

Exploring the Influences of the *Felis catus* on Animal Rescue Costs Using Propensity Score Matching and Regression Analyses

Sophie Berkowitz

December 17, 2021

Abstract

Since the rise of climate change, anthropogenic actions have posed a threat to planetary boundaries, specifically biodiversity loss. Currently, half of the world's species are threatened with extinction by mid-century as a result (Center for Biological Diversity, 2015). One such anthropogenic cause that has threatened biodiversity is the introduction of the domestic cat as a worldwide invasive species. This report will assess animal rescue costs by the London Fire Brigade (LFB), specifically investigating whether a causal relationship exists between an incident involving a cat and the cost of the rescue services. We apply propensity score matching (PSM) in order to arrive at causal estimates and address selection bias of the observational data. The objective is to derive the treatment (Cat) and comparison (not Cat) groups via single nearest neighbour matching of each observation's probability to belong in the treatment group. These probabilities, or propensities, are based on various predictors, which are calculated through logistic regression analysis. Finally, a linear regression analysis that assesses the influence of the treatment group versus the comparison group on incident notional cost is performed to test our hypothesis. Our findings suggest that cats cause greater incident costs relative to other species. Therefore, tax payers and policy makers should support cat regulation in order to cut municipal costs and more importantly, protect wildlife species.

Keywords: Causal Inference, Propensity Score Matching, Invasive Species, Cats, Observational Data, London, Climate Change.

I. Introduction

Extinction, the natural phenomenon that occurs when species no longer breed, which leads to termination, has increased dramatically since the rise of climate change (Thomas et al., 2004). Unlike past extinctions that were triggered by natural causes such as asteroid strikes or severe weather disasters, the current extinction crisis is a product of anthropogenic causes (Cahill et al., 2013). Anthropogenic causes are defined as activities deriving from humans, such as extracting fossil fuels or hunting wild animals (Cahill et al., 2013). Consequently, up to 50% of the world's species are at risk of extinction by 2050 (Center for Biological Diversity, 2015). Maintaining biodiversity by preventing further species' extinction plays a vital role in reinforcing ecosystem resilience and protecting the livelihood of the things that humans depend on, from food to the economy. Over twenty years ago the domestic cat, or *Felis catus*, was proclaimed by experts as a worldwide invasive species that have been introduced by humans and therefore pose as an additional consequence of anthropogenic actions (Lowe et al., 2000). They are estimated to be responsible for the deaths of up to 23 billion mammals per year (Loss et al., 2013).

Despite the risks that cats pose to wildlife and furthering species endangerment, regulation on outdoor cat populations is essentially non-existent due to pushback from animal welfare agencies against vital environment platforms (Loyd & Devore, 2010). This report will evaluate a dataset containing information regarding

animal rescue incidents performed by the London Fire Brigade (LFB) and explore whether the animal group, cats, influence costs disproportionately while accounting for various observational traits. The London Fire Brigade is funded by municipal taxpayer money and the central government (LFB, 2009). Therefore, outlining the excessive costs of outdoor cats on the citizens should incentivize the jurisdiction to focus their resources on preventing cat predation and further species endangerment in London. Additionally, if cat costs appear low, then this might imply a necessity to impose additional taxes on cat related animal service incidents attended to by the LFB in order to force society to internalize the externality of outdoor cat dangers on wildlife. This analysis will contribute to the plethora of academia surrounding the costs of anthropogenic causes on planetary boundaries and furthering the impacts of climate change with the goal of encouraging systemic change. **We will test whether a causal relationship exists between an incident being cat related and the cost of incident services via propensity score matching and regression analyses.** Regression analysis helps us gain further insight into understanding the significance of various variables and how they influence each other, and the response variable. Meanwhile, propensity score matching is used to infer causality in observational data by emulating the role that treatment and comparison groups play in experimental data. While thousands of species are threatened with extinction, conservation efforts have rescued many from extinction (Palomares et al., 2011). The goal of this research is to provide statistical insights that inform the implementation on outdoor cat regulation for the sake of wildlife protection and mitigating climate change.

II. Data

Outside of the traditional fire incidents that the London Fire Brigade (LFB) in the United Kingdom, routinely responds to, they also attend special services, or non-fire incidents. Many of the non-fire incidents regard animals that require rescuing or aid (LFB, 2009). This dataset consists of information on each animal-related incident that the LFB has responded to since January 2009 up until the present. It was also published on June 29, 2021 in TidyTuesday, a github data project that aims to improve the understanding of data analysis among R users (Mock, 2021). The original dataset was retrieved from the London Datastore, which is a data-sharing portal operated by the government that shares key information regarding the city of London’s inner workings (LFB, 2009).

a. Data Collection Process

Refining the dataset into a subsetting collection of information is a vital component of the data cleaning process because it promotes productivity of the subsequent analysis by improving overall clarity. It reduces noise in the dataset by removing non applicable or incomplete data, by getting rid of unimportant variables. Additionally, cleaning the data involves purifying the relevant variables to make them more consistent and readable for both the reader and the computing software. After obtaining the complete dataset, it was subsetting into a more refined version that included the variables of interest for this analysis. These pertinent variables are as follows – incident notional cost, animal group parent, latitude, postcode district, total pump hours, and calendar year (LFB, 2009). Animal group parent, which lists the type of animal that the fire incident regarded, was converted to a categorical variable by classifying the groups as levels or factors. This mechanism allows the computing software, R, to more efficiently classify the variables into their respective bins. Postcode district consisted of 274 postcodes that exist in London. In order to make the dataset and variable more digestible, the postcodes were reduced to their region, such as south or east, by keeping only the first character of the postcode district (Figure 1). Then, the variable was converted to categorical levels. Initially, the numeric variables: incident notional cost, latitude, and pump hours total were listed as characters instead of numbers. Therefore, these variables were all converted into numeric data.

Due to various reasons, such as incompleteness or computer error, NA and “NULL” values appeared throughout the dataset. All rows, or incidents, that included these values were omitted from the dataset in order to exclude incomplete or inaccurate data. This data hinders the analysis as it constricts the sample size and might neglect information addressing our hypothesis. It is possible that this neglected data demonstrates

incidents that have certain costs or are related to certain animal groups, such as cats. Hence, this might introduce bias into our analysis and acts as a limitation.

