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,

https://doi.org/10.1017/9781316544778.002 Published online by Cambridge University Press

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

https://doi.org/10.1017/9781316544778.002 Published online by Cambridge University Press

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

https://doi.org/10.1017/9781316544778.002 Published online by Cambridge University Press

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,

https://doi.org/10.1017/9781316544778.002 Published online by Cambridge University Press

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.

https://doi.org/10.1017/9781316544778.002 Published online by Cambridge University Press

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	3000	4863.123
Number of insubordinate students	0	16	88.76	9608	118589.792
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, 3000 and 9608, 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,

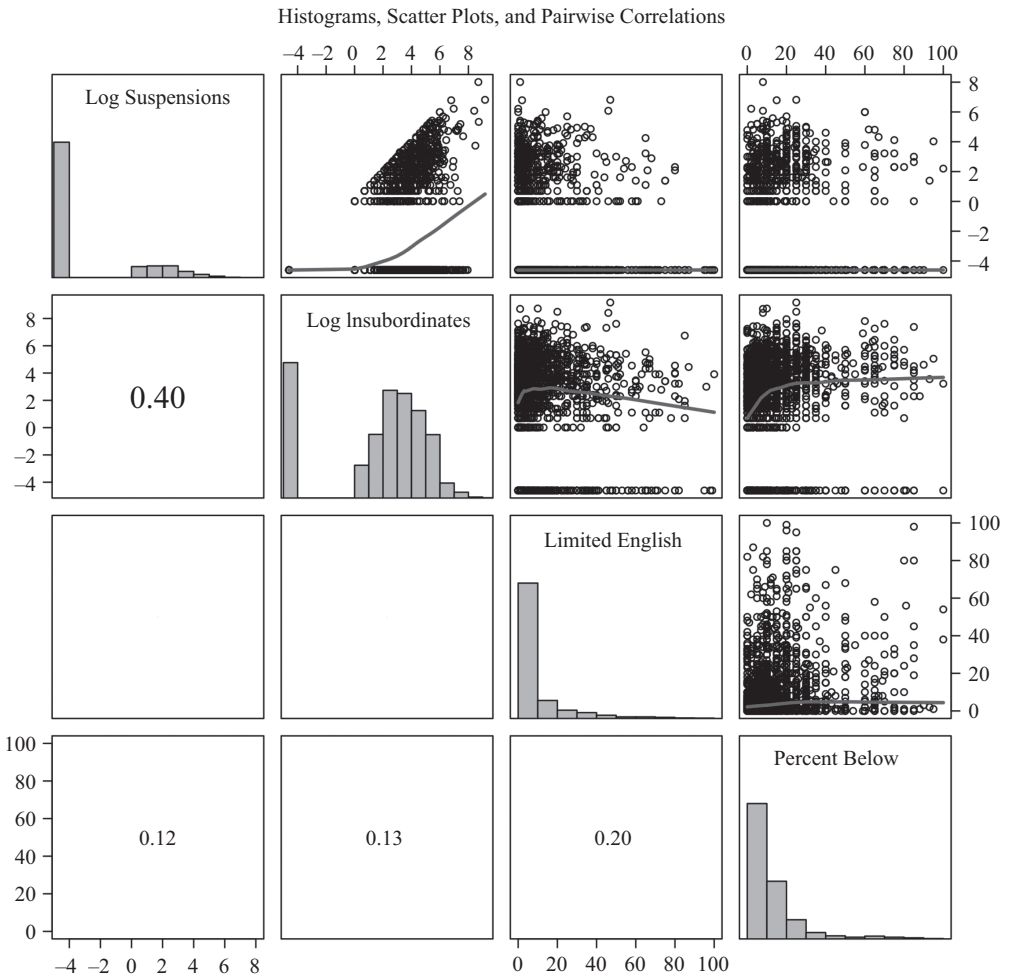

Figure 1.1 Scatter plot matrix of School Survey on Crime and Safety continuous variables, including histograms on the diagonal, pairwise Pearson correlations, and smooth loess curves.

