1

The Data Sets

1.1 Introduction

This handbook is designed to provide an accessible introduction to statistical modeling techniques appropriate for data that are non-Gaussian (not normally distributed), do not have observations independent of each other, or may not be linearly related to selected predictors. The discussion relies heavily on data examples and includes thorough explorations of data sets, model construction and evaluation, detailed interpretations of model results, and model-based predictions. We intend to provide readers with a sufficiently thorough and understandable analysis process such that the techniques covered in this text can be readily applied to any similar data situation. However, it is important to understand that we use specific data sets with the various models strictly for demonstrative purposes. The outcomes we present are not to be assumed as definitive representations of information contained within the data sets.

Throughout the text, we will use four data sets (each described in this chapter) to exhibit the analytical methods including exploration of the data, building appropriate models (Chapter 2), evaluating the appropriateness of the models, output interpretation, and predictions made by the models. The purpose of using the same data sets throughout is to show that multiple methods can be applied to similar or identical variables of interest, possibly resulting in different conclusions. Consistent use of the same data sets should maintain data familiarity. After reading this first chapter, the intention of every data analysis throughout the remainder of the text should be understood. The modeling methods that are applied to the data sets are models for responses with constant variance (Chapter 3), responses with nonconstant variance (Chapter 4), discrete categorical responses (Chapter 5), models for count responses (Chapter 6), responses that are time-dependent (time-to-event data in Chapter 7, and outcomes collected over time in Chapter 8), and models for which variables that cannot be measured directly but are represented by variables that are measurable (Chapter 9). The last chapter, Chapter 10, is a guide to matching data sets to model types.

The following are brief introductions to each data set. These introductions are followed by descriptions of the data exploration methods that provide the details of the data sets needed to match them to the various models specifically designed for non-Gaussian and correlated data. Throughout the handbook, including the exploratory data analysis in this chapter, we use the functions and procedures, respectively, in the R (R Core Team, 2016) language and environment, and the SAS software, ©2016, SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA. All R and SAS commands used to produce the output discussed in this handbook are available.

1.1.1 The School Survey on Crime and Safety

The National Center for Education Statistics used the School Survey on Crime and Safety to record data from US public schools, addressing issues regarding safety in and around American public schools. Data for the 2007–2008 wave were collected from a stratified sample of 3 484 regular public schools. Variables of interest cover topics including school policies and facilities such as the presence of a school uniform policy and the use of metal detectors; school training and services such as teacher training on discipline policies and availability of counseling for students; and other variables such as the crime level of the areas surrounding school locations. We are interested in using these data to answer questions about school culture issues such as bullying and suspensions due to insubordination. The data can be downloaded from https://nces.ed.gov/surveys/ssocs/data_products.asp.

1.1.2 The Framingham Heart Study

The Framingham Heart Study was initiated to study common risk factors associated with cardiovascular disease, and to follow this disease's development over a long period of time for a large sample of participants. Data were collected from an initial cohort of 5 209 men and women between 30 and 62 years of age as of the time of a baseline physical examination. Data were also collected from two follow-up examinations performed two years and four years after the initial baseline measures. Variables of interest include patient demographic information such as age and sex; patient behaviors such as the number of cigarettes used per day; whether the patient uses blood pressure medication; and patient physiological measures such as systolic blood pressure, presence of diabetes, and body mass index. We shall analyze these data to answer questions about indicators of heart disease such as evidence of hypertension. The data can be downloaded from https://biolincc.nhlbi.nih.gov/teaching/.

1.1.3 Fire-Climate Interactions in the American West

Fire-Climate Interactions in the American West data since 1130 were obtained from the World Data Center for Paleoclimatology, Boulder, Colorado, and the National Oceanic and Atmospheric Administration (NOAA) Paleoclimatology Program (Trouet et al., 2010). These data were collected to assess whether climate is considered the main driver of wild fires in the American West. The data are on core samples from a variety of trees (Pipo, Psme, and Cade) from specific sites within each of the regions covering California, southern Oregon, and western Nevada. The regions are the Pacific Northwest (PSW), Northern California (NC), Interior West (IW), and the Southwest (SW). The time span is from 1130 through 2004. The core samples were examined to identify tree rings with fire scarring as an annual presence or absence of wild fires. We will use these data to predict indicators of fire scarring across regions and years. The data can be downloaded from www.ncdc.noaa.gov/paleo/study/10548.