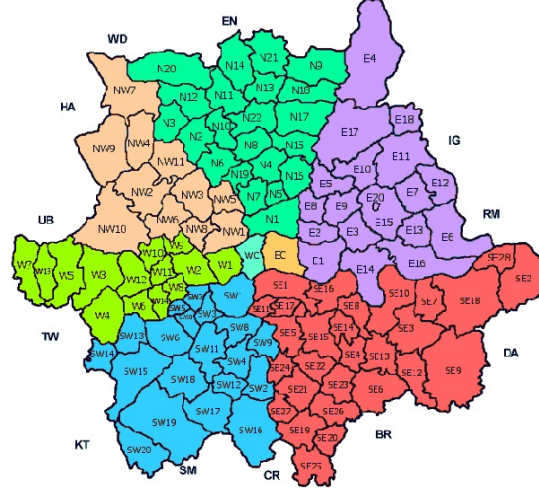


Figure 1. Photo by Michael Kourtis from L-Universita ta Malta. The figure is representative of the city of London, UK, postcode system. The postcode corresponds to the area designation with either one or two letters then a number. The central portion of London is divided into these regions: EC, WC, W, E, N, NW, SE, SW. The data cleansing process confined the regions to N, E, S, W, for North, East, South, West, respectively.

b. Data Summary

The relevant variables subsetting into a smaller dataframe for this analysis are incident notional cost, animal group parent, latitude, postcode district, pump hours total, and calendar year. Incident notional cost is a continuous variable that represents the total cost of the animal related incident (LFB, 2009). Therefore it demonstrates the financial impact that it had on the city of London, as it was paid for with both taxpayer and government funding. In this analysis, incident notional cost will be used as the response, or outcome, variable. This variable is of interest, since we are investigating whether cats disproportionately affect costs of LFB rescue incidents, relative to other animal groups. If that is true, then this analysis should contribute to the notion that outdoor cat management would be necessary to manage this nuisance. If this is not so, then it might be necessary for the city to impose an additional tax on cat incidents in order to force the city to internalize the externality. Animal group parent is a categorical variable that indicates the type of animal that required rescuing at the time of the incident. Twenty eight animal groups are included in the data. This analysis emphasizes the role of cats in animal rescue incidents, therefore this variable allows us to highlight specifically, the cat incidents and their corresponding characteristics. This variable will allow us to create a treatment variable for the propensity score matching by separating the cat group from the remaining 27 animal groups.

Postcode district is a categorical variable that demonstrates in what region in London the incident occurred. This variable is of interest for the analysis because a relationship might exist between animal incidents and the postcode district that they occurred in. Specifically, it is possible that cat incidents occur in specific districts based on various factors, such as topography or density. Latitude is a continuous variable that corresponds to the north-south coordinates where the animal incident occurred on the earth. Total pump hours is a continuous variable that reflects the total hours that the LFB attended the incident for, as pump is the number of trucks. Various animal groups or incidents might require longer rescue times. For instance, feral animals might be harder to rescue, as they are generally less comfortable around humans. Calendar year is the year that the incident occurred in, a continuous variable. Based on the data, it was reported that LFB animal rescues rose 20% in 2020, the pandemic year (Kampfner, 2021). Therefore, this variable was

included based on contextual significance and the role that the pandemic, or other timely events, might play in animal incidents.

Table 1: The following Table 1 demonstrates the numerical summaries of the variables that will be used in the propensity score matching and regression analyses. The summaries of the continuous variables, incident notional cost, latitude, pump hours total, and calendar year, indicate the Minimum, 1st Quartile, Median, 3rd Quartile, and Maximum values. The Maximum and Minimum reflect the greatest and lowest values, respectively for each variable. The 1st Quartile value is the cutoff where 25% of the data is located under that value. Likewise, the 3rd Quartile is the cutoff where 75% of the data lies below that value. The median is the value that distinguishes the upper 50% of the data from the lower 50% of the data. The mean, or average, is calculated by finding the sum of all elements and dividing it by the total number of elements. The summaries of the categorical variables – animal group parent and post code district, demonstrate the six most common categories for the variable. The last row, (Other), groups the remaining categories with each other and reflects this count. The most prevalent category of animal group parent, cats, reflects that the dominant animal group that accounts for LFB animal service incidents is significantly more presiding than any other animal group. The NA values that exist in the table as elements of the numerical values simply appear due to the presence of a 7th row of the table necessary from the categorical values. (continued below)

IncidentNotionalCost	AnimalGroupParent	Latitude
Min. : 0.0	Cat :1443	Min. : 0.00
1st Qu.: 260.0	Bird : 849	1st Qu.:51.46
Median : 326.0	Dog : 803	Median :51.52
Mean : 379.8	Fox : 231	Mean :51.48
3rd Qu.: 346.0	Horse : 190	3rd Qu.:51.57
Max. :3480.0	Deer : 127	Max. :51.69
NA	(Other): 209	NA

PostcodeDistrict	PumpHoursTotal	CalYear	treatment
S :893	Min. : 0.000	Min. :2009	Min. :0.0000
E :654	1st Qu.: 1.000	1st Qu.:2012	1st Qu.:0.0000
N :630	Median : 1.000	Median :2015	Median :0.0000
W :249	Mean : 1.259	Mean :2015	Mean :0.3746
T :220	3rd Qu.: 1.000	3rd Qu.:2018	3rd Qu.:1.0000
R :214	Max. :12.000	Max. :2021	Max. :1.0000
(Other):992	NA	NA	NA

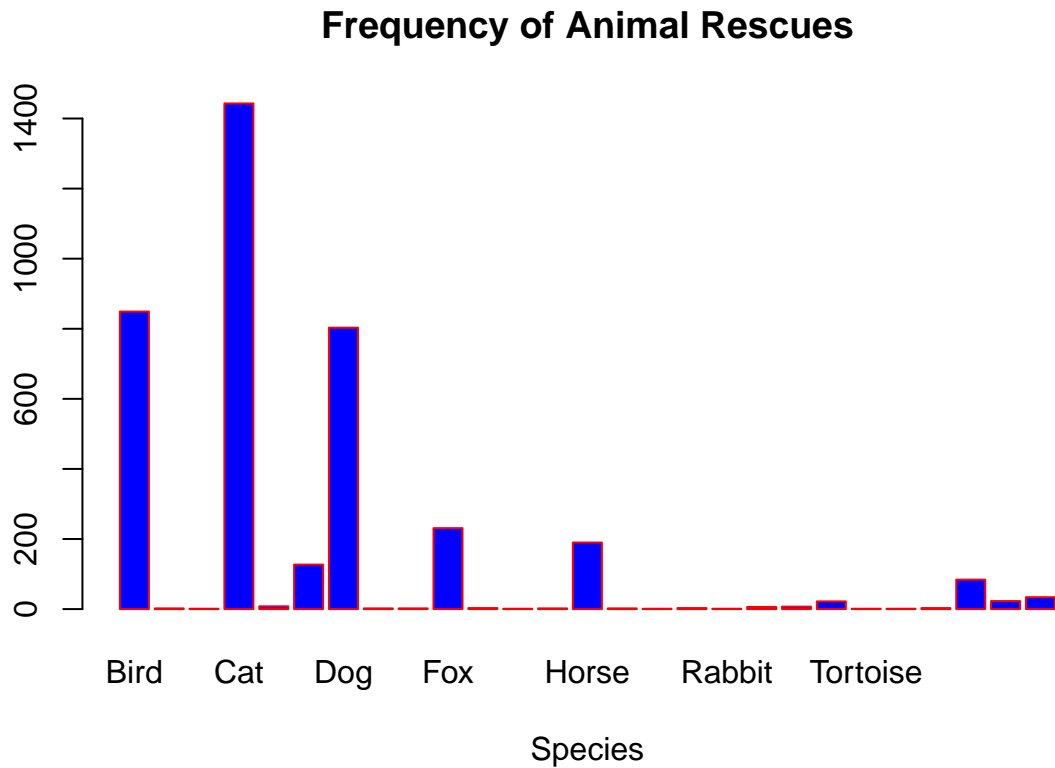


Figure 2. The barplot corresponds to the counts of animal incidents attended to by the LFB from 2009 until the present for each animal group. It is evident that there are certain dominant species that account for the majority of the animal rescues performed. Specifically, cats appear to be responsible for over 3,500 of the incidents, where the second most dominant group is birds, which account for less than half of cats at ~1500 incidents. It is worth investigating the role that the dominant group, cats, play in this context, as they are considered a lethal invasive species. Many of the 28 species are not explicitly labelled on the barplot for reasons of redundancy and to highlight the more prevalent groups.