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.

https://doi.org/10.1017/9781316544778.002 Published online by Cambridge University Press

Table 1.2 *School Survey on Crime and Safety descriptive statistics for categorical variables.*

Variable	Levels	Number	Percent
Bullying	Never	40	1.6%
	On occasion	1 187	46.4%
	Monthly	547	21.4%
	Weekly	498	19.5%
	Daily	288	11.3%
Uniforms	Yes	377	14.7%
	No	2 183	85.3%
Metal detectors	Yes	65	2.5%
	No	2 495	97.5%
Tipline	Yes	901	35.2%
	No	1 659	64.8%
Counseling	Yes	2 406	94.0%
	No	154	6.0%
Discipline training	Yes	1 792	70.0%
	No	768	30.0%
Behavioral training	Yes	1 969	76.9%
	No	591	23.1%
Crime	Low	1 922	75.1%
	Moderate	494	19.3%
	High	144	5.6%

Table 1.3 *School Survey on Crime and Safety contingency table of categorical predictor counts, by bullying level.*

		Bullying				
		Never	On occasion	Monthly	Weekly	Daily
Total		40	1 187	547	498	288
Uniforms	Yes	10	174	83	66	44
	No	30	1 013	464	432	244
Metal detectors	Yes	1	35	10	12	7
	No	39	1 152	537	486	281
Tipline	Yes	12	405	190	184	110
	No	28	782	357	314	178
Counseling	Yes	30	1 115	509	476	276
	No	10	72	38	22	12
Discipline training	Yes	33	845	361	359	194
	No	7	342	186	139	194
Behavioral training	Yes	30	937	397	395	210
	No	10	250	150	103	78
Crime	Low	31	938	416	350	187
	Moderate	6	189	111	117	71
	High	3	60	20	31	30

https://doi.org/10.1017/9781316544778.002 Published online by Cambridge University Press

Table 1.4 *School Survey on Crime and Safety descriptive statistics for number of suspensions, by categorical predictors.*

Variable	Levels	Suspensions average	Suspensions variance
Uniforms	Yes	8.777	1 407.280
	No	7.692	5 460.686
Metal detectors	Yes	19.185	4 006.028
	No	7.557	4 883.633
Tipline	Yes	12.646	1 2011.913
	No	5.248	966.259
Counseling	Yes	8.079	5 150.157
	No	4.312	369.615
Discipline training	Yes	9.317	6 758.223
	No	4.434	427.555
Behavioral training	Yes	8.381	6 052.134
	No	6.090	901.265
Crime	Low	6.009	5 217.558
	Moderate	13.065	4 072.588
	High	14.569	2 708.778

Table 1.5 *School Survey on Crime and Safety descriptive statistics for continuous variables, by bullying level.*

		Insubordinates	Limited English	Below 15th
Bullying	Never	18.35(2 535.82)	10.48(357.13)	16.88(402.63)
	On occasion	64.95(33 029.27)	8.32(201.03)	12.77(201.62)
	Monthly	98.11(99 122.63)	8.46(226.71)	12.97(159.04)
	Weekly	104.20(138 104.02)	8.74(220.84)	15.10(215.89)
	Daily	152.24(485 167.10)	10.64(240.38)	16.69(275.67)

Table 1.4 shows descriptive statistics of the number of school suspensions for each group of the categorical predictors. Generally, the mean numbers of suspensions tend to be noticeably higher for schools with each of metal detectors, tiplines, counseling services, training on discipline policies and in areas of moderate or high crime. The differences in mean suspensions for uniform policies and positive behavioral training do not appear to be as large.

Table 1.5 shows means and variances of continuous predictors across levels of school bullying. Average values of insubordinate students increase as expected across levels of bullying frequency, although the percentage of students with limited English language proficiency remains relatively stable. The percentage of students below the 15th percentile in standardized testing shows some increase for weekly and daily bullying as compared to

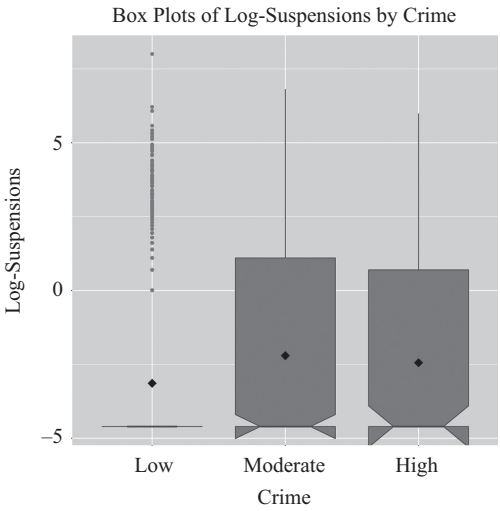

Figure 1.2 Box-whisker plots of log-suspensions for each crime level in the School Survey on Crime and Safety.

lesser levels of bullying. We can see that student bullying is not represented as a continuous variable, and therefore it would not be appropriate to model this outcome using normal linear regression techniques.