1.1.4 English Wikipedia Clickstream Data

The clickstream data from the desktop version of the English Wikipedia were extracted from the logs of internet servers. The data are sequences of user-selected web addresses and links. Wikipedia (Wulczyn and Taraborelli, 2015) makes clickstream data available from its request logs, and we use the February, 2015 data. The data set includes only requests for articles in the main namespace. Pairings of the referring and requested sites with fewer than ten observations were removed from the data set by Wikipedia analysts.

The February, 2015 English Wikipedia Clickstream data set includes requests for redlinks (failed links), sorts out redirects, and has a field indicating whether the referrer and requested site pairings represent a link, a redlink, or a search. We intend to use these data to investigate the frequencies of pairings and the factors relating to redlinks. For examples of working with the February, 2015 release of the data, see this blog post: http://figshare.com/articles/Wikipedia_Clickstream/1305770. The data can be downloaded from http://ewulczyn.github.io/Wikipedia_Clickstream_Getting_Started/.

1.2 Exploratory Data Analysis

Data are used to drive, support, or provide understanding of a wide variety of human activities. We often use existing data to suggest patterns that predict human behavior such as spending habits, health care access, voting outcomes, among many others. Available data are used to allocate expensive resources, or make decisions that may affect the quality of human lives, including survival. The costs of making erroneous inferences can be enormous. Understanding data collection methods, contents, and quality is a prerequisite to utilizing the data set for analysis. The first step in understanding the information contained within a data set is a thorough exploration of the set's variables' structures and contents. This exploratory data analysis (EDA) differs from data management, which is concerned more with data collection, organization, and quality attributes including accuracy, consistency, and completeness.

EDA, in the context of model-building, provides a guide as to what model types may be appropriate for the data set under consideration. A few common exploratory analyses are distribution investigation, frequencies of variable levels, variable correlations, and data summary statistics. Distribution investigation, when applied to prospective dependent or response variables, suggests whether they may satisfy the independence and distribution assumptions of specific models. Frequency analysis gives the number of levels within variables and by variables. These frequencies often suggest the use of, for example, indicator variables. Within- and across-variable correlations can indicate autocorrelation and possible issues with multi-collinearity. Data summaries offer measures of central tendency, ranges, and quantiles. These attributes help show if a chosen model's predictions are commensurate with the observed data.

EDA assists us in choosing models appropriate for a specific data set. EDA for model selection includes response-to-predictor relationships, predictor-to-predictor associations, response-by-predictor clustering, to name a few characterizations. A critical aspect of these characterizations is which distributions the data set variables may follow. The response variable distributions are particularly crucial to model type identification and selection.

The following is a review of common EDA methods and tools.

1.3 Gauss-Markov Assumptions

In statistics, the Gauss-Markov Theorem, named after Carl Friedrich Gauss and Andrey Markov, states that in a model with linearly related coefficients for which the errors have an expected value of zero, are uncorrelated, and have constant variance; the ordinary least squares (OLS) method produces the best linear unbiased estimators (BLUE) of the coefficients. Here "best" means the estimators have the smallest variance as compared to any other unbiased and linear estimators. The errors do not need to be normal, nor do they need to be independent and identically distributed. They need only be uncorrelated with mean zero and homoscedastic with finite variance. The requirement that the estimator be unbiased is mandatory as biased estimators may exist with smaller variance. Biased estimators can lead to unrealistic and unusable model outcomes.

In general we construct linear regression models under the expectation that the Gauss-Markov (G-M) assumptions hold, or that remedial measures may be taken to transform the data to approximate the G-M assumptions. EDA is a tool by which we may determine the compliance of such transformations to accommodate the G-M assumptions.

The following sections describe common EDA techniques that we shall reference throughout this handbook.

1.4 Data Summaries and Tables

Data summaries include descriptive statistics such as the mean, median, mode, range, minima, and maxima. These statistics generally are measures of central tendency and dispersion, variance, or spread. Continuous data may also be summarized using quartiles or percentiles. Partitions of this sort indicate how much grouping may exist in continuous data, and whether the two ends have a paucity or abundance of observations which indicate tail thickness. The values of the medians and means may suggest a variable has a symmetric distribution when the mean and median are equivalent, or skewed otherwise. Data scales are apparent from the minima and maxima.