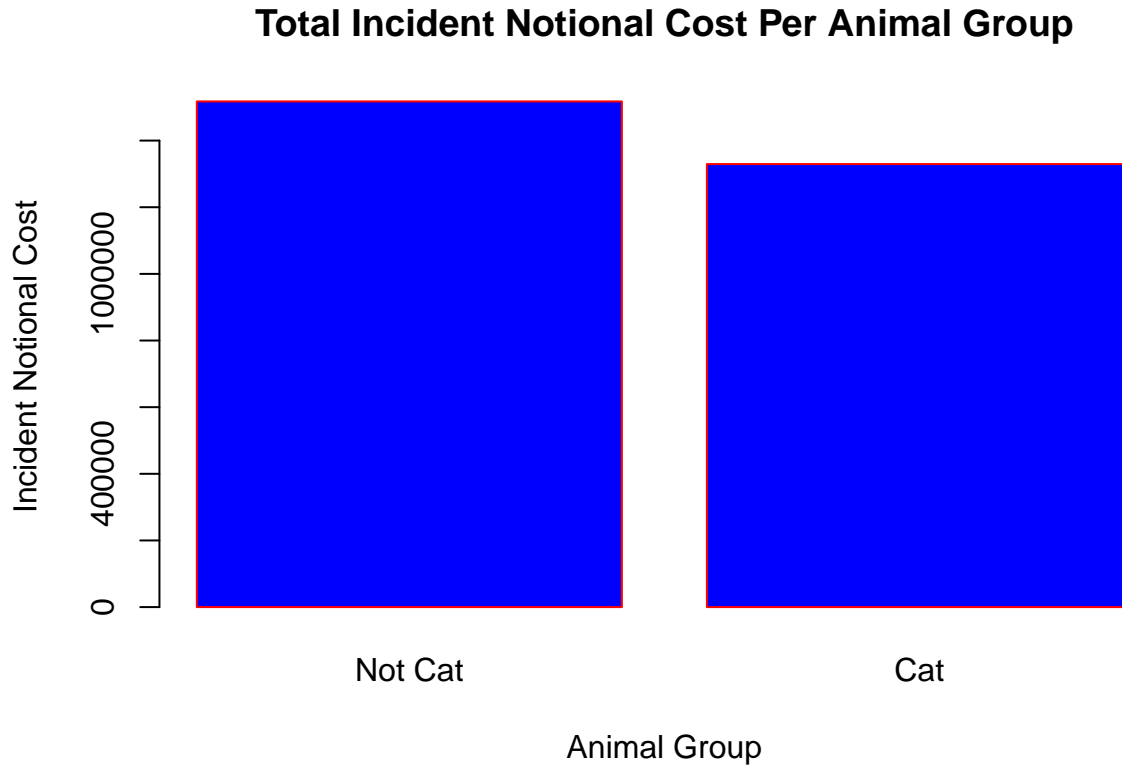


Figure 3. The plot reflects the total costs of the treatment group versus the control group. These values were obtained by grouping each incident by treatment or comparison group and calculating the sum of the costs to return one cost for each group. The treatment group includes the cat data and the control group consists of all of the remaining animal groups that are not cats, approximately 27 groups. Interestingly, the sum of cat costs amounts to nearly the sum of all remaining animal groups. In other words, only one species costs as much as all other 27 species combined. Even if individual incident costs are approximately similar among all species, the total species cost of cats alone is worth noting.

All analysis for this report was programmed using R version 4.0.4.

III. Methods

a. Propensity Score Matching

Experimental studies commonly involve the introduction of a treatment, randomly assigned to subjects in order to establish causal treatment effects. Randomly assigning participants to either the treatment or control group aids in offsetting the effects of various factors that might influence the research question, such as confounding variables (Caliendo & Kopeinig, 2008). Traditionally, observational data does not allow for the assignment of subjects to treatment or control groups, as the predictor variables are not manipulable by the researcher. Various techniques, such as Propensity Score Matching (PSM), permit us to infer causality in observational data and neutralize issues within observational data such as confounding. PSM essentially derives treatment and control groups within observational data by matching subjects based on various criteria and their likelihood to be in the treatment group. Using PSM, we assume that the observations within the data set possess the same global characteristics and perform a comparative analysis between the treatment and comparison group, similar to one for experimental data. It assigns a probability, or propensity to be

in the treatment group, to each observation based on the selected criterion. This propensity is retrieved through logistic regression, assigning the treatment as the response and the observational characteristics as the predictors:

b. Logistic Regression Analysis

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1x_1 + \dots + \beta_nx_n + \epsilon$$

Where p reflects the propensity of an observation being in the treatment (Cat) group. $\beta_1 \dots \beta_n$ are the logistic regression coefficients corresponding to the slope, or change in log odds for each one unit increase in the coefficients' independent variable, which serves as the criterion for the propensity derivations. β_0 is the intercept of the logistic regression line. The categorical variables, such as postcode, are inputted as indicator variables, where a reference category is assigned to one category that all of the remaining categories are compared to. These remaining categories are each assigned to a random variable and their coefficients demonstrate the difference in intercepts between the reference category and the category corresponding to the coefficient. When the categorical random variable is true for each observation, it will be set to 1 and the remaining categories set to 0. The continuous variables are each assigned to a coefficient multiplied by the respective predictor value.

Logistic Regression Analysis involves 5 assumptions – 1) Dependent variable is binary, 2) Observations are independent of each other, 3) No multicollinearity, 4) Linearity of independent variables & log odds, 5) large sample size (Peng et al., 2002). The first assumption indicates whether the treatment corresponds with the appropriate outcome type according to logistic regression, which assumes a binary outcome (e.g., Yes or No). Next, independence of observations requires that all animal incidents be independent of each other, meaning that they are not influenced by or related to each other. Independence can be evaluated based on contextual data or through the use of residual plots, where a randomly scattered residual plot implies independence. Multicollinearity refers to data that consists of heavily correlated independent variables (Peng et al., 2002). When this occurs, the estimated coefficients will be less precise and the model itself will be less powerful as a result. Multicollinearity can be checked using the Variance Inflation Factor (VIF), which evaluates the degree of multicollinearity between predictors (Fox & Weisberg, 2018). Logistic Regression rests upon the notion that the relationship between the log-odds of the outcome and the independent variable is linear. This can be evaluated visually via a scatterplot of the variable against the log-odds of the outcome. If the plot is non-linear then the assumption is violated. It is also worth investigating influential points, such as outliers, and where they affect the model. We can check this using standardized residual plots, which quantify the magnitude of the residuals in terms of the standard deviation. Therefore the outliers will appear more obvious.

In order to select the optimal model, we will consider various model selection methods that aid in selecting significant variables. The Akaike Information Criterion (AIC) estimates prediction error of the logistic regression model by selecting the model that describes an unknown operating model (Seber & Lee, 2012). It explores how compatible the model is with the data it was generated from. Typically, the lower the AIC value is, the better fitted the model is. A model's p-value provides insight on overall significance, indicating the probability of reproducing our data given the model. We can also investigate each coefficient's p-value, which tests the null hypothesis that assumes that the coefficient is 0. Similar to AIC, the lower the p-value, the better, as it indicates that the likelihood that the data fits the null hypothesis is very low. A p-value below the critical value 0.05 is commonly used when assessing models. The F-statistic evaluates the ratio of variance of the model in question to the model's unexplained noise. Unlike AIC and the p-values, a high F-statistic indicates a stronger model.