Figure 1.2 shows box-whisker plots of log-suspensions by area crime level. Schools in areas with low crime show the lowest typical numbers of suspensions, but also the largest number of possible outliers. The box-whisker plots of log-suspensions for schools in areas of moderate and high crime are similar to each other, but still show evidence of skewness to the right. Overall it appears the number of annual suspensions is highly skewed to the right and may have an excess of observations of 0, and therefore normal linear regression models would not be appropriate to model suspensions as an outcome.

1.8.2 Framingham Heart Study Data

We are interested in making conclusions about hypertension, including the prevalence of hypertension and the time until the onset of hypertension, using the data available from the Framingham Heart Study. The data comprise three waves of collection, each separated by two years. Some variables in the data set indicate measures specific to each wave, while others indicate presence or absence of a property over the entirety of the three-wave data set. For example, "PREVHYP" refers to each individual examination and may change over periods of observation, while "HYPERTEN" is an indicator of what was observed over the entire course of data collection. While there are numerous variables available in the public Framingham Heart Study data, we are concerned with only a few. Details of the processes involved with each physiological measure are available in the official documentation.

- RANDID: ID, a unique identification number for each participant.
- PERIOD: Period, examination cycle (1, 2, and 3).
- SEX: Sex, participant sex (1 for male, 2 for female)

https://doi.org/10.1017/9781316544778.002 Published online by Cambridge University Press

Table 1.6 *Framingham Heart Study descriptive statistics for continuous variables.*

Variable	Minimum	Median	Mean	Maximum	Variance
Time of hypertension	0	2429	3599	8766	11996207.990
Total cholesterol	107.0	238.0	241.2	696.0	2059.309
Age	32.00	54.00	54.79	81.00	91.325
Cigarettes	0.00	0.00	8.25	90.00	148.000

- AGE: Age, the age, in years, of each individual at examination.
- TOTCHOL: Cholesterol, serum total cholesterol (mg/dL) at examination.
- DIABETES: Diabetes, an indicator of diabetic at examination.
- CIGPDAY: Cigarettes, the number of cigarettes smoked each day.
- PREVHYP: Hypertension, an indicator of prevalence of hypertension at examination.
- HYPERTEN: Hypertension, an indicator of whether each participant showed evidence of hypertension during any exam.
- TIMEHYP: Time to Hypertension, the number of days from the baseline examination to the first indications of hypertensive, or the number of days until the final contact is made with the participant if indications of hypertension are never recorded.

Continuous variables of interest include the time to hypertension, total cholesterol, and age. Noncontinuous variables of interest include sex, cigarettes per day, diabetic, and hypertensive. Table 9.2 provides basic descriptive statistics about each continuous variable. We note that some patients have a time of hypertension of 0 days, indicating presence of hypertension at the baseline exam. Both the minimum and the median of cigarette use are 0, indicating that at least half of the patients in the study do not use cigarettes on a daily basis, and suggesting this variable is skewed to the right.

The scatter plot matrix shown in Figure 1.3 gives a visual indication of possible relationships among the continuous variables. Many of the scatter plots fail to show evidence of strong relationships. Based on the plot in the first row, fourth column, it appears the most extreme values of total cholesterol are observed with the smallest times until hypertension. The plot in the first row, third column shows the variation in total cholesterol to decrease with the number of cigarettes smoked per day, and the plot in the second row, third column shows the number of cigarettes smoked per day to peak slightly before age fifty, and decrease with age. The small values of correlation in the lower panel of the figure support the lack of strong associations.

Histograms in Figure 1.3 show total cholesterol and age to be relatively symmetric, while cigarettes smoked per day is skewed to the right and time until hypertensive is bimodal, with common values toward the low end and high end of times.