Further summary of discrete and categorical variables may be made by generating two-way, three-way, or multi-way tables. For example, we may need to understand not just the counts of the levels within a categorical variable, but also the counts of various combinations of the levels of two or more categorical variables. Tables of categorical variable levels can contain not just the counts of levels, but also the percentages, fractions, or proportions based on the counts. As we shall see in Chapter 5, these tables may be used for assessing model fit.

1.5 Graphical Representations

The main graphical methods for exploring data are plots of distributions, response-to-predictor relationships, and predictor-to-predictor associations. The plots representing distributions include histograms, quantile-quantile (Q-Q) plots, and box-whisker plots. These plots suggest shape, including symmetry, modality, locations of central tendency, the amount of spread, and possible outliers. The most common plot for depicting associations between pairs of variables is the scatter plot. Machine learning such as ensemble learning utilizes clustering visualizations to depict variable grouping patterns. Ensemble learning and many other plot types for representing data behaviors are beyond the scope of this handbook.

1.5.1 Histograms

Histograms allow the examination of the distributional characteristics of a numeric variable using a specially constructed bar chart. Typically, a single variable is divided into groups, often called bins, the size of which defines the width of the bar. There are a number of ways these bins' widths can be determined, and we leave it to the reader to investigate the types of algorithms used by the software package being used. Once the bin width is set, the number or proportion of observations that fall within each bin is used to construct the height of the corresponding bar.

Each bar's height, as determined by the observation count, is divided by the total number of observations. The bar height now represents the fraction of the total number of observations centered on each bar within the width of the bar. How the adjacent bar heights are distributed will usually show symmetry or skewness, either to the left or to the right. The height of the left-most and right-most bars may suggest unusual tail thickness. A curve is often superimposed over the bars that represents a best fit probability density function.

The histograms suggest distribution characteristics, but when used in conjunction with descriptive statistics, other EDA plots, and distribution fit assessment statistics, they give information needed to choose which model will best fit the data.

1.5.2 Q-Q Plots

A Q-Q plot is a graph of one set of quantiles against another set. A quantile for any distribution; whether a normal, Poisson, or no apparent named distribution; is an element of equally-spaced ranks resulting from ordering the data from lowest to highest, followed by summing these ranks and dividing by the sample size. Often the data quantiles (the vertical axis, or ordinate) are plotted against the quantiles of a normal distribution (the horizontal axis, or abscissa). Deviations from, say, the normal quantile line suggest a non-Gaussian distribution.

1.5.3 Box-Whisker Plots

The box-whisker plot is a useful tool to partition a continuous or ordinal variable (e.g., a response for a model) into groups defined by some other discrete, prediction variables in the data set. The plot identifies, by group, asymmetries in the response, relative positions of the response quartiles, and possible extreme values. Each group's box-whisker plot is composed of five parts: (1) the box, (2) the horizontal median line inside the box, (3) a mean marker inside the box (though not a standard practice), (4) upper and lower whiskers extending from the box, and (5) indicators of observations above the upper whisker or below the lower whisker.

The five box-whisker plot parts are described as follows:

1. The vertical axis of the plot ranges from the smallest to largest values of the continuous or ordinal data variable. The horizontal axis ranks (generally, defined by the user) the order of the groups represented by the box-whisker plots. The box contains the portion of the range in which 50% of the data lie within a given group. The lower box boundary is the $Q_1 = 25$th percentile, and the upper boundary is the $Q_3 = 75$th percentile. The difference gives the range in which 50% of the data lie, and is known as the inter-quartile range (IQR). The width of the box sometimes is used as a conceptual measure of the sample size for each group.
2. Within each box is a horizontal line that represents the median value location of the data; viz., the 50th percentile location. Often this line connects notches with end points on the left and right sides of the box. The notches represent an approximate 95% confidence interval about the median value. The median value indicates that approximately half the data values lie at or below the median, and approximately half the data values lie at or above the median.
3. The mean marker is not always used, but when it is, it represents the location of the mean relative to the box. If the marker seems significantly shifted from the median line, an asymmetric distribution of the data is likely.
4. The upper and lower whiskers represent a bound in which approximately 90% of the data lie. The whisker lengths are found as $Q_1 - 1.5 \times IQR$ and $Q_3 + 1.5 \times IQR$, where IQR is as described in Part (1). Differing lengths of the upper and lower whiskers suggest nonsymmetric distributions.
5. Finally, the locations of values above and below the whiskers identify possible extreme values. Caution is advised before unequivocally designating these extreme values as outliers, as many probability distributions are skewed, and will allow their inclusion. The skewness results from the response possibly being from a nonsymmetric probability distribution. For example, count data often follow nonsymmetric distributions (e.g., Poisson), and the markers are not usually outliers.