In this analysis, the propensity score reflects the propensity of an animal service incident to involve a Cat. The outcome of interest is the incident notional cost. Once the propensities are returned, the pairs are matched between a subject in the treatment group and a subject in the comparison group that should have similar propensities of being in the treatment group. The matching will be retrieved via nearest neighbour matching, which pairs the observations based on their closest unmatched propensity score (Gelman & Hill, 2006). The goal of propensity score matching is to retrieve a treatment (Cat) and comparison group (not

Cat) that look similar based on observable characteristics such as latitude, postcode, or total pump hours. If we fail to achieve similar looking groups that have a balance of characteristics, then the matches might've been poorly produced. The balance between treatment and comparison group can be evaluated by comparing the means of each group for the predictors, analyzing the histograms of each group for the predictors, and more. If the matches are satisfactory, we can evaluate our outcome by performing a linear regression on treatment against incident notional cost.

c. Linear Regression

We can use linear regression analysis to examine the effect of being in treatment group on the incident notional cost. In the regression model built subsequent to the propensity score matching, the response variable is the incident notional cost and the predictor variable is the indicator variable that reflects whether the observation is in the treatment group (Cat), or not. Linear regression is a statistical tool to investigate whether a response variable and predictor variables possess a statistical linear relationship. If so, then the model can predict the expected value of the response variable, incident cost, conditional on a value of the explanatory variable, treatment, alternatively known as the conditional mean response. Since observational and experimental data is imperfect, the difference between the independent realizations of the random variable Y and the conditional mean response, is reflected using the error term. The linear regression equation can be denoted as:

$$y = \beta_0 + \beta_1 x + \epsilon$$

Similar to the logistic regression equation, β_0 is the intercept of the regression line. β_1 represents the magnitude of the difference in incident notional cost between the treatment (Cat) and comparison group (not Cat). $x = 1$ corresponds to observations when the incident involves a Cat. $x = 0$ corresponds to observations when the incident doesn't involve a Cat. We substitute these values into the linear regression equation to compute fitted values for each dummy variable.

Linear Regression makes assumptions similar to logistic regression. Firstly, there should be a linear relationship between the predictor variable and the dependent variable. In other words, the conditional mean of Y given X is a linear function of X . However, since this analysis incorporates a dummy variable as the independent variable, the linearity assumption is not required as dummy variables are automatically linear (Seber & Lee, 2012). The second assumption evaluates whether the errors, ϵ are independent. In other words, the difference between a response and the mean doesn't depend on the distance of any other response. Independence of errors can be evaluated using a plot of the residuals versus the fitted values. There should exist a correlation of 0. The next assumption investigates whether the errors exhibit constant variance. This holds when the plot of the standardized residuals appear uniform with no identifiable pattern. Finally, the errors should be normally distributed, which is checked using a normal Q-Q plot that evaluates whether two quantiles, the sample and the theoretical, are normally distributed.

IV. Results

a. Logistic Regression Model Selection

In order to observe the role that cats played in animal related incidents against all other species combined, the data was filtered to separate the treatment data group, Cats, from the rest of the animal group data. The logistic regression analysis to evaluate each observation's predicted propensity to involve a Cat or not was performed twice. The first, full, model incorporated all of the predictor variables – latitude, post code district, pump hours total, and calendar year (Table 2). Based on the model's AIC and each variable's statistical significance, a second model was performed, which only kept postcode district and pump hours total. In order to select the optimal model, AIC, residual deviance, and vif were considered (Table 2).

Table 2. Model statistics for the full model and adjusted model outputs. The full model includes all predictor variables included in the subsetted data frame created for this analysis. The adjusted model eliminated the predictor variables that did not appear to be significant. Significant variables, pump hours total and post code district remained in the model. The AIC is a measurement of the model's goodness of fit, where a small value indicates a better model. The Residual Deviance demonstrates the degree to which the outcome variable, or treatment, can be forecasted by the model, given its predictors. The lower the Residual Deviance, the better. *SignificantPredictorVariables/TotalPredictorVariables* indicates how many of the model's predictors deemed significant versus the total predictors in the model. The vif approximation reflects a value for how the vif for the predictors in each model scored. All variables consisted of vif scores near 1, demonstrating a lack of multicollinearity for both models.

	AIC	Residual Deviance	Significant Predictor Variables / Total Predictor Variables	vif
Adjusted Model	5049.7	5019.7 on 3837 df	3/14	~1
Full Model	5051.9	5017.9 on 3835 df	3/16	~1

Logistic Regression Equation:

$$\begin{aligned}
\log\left(\frac{p_{Cat}}{1 - p_{Cat}}\right) = & -0.018654 - 0.301983x_{PumpHoursTotal} - 0.335333x_{PostCode\ C} \\
& -0.412397x_{PostCode\ D} - 0.412397x_{PostCode\ D} \\
& -0.111813x_{PostCode\ E} - 0.205498x_{PostCode\ H} \\
& -0.076725x_{PostCode\ I} - 0.68805x_{PostCode\ K} \\
& -10.641451x_{PostCode\ M} + 0.160176x_{PostCode\ N} \\
& -0.353094x_{PostCode\ R} - 0.008019x_{PostCode\ S} \\
& -0.460413x_{PostCode\ T} - 0.325926x_{PostCode\ U} - PostcodeDistrictWx_{PostCode\ W}
\end{aligned}$$

Model Assumptions:

The dependent variable is our treatment variable, which indicates whether the observation involves a cat, or not. Therefore, it is binary because it only takes in the values 0 (not Cat) or 1 (Cat). Multicollinearity is explored in Table 2, demonstrating no highly correlated predictor variables. The large sample size assumption is met, as the sample size contains 3,852 observations. The linearity of independent variables and log-odds assumption is explored in Figure 7, and ultimately failed for many variables due to the behavior of the plots. Many of the variables that were incorporated into the initial logistic regression model, such as calendar year and latitude, violate this linearity assumption, which supports the decision to remove them from the selected model. Only the continuous variable, pump hours total is extended into the logistic regression model, which eliminates the possible issues brought with the remaining continuous variables (Figure 7). However, total pump hours, appears to exhibit a less linear and more quadratic relationship in this plot, which also violates the assumption and suggests the necessity of a transformation of this variable (Figure 7).

b. Propensity Score Matching

The logistic regression model allowed us to predict the propensity of each observation being in the treatment group (involving Cats), given the predictors postcode district and total pump hours. Using the predicted propensity, the matching was performed between incidents that involved a cat and incidents that didn't involve a cat, in order to match the treatment and control group. The matches were evaluated via single nearest neighbor matching, where the score represents the observation's predicted propensity and the vector of indicators is the treatment variable that indicates whether the observation is in the treatment group (Gelman & Hill, 2006). The data was reduced to contain only the paired matches for reasons of clarity and conciseness, as this will be the only relevant information for the remaining analyses. This new sample contains 2,886 observations, which theoretically should contain an equal number of treated (Cat) and non-treated (not Cat) incidents.

