attention to influential observations, we calculated the Cook's Distance for each observation and plotted them against the threshold value of 4 / n, which corresponds to the median of the F distribution with (p, n - p) degrees of freedom (**Fig. 10**). There are 10 observations in total that surpass the threshold, and only three that were over twice as high. However, this is a conservative threshold value, and no observations surpassed a Cook's Distance of 1, a commonly used cutoff which indicates an observation has a large influence on the model. The log-transformation of our outcome variable likely helped in addressing any influential observations, though it was initially performed to remedy the non-normality of our errors.

Fig. 10: Cook's Distance



Further to this, we previously employed the Shapiro-Wilk test to assess the normality of our data. The result of such a test could indicate whether or not a robust regression model might be more appropriate to address concerns about the data being from a heavier-tailed distribution than the normal (i.e. a larger number of outlying observations). The test yielded a p-value of 0.7877, indicating that we would fail to reject the null hypothesis that our data is normally distributed. Therefore, a robust regression was deemed unnecessary. By our assessment, no remedial measures were required to address influential observations.