1.5.4 Scatter Plots

Scatter plots represent the paired relationship between two variables. Three variables may be graphed in a single scatter plot, but they can be challenging to view with the possible exception of geographical plots. Geographical plots may show, e.g., a map of the United States with bars in each state representing a quantity such as health care costs. However, we focus on the relationships between pairs of variables as depicted in x-y plots; i.e., one variable is plotted on the horizontal axis and the other is plotted on the vertical axis.

Information depicted in two-way x-y scatter plots includes how one variable responds to changes in another, whether one variable has differing levels of variability at various levels or locations of another variable, and whether the levels of one variable tend to group on specific levels of another. Each observation is plotted as an x-axis and y-axis pair. The entirety of the observations plotted as x-y pairings gives a sense of the shape of the paired observations. The functional shape of the pairing may be depicted by a smoother such as the loess or spline smoothers. The shape is also dictated by the spread around a functional curve such as a straight line. The shape may be nonlinear, such as a quadratic function. The data points may appear equally spread around the function which suggests homogeneous variance.

When a variable is dependent on time, the variable-versus-time scatter plot is known as a time-series plot or a time plot. A time-series plot may show if the variable has constant variance through time, whether there is a trend such as a steady increase, or a nonlinear change through time such as a sinusoid. There is a class of statistical models designed specifically to analyze time-series data, but we do not consider it in this text.

Scatter plots, then, are a descriptive form of bivariate analysis. They suggest if a transformation of one variable results in a linear association rather than a curvilinear relationship, and whether a transformation converts nonhomogeneous variance to near homogeneous variance.

1.6 Pairwise Correlation

While the scatter plots give graphical representations between pairs of variables, the independence between these same pairs may be numerically evaluated using pairwise correlation. Pairwise Pearson correlation is a numeric measure of the linear association between two variables. Multiple variable pairs may be combined into a matrix for convenience. The correlation values are independent of scale (as opposed to covariance). This means that a variable with a large range can be correlated to a variable with a much smaller range, preserving the integrity of the correlation, even when the two ranges are measured with different units. It is critical to note that pairwise correlation has meaning only if the variable pair have a linear relationship.

Pairwise Pearson correlations range from $-1 \leq \hat{\rho} \leq 1$, where $\hat{\rho}$ is the estimated value of the correlation. (Often the correlation coefficient is denoted by r.) The closer $\hat{\rho}$ is to the extremes, the stronger is the linear correlation of a pair of variables. However large the absolute value of the correlation coefficient is, it may lack statistical significance due to such conditions as sample size. Therefore it is useful to also generate a significance level statistic such as a p-value.

Many data analysts suggest that only relatively large values of $\hat{\rho}$ should be used in, e.g., a linear regression model, when the correlation is between the model response variable and a candidate predictor variable. However, this reasoning is fallacious for three reasons: (1) the linearity of the correlation may be in question, (2) the pairwise behavior may change in the presence of variability due to multiple predictors, and (3) the intent behind an effects model and a prediction model are not always the same. Hence, we should always test for linearity in response-predictor relationships, we should never conclude a predictor has no influence on a response until it is tested in the multi-predictor environment if more than one predictor is used, and we must remember that effects models essentially identify the predictors that minimize model unexplained-outcome variation whereas predictive models may give more robust predictions if so-called pairwise noncorrelated response-with-predictors are included.

1.7 Machine Learning Pattern Recognition

Particularly for large data sets (the so-called big data), the methods of pattern recognition may prove beneficial. Pattern recognition methods are used in the discipline known as data mining, and include such techniques as cluster analysis, random forests, lift charts, regression trees, neural networks, nearest neighbors, and support vector machines, to name a few. Descriptions of these methods and techniques is beyond the scope of this book, but can be useful for identifying variable roles for modeling.

1.8 Exploring the Data Sets

We first examine the School Survey on Crime and Safety data, then explore the Framingham Heart Study data, followed by the Fire-Climate Interactions in the American West, and we finish with the English Wikipedia Clickstream data.