Matching Evaluation:

Once the data was matched to incorporate pairs of animal incidents involving cats and not involving cats that have the same propensity to be either species, the quality of this matching was assessed. The goal of this evaluation was to address whether the reduced data frame of the matched observations were truly similar to each other subsequent to accounting for the treatment group. We explored the balance of observable traits between the two groups, accounting for variables such as latitude or total pump hours. The side-by-side histogram showing the the comparison and treatment group spread of latitude demonstrates a strong quality of matching as the distributions appear approximately analogous (Figures 8 & 9). A two-sample t-test is commonly used when evaluating whether the population means between two groups are equal. Therefore, a two sample t-test was performed to investigate the balance among total pump hours between both groups. The means of pump hours total were 1.160 and 1.168 for the comparison and treatment group, respectively. The null hypothesis of the t-test assumed that the true difference in means between the treatment group and the comparison group would be zero, or there would be no difference. Since the 95% confidence interval included zero, we could expect there to be no difference between the two groups. Additionally, the p-value was 0.6918, meaning we fail to reject the null hypothesis and can assume that the propensity score matching was successful.

Linear Regression:

$$\hat{y} = 328.420 + 23.549x_{treatment} + e_i$$

Model Interpretations:

- The incident notional cost when cats are not present at the incident, or the predictor is equal to 0, is €328.420
- When cats are involved with an incident, incident notional cost is €23.549 greater than when all other species are involved with an incident.

Table 3. The summary output of the linear regression model indicates high significance of the intercept of regression equation. In the context of the dataset used in this analysis, a significant intercept demonstrates that incident notional cost for the comparison group, which includes animals that aren't cats, is significantly different from zero. More importantly, the treatment coefficient appears significant with a p-value of less than 0.001, meaning that the average difference in incident notional cost between the comparison and treatment group is significant. The model had an incredibly low R^2 and an incredibly high AIC, where both are indicative of a poor model. However, when working with indicator variables in linear regression, both of these metrics are expected to be lower and higher for R^2 and AIC, respectively, therefore these are less significant for our model. Additionally, the model's p-value of 0.00002615 indicates that the explanatory variable, treatment, is linearly related to incident notional cost, as it is significantly less than the critical value 0.05.

	(1)
(Intercept)	328.420 ***
	(3.954)
treatment	23.549 ***
	(5.592)
N	2886
R2	0.006
logLik	-18558.505
AIC	37123.011

*** p < 0.001; ** p < 0.01; * p < 0.05.

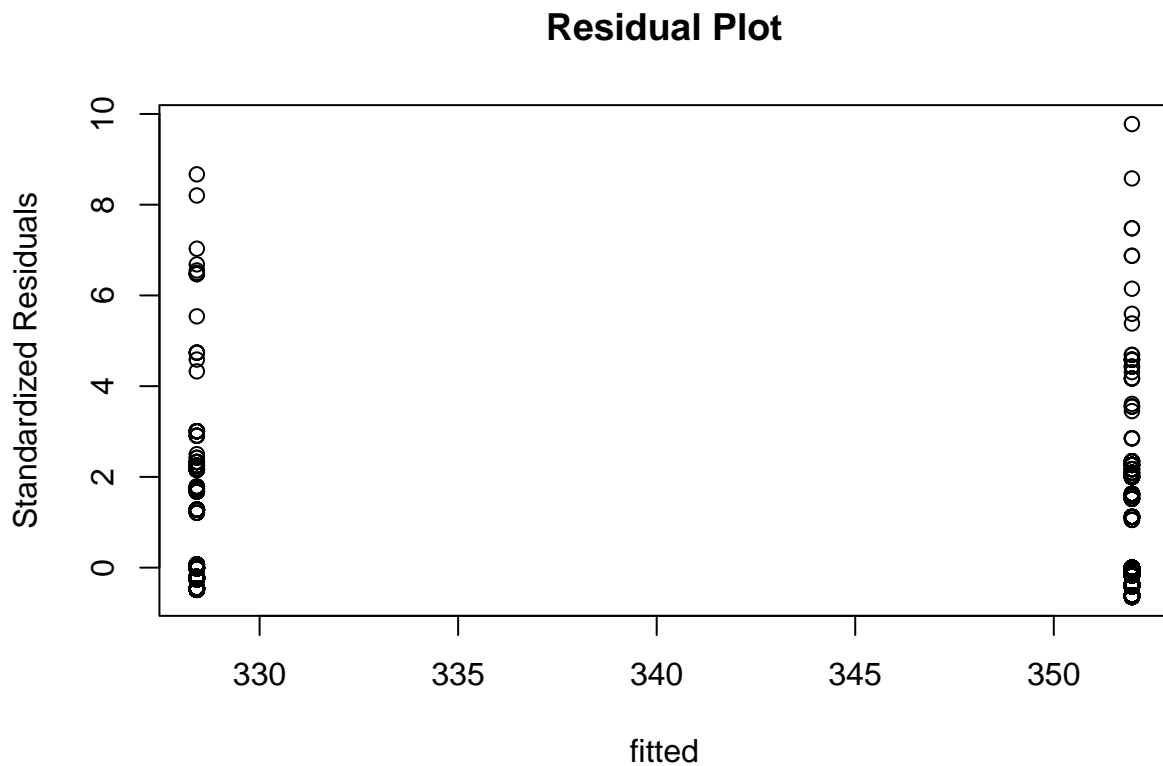


Figure 4. The residuals vs. fitted values plot demonstrates a scatter plot showing the residuals of the linear regression surface's residuals on the y-axis and the estimated response on the x axis. This plot allows us to visualize the homogeneity of variance by comparing the spread between the comparison and treatment groups. It appears that the comparison group located at the lower end of the x-axis exhibits greater spread than the treatment group, as the fitted treatment residuals are clustered closer together. However, there still appears to be approximate symmetry between the two groups.

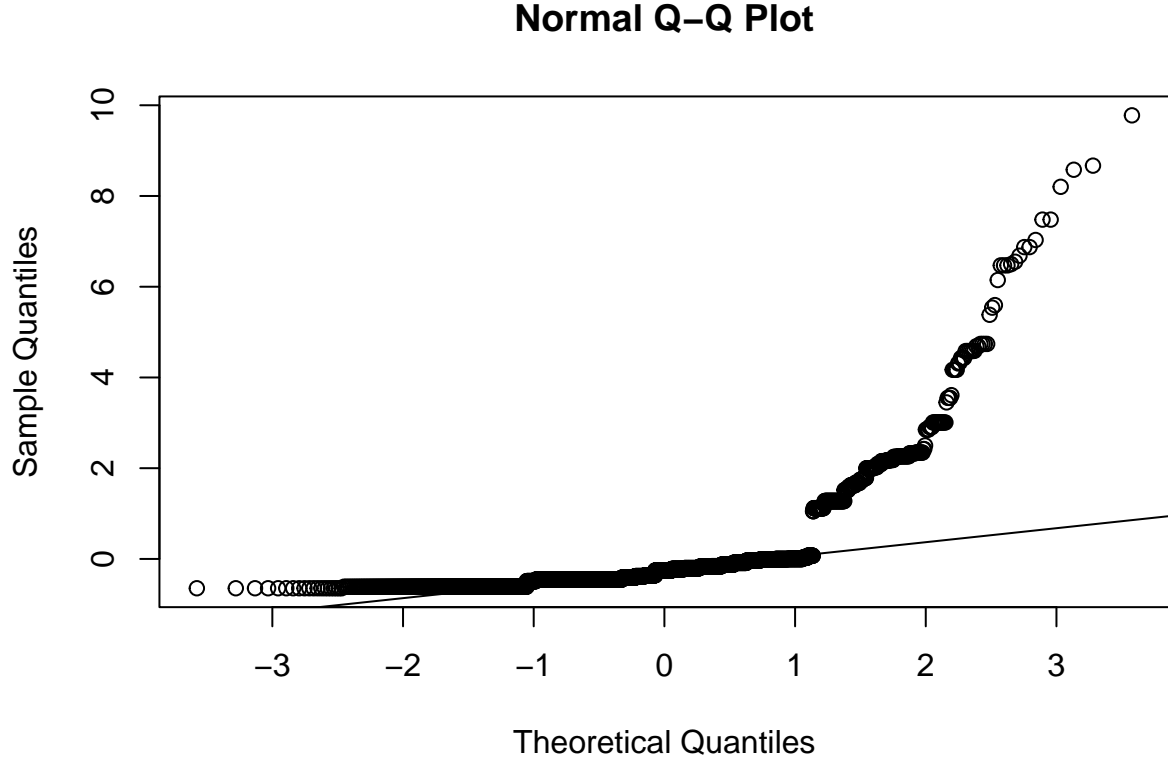


Figure 5. The Normal Q-Q plot compares the probability distribution of the sample quantiles versus the theoretical quantiles. This visualization helps to identify if both datasets belong to the same family of probability distributions, which in this case is the normal distribution. The q-q line signifies where the data points should fall theoretically. The data that lies along this line indicates that these points do belong to the same distribution. However, most of the data points exhibit behaviour that does not fall along the line. The plot appears to possess a heavy upper tail, implying greater variation at those points. Additionally, the q-q line should possess a steeper slope that exists more diagonally across the plane, which is where the data points should also be. This behavior violates the normality assumption of linear regression.

V. Conclusions

As anthropogenic factors continue to aggravate various consequences of climate change, such as biodiversity, it is vital to consider how municipalities can mitigate these effects. Since the turn of the century, the domestic cat has been considered by experts as an invasive species, responsible for millions of wildlife deaths annually (Loss et al., 2013). In spite of the threats that this species poses, there is a lack of global effort to manage cats, preventing their further destruction. This report investigates data provided by the London Fire Brigade (LFB) on animal rescue incidents from 2009 until the present. The LFB services are funded by city tax payers and the central government, implying that the burden that cats pose on rescue incidents is imposed on these taxpayers and the government. This report investigates whether a causal relationship exists between an incident being cat related and the cost of incident services using propensity score matching and regression analyses. If so, then this analysis should incentivize the city to impose further regulation on outdoor, invasive cats given the proof that they are costing the city so much.

The variables used in the analysis include incident notional cost, animal group parent, latitude, postcode district, pump hours total, and calendar year. These variables were incorporated into the initial report due to their contextual significance. The treatment variable took in values 1 or 0 if the incident involved a cat

or not, respectively. Propensity score matching is a statistical tool that allows researchers to infer causality in observational data by deriving a treatment and comparison group via matching subjects based on their likelihood to belong in the treatment group, while accounting for various predictors. The initial propensity score matching involved a logistic regression analysis that incorporated the treatment variable as the response and various observational characteristics as the predictors in order to determine an observation's propensity to be in the treatment group. A model selection process involving AIC and significance tests was used to select the optimal model and predictors. Once this logistic regression model was obtained, the matches were performed using single nearest neighbor matching between a subject in the treatment group and in the comparison group, where each pair had a similar propensity to be in the treatment group.

In order to ensure that the matching was of high quality, various two-sample t-tests among observational characteristics and between the two groups were performed. Additionally, side-by-side histograms were visualized to confirm balance between the two groups. These results validated a high quality of matching and overall balance. Finally, the linear regression model was run to evaluate the influence of an incident involving a cat on incident notional cost. The summary statistics of the linear regression analysis revealed that there is a significant difference in the incident notional cost between the comparison and treatment groups.

The analysis performed in this report supported the hypothesis that a causal relationship exists between an incident being cat related and the cost of incident services. Additionally, these findings indicate that cat related animal services result in greater costs compared to non-cat related incidents. The presence of cats in the outdoors poses a threat to biodiversity loss and species endangerment. This analysis provides insights into the costs of cats for the LFB, as both their incident costs and their overall costs are significantly higher than other species (Figure 4). This connection between the influence of cats and the cost of cats should incentivize policy makers to impose stricter regulation on outdoor cats, as they are imposing greater costs on both the city and the livelihood of wildlife.

a. Weaknesses

Many of the predictor variables exhibited skew, which might've influenced the validity of the regression analyses as this violates key assumptions. Therefore, performing transformations on these variables using methods such as BoxCox, might improve the strength of the subsequent models in order to produce a more powerful model. Additionally, this model failed to consider influential points, which also affect the regression surface.

b. Next Steps

Next steps would include applying further transformations on skewed variables and generally improving the regression models. Additionally, we are interested in applying these statistical methods in regions where invasive species regulation has been enacted in order to investigate the success of these policies. Instead of applying propensity score matching, we plan to infer causality through the use of regression discontinuity design (RDD). RDD would be used to evaluate the effect of the invasive species regulation on cost while accounting for various observational characteristics. It is valuable to explore the success of regulation that this report proposes through the use of statistical methods that this report uses.

c. Discussion

Social reform, such as climate policy, is often primarily motivated by financial impacts in market based economies, such as the United Kingdom (Urry, 2015). According to experts, invasive species cost the United Kingdom almost €20 billion per year (IBC, 2014). The analysis in this report sheds light on the costs of the invasive species, cats, on London Fire Brigade Rescues, which are primarily paid for by the city and its taxpayers. Our results indicate that cats cause greater animal incident costs than other species. Therefore, imposing additional regulation that manages the prevalence of city cats should decrease their costs on the

city and their influence on biodiversity loss. The fate of invasive species that were introduced by humans is also in the hands of humans. Effective regulation has the potential to mitigate the extinction of many species by 2050 and aid in addressing the climate crisis.