1.8.1 School Survey on Crime and Safety Data

We are interested in constructing models to make conclusions about general student behavioral problems, including bullying by students and suspensions for insubordination, using the 2007–2008 School Survey on Crime and Safety. Data were collected at the school level during the one-year period from 2007 to 2008, meaning that each variable represents an attribute of a school and not of any individual student. While the data set contains hundreds of variables, we have narrowed our focus to a few school characteristics.

- C0514: suspensions, the number of suspensions due to insubordination during the year.
- C0134: uniforms, an indicator of whether the school requires students to wear uniforms.
- C0116: metal detectors, an indicator of whether students must pass through metal detectors.
- C0188: tipline, an indicator of whether the school maintains a “hotline" or “tipline" for students to report problems.
- C0178: counseling, an indicator of availability of counseling or social work for students.
- C0562: crime, the crime level in the location of the school (low, moderate, or high).
- C0268: discipline training, an indicator of the availability of teacher training on discipline policies.
- C0276: behavioral training, an indicator of the availability of teacher training on positive behavioral interventions.
- C0508: insubordinates, the number of students involved in insubordination during the year.
- C0526: limited English, the percent of students with limited English language proficiency.
- C0532: below 15th, the percent of students who scored below the 15th percentile on standardized tests.
- C0376: bullying, how often student bullying occurs during the year in question (never, on occasion, monthly, weekly, or daily).

Continuous predictors of interest include the frequency of insubordinate students, the percentage of students with limited English language proficiency, and the percent of students below the 15th percentile on standardized tests. Noncontinuous predictors of interest include the level of crime in the area where the school is located (low, moderate, high), and indicators of whether students are required to wear uniforms, whether students pass through metal detectors, whether the school has a tipline to report problems, whether student counseling is available, whether teachers have training in discipline policies, and whether teachers have training in positive behavioral interventions.

Table 1.1 *School Survey on Crime and Safety descriptive statistics for continuous variables.*

Variable	Minimum	Median	Mean	Maximum	Variance
Number of suspensions	0	0	7.852	3 000	4 863.123
Number of insubordinate students	0	16	88.76	9 608	118 589.792
Percent with limited English	0.00	2.00	8.727	100	217.387
Percent below 15th percentile on tests	0.00	10.00	13.77	100	208.417

Table 1.1 shows basic descriptive statistics for the continuous variables of interest. Using this table we can see there are schools reporting 0 for each variable, and there are also schools reporting 100% for the two percentages. The number of suspensions and the number of insubordination events show clear evidence of skewness to the right, as the mean exceeds the median, and the maximum values, 3 000 and 9 608, respectively, are much greater than the median values. The median of 0 for the number of suspensions implies that at least half of the schools reported no suspensions during the year of interest.

The scatter plot matrix shown in Figure 1.1 gives a visual indication of possible relationships among the continuous variables. The histograms along the diagonal show evidence of skewness to the right in the percent of students with limited English language proficiency and also in the percent of students below the 15th percentile on standardized tests. Due to expected skewness to the right, both the number of suspensions and the number of insubordinate students were log-transformed (using the logarithm of suspensions+0.01 and the logarithm of insubordinates+0.01 to avoid undefined values from taking the logarithm of 0). The histograms from these two log-transformed variables show the smallest value to be the mode, which indicates that 0 suspensions and 0 insubordinates are the most common response for each.

The plot in the first row, second column of Figure 1.1 shows a reasonably strong relationship between the logarithm of suspensions and the logarithm of insubordinate students, as expected, and is supported by the relatively large value of 0.40 in the pairwise linear correlation. This scatter plot also shows an apparent diagonal "border" above which no observations are plotted. This is reflective of the fact that the number of suspensions does not exceed the number of insubordinate students. The remaining relationships appear relatively weak, and are affected by the large number of zeros for both number of suspensions and number of insubordinate students. In fact, the estimates of linear correlation between the percent of students with limited English language proficiency and both log-suspensions and log-insubordinates are so small as to not appear in the plot (0.006 and 0.005, respectively).