Bibliography

- Allaire, J.J., et. el. *References: Introduction to R Markdown*. RStudio. <https://rmarkdown.rstudio.com/docs/>. (Last Accessed: January 15, 2021)
- Cahill, A. E., Aiello-Lammens, M. E., Fisher-Reid, M. C., Hua, X., Karanewsky, C. J., Yeong Ryu, H., ... & Wiens, J. J. (2013). *How does climate change cause extinction?*. *Proceedings of the Royal Society B: Biological Sciences*, 280(1750), 20121890.
- Caliendo, M., & Kopeinig, S. (2008). *Some practical guidance for the implementation of propensity score matching*. *Journal of economic surveys*, 22(1), 31-72.
- Center for Biological Diversity. “*The Extinction Crisis*” Web Accessed March 20, 2015.
- Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how*. Springer Science & Business Media.
- Fox, J. and Weisberg, S. (2018) *An R Companion to Applied Regression, Third Edition*. Sage.
- Gelman, A. & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.
- Grolemund, G. (2014, July 16) *Introduction to R Markdown*. RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: January 15, 2021)
- Insurance Bureau of Canada (IBC). (March, 2014). *Reducing the Fiscal and Economic Impact of Disasters*. Retrieved from <http://www.ibc.ca/qc/resources/studies/reducing-the-fiscal-and-economic-impact-of-disasters>
- Kampfner C. (2021). *Animal rescues by London fire brigade rise 20% in pandemic year*. The Guardian. Retrieved from <https://www.theguardian.com/world/2021/jan/08/animal-rescues-london-fire-brigade-rise-2020-pandemic-year>
- London Fire Brigade (LFB). 2009. *Animal rescue incidents attended by LFB*. London Datastore. Retrieved from <https://data.london.gov.uk/dataset/animal-rescue-incidents-attended-by-lfb>
- Loss, S. R., Will, T., & Marra, P. P. (2013). *The impact of free-ranging domestic cats on wildlife of the United States*. *Nature communications*, 4(1), 1-8.
- Lowe, S., Browne, M., Boudjelas, S., & De Poorter, M. (2000). *100 of the world's worst invasive alien species: a selection from the global invasive species database* (Vol. 12). Auckland: Invasive Species Specialist Group.
- Loyd, K. A. T., & DeVore, J. L. (2010). *An evaluation of feral cat management options using a decision analysis network*. *Ecology and Society*, 15(4).
- Mock, T. (2021). *TidyTuesday: Animal Rescues*. Retrieved from <https://github.com/rfordatascience/tidyuesday/blob/master/data/2021/2021-06-29/readme.md>
- Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2002). *An introduction to logistic regression analysis and reporting*. *The journal of educational research*, 96(1), 3-14.
- Palomares, F., Rodríguez, A., Revilla, E., LÓPEZ-BAO, J. V., & Calzada, J. (2011). *Assessment of the conservation efforts to prevent extinction of the Iberian lynx*. *Conservation Biology*, 25(1), 4-8.
- Seber, G. A., & Lee, A. J. (2012). *Linear regression analysis* (Vol. 329). John Wiley & Sons.

Thomas, C. D., Cameron, A., Green, R. E., Bakkenes, M., Beaumont, L. J., Collingham, Y. C., ... & Williams, S. E. (2004). *Extinction risk from climate change*. *Nature*, 427(6970), 145-148.

Urry, J. (2015). *Climate change and society*. In *Why the social sciences matter* (pp. 45-59). Palgrave Macmillan, London.

Appendix

A1: Ethics Statement

Maintaining proper ethics in statistical analyses is as important as any result. Reinforcing objectivity, transparency, and following the scientific method was a key aspect of this analysis. One such ethical consideration that was made within this report, was maintaining the anonymity of any subjects involved in the incidents, ranging from animal owners to employees of the LFB. Evidently, the final linear regression analysis of this report did conclude that there was a difference between the treatment and comparison group in costs. Consequently, this conclusion supported the hypothesis of this report. Nevertheless, it was crucial that there was no manipulation of the data or analyses to prove a point. Finally, this research could not have been performed without the literature and progress of experts in various academic fields such as climate science, economics, and statistics. This report assigned full credit to all experts whose ideas were referenced and used to form the argument that this report was based on.

A2: Materials

```
## Rows: 3,852
## Columns: 11
## $ IncidentNotionalCost <dbl> 999, 3480, 2980, 2340, 2340, 2340, 2768, 2664, 22~
## $ AnimalGroupParent    <fct> Unknown - Animal rescue from water - Farm animal,~
## $ Latitude             <dbl> 51.47184, 51.61431, 51.31007, 51.60176, 51.60147,~
## $ PostcodeDistrict     <fct> M, E, T, E, E, E, D, U, T, D, U, E, D, K, K, R, R~
## $ PumpHoursTotal       <dbl> 3, 12, 10, 9, 9, 9, 8, 8, 7, 7, 7, 7, 6, 5, 5, 6,~
## $ CalYear              <int> 2019, 2013, 2015, 2011, 2011, 2011, 2011, 2020, 2018, 2~
## $ treatment            <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0~
## $ predicted_propensity <dbl> 9.482496e-06, 2.288192e-02, 2.934369e-02, 5.47686~
## $ match.ind            <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 13, 12, 15, 14, ~
## $ cnts                 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0~
## $ pairs                <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 1, 1,~
```


Section 2: Supplementary Plots

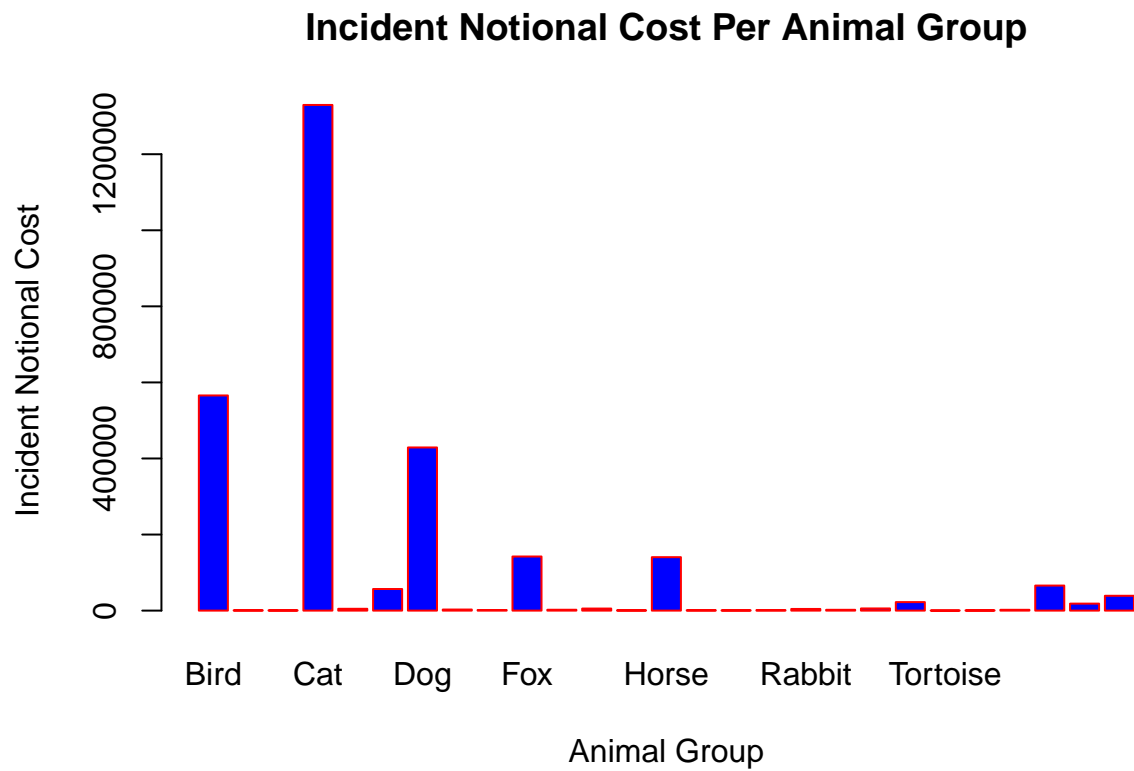


Figure 6. The bar plot corresponds to the total cost of incidents per animal group. These values were obtained by grouping each incident by animal group and calculating the sum of the costs to return one cost per animal group. Based on the plot, cats exhibit an exceedingly high total cost compared to the remaining animal groups. This is worth considering, as the individual incident costs for cats are below average for all animal groups, yet the total cost is so high.