Table 1.2 shows the frequencies associated with each level of each categorical variable of interest in the data. Bullying has been recorded by schools as happening never, on occasion, monthly, weekly, and daily. The table shows most schools report bullying on occasion or monthly, but that daily bullying is more prevalent than no bullying at all. Most schools have no uniforms, no metal detectors for students to pass through, and no tipline to report issues,

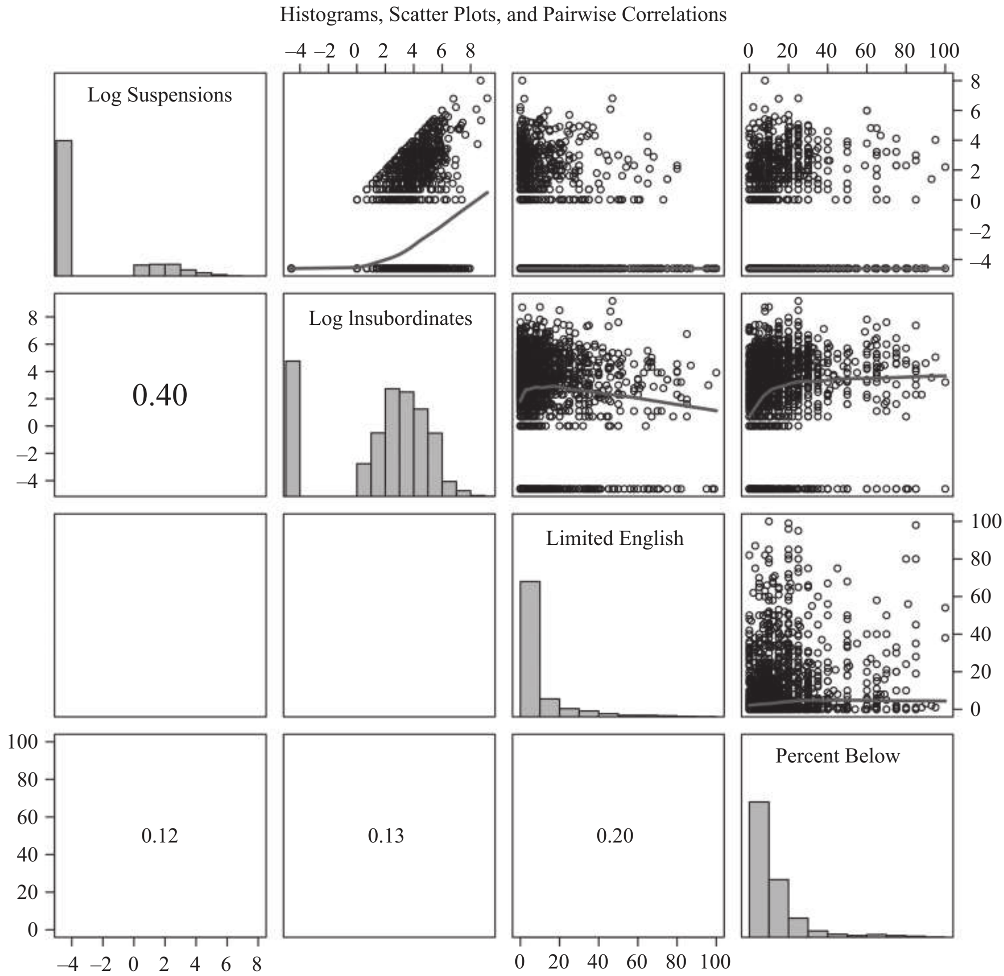

Figure 1.1 Scatter plot matrix of School Survey on Crime and Safety continuous variables, including histograms on the diagonal, pairwise Pearson correlations, and smooth loess curves.

but most schools have counseling available to students, are in locations of low crime, and have teachers trained in discipline policies and positive behavioral interventions.

In order to investigate the prevalence of predictor characteristics with school bullying, Table 1.3 shows cross-classification counts of schools that show combinations of bullying levels with the categorical predictors of interest. Counts can be used to describe patterns of variables of interest across bullying levels. For example, the "never" group has 30 schools with counseling and 10 without, a ratio of $30/10 = 3.00$, while the "on occasion" group increases to $1\,115/72 \approx 15.49$, and also ≈ 13.39, ≈ 21.64, and ≈ 23.00 for "monthly," "weekly," and "daily," respectively. The relative proportion of schools with counseling services available to students increases with frequency of student bullying; however, this change does not appear to follow a straight-line trend. Similarly, the proportion of schools in areas of moderate crime shows an increase across levels of bullying.