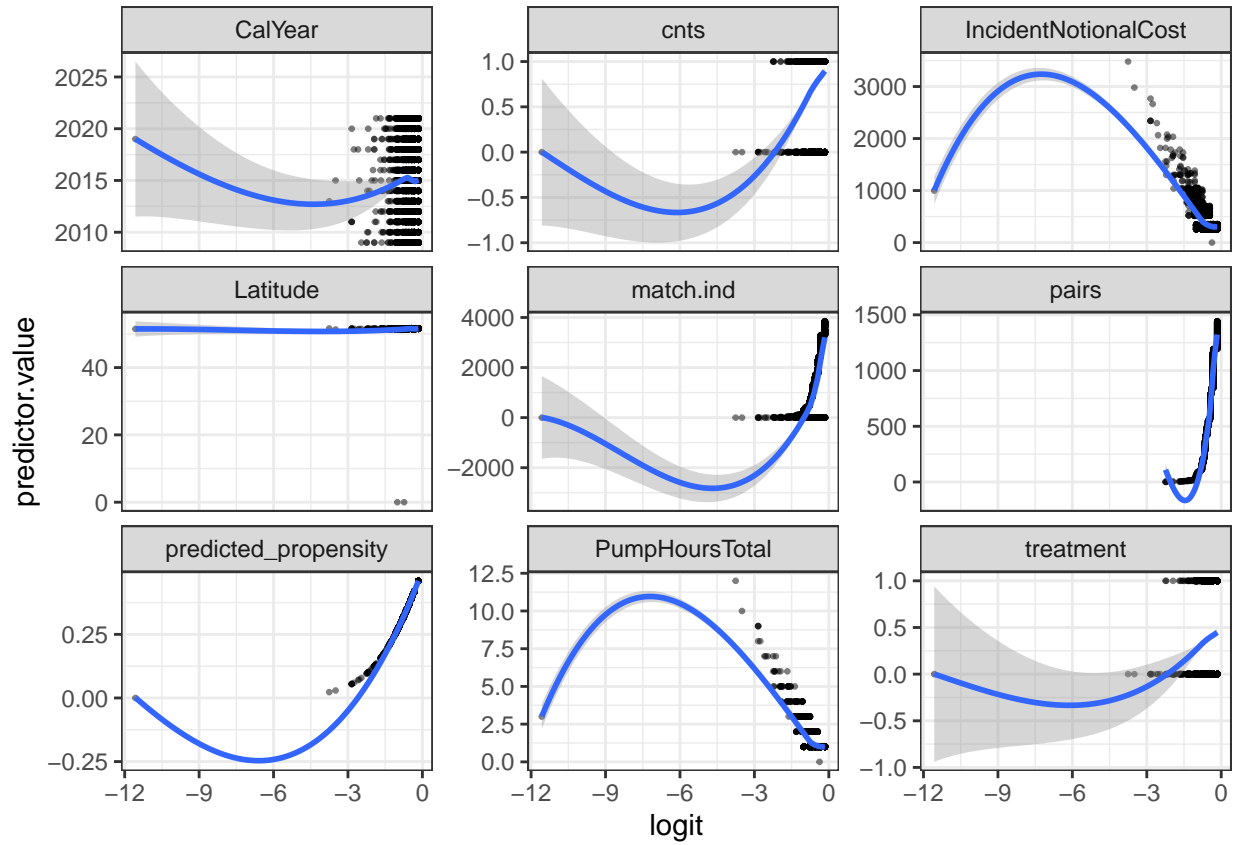
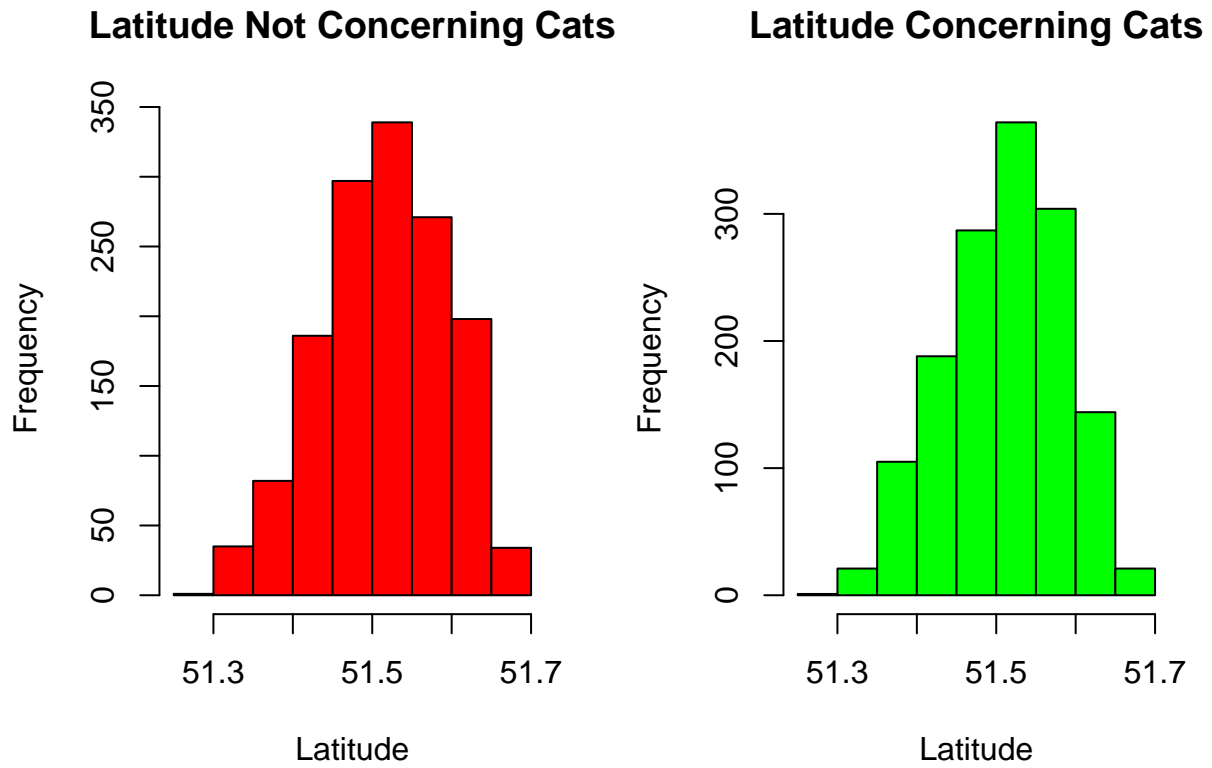


Figure 7. The smoothed scatter plots of the possible predictor variables indicate that none of the variables are linearly associated with the the outcome of an observation involving a cat in the logit scale. This lack of linearity implies a potential requirement for transforming the relevant variables. Additionally, the model might require additional analyses such as higher power terms.



Figures 8 & 9. The left, or red, histogram visualizes the distribution of latitude for animal incidents not involving cats. The right, or green, histogram demonstrates the spread of latitude for animal incidents involving cats, otherwise known as the treatment group. This side-by-side histograms demonstrates a balance of the observable trait, latitude, between the treatment and comparison groups since the spreads are nearly identical. This balance implies a high quality of matching.