Table 1.7 shows the proportions of specific outcomes associated with each categorical variable. The proportions suggest that around three-quarters of all exams resulted in evidence of hypertension, while only around 5% of exams showed evidence of diabetes. However, we cannot extend these simple descriptive statistics to make statements about percentages of patients, as the data were collected over time and represent multiple observations for each participant. Multiple observations on an individual, when ignored,

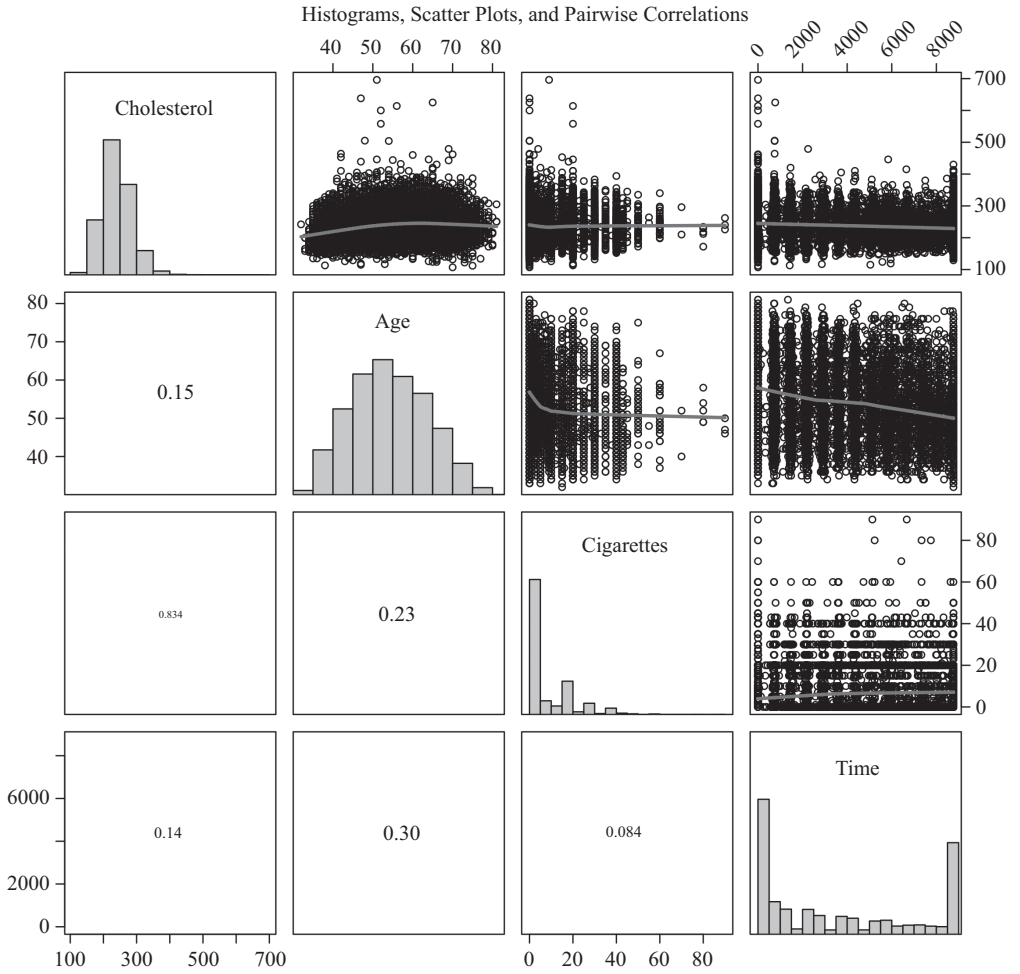

Figure 1.3 Scatter plot matrix of Framingham Heart Study continuous variables, including histograms on the diagonal, pairwise Pearson correlations, and smooth loess curves.

can bias the interpretations associated with descriptive statistics such as correlation and contingency table counts.

Table 1.8 shows cross-classified counts: the number of individuals with and without evidence of hypertension who showed diabetes and who were male or female. These splits in counts show us that there are more females than males with evidence of hypertension. Most of the individuals with diabetes show evidence of hypertension, although most of the individuals with hypertension are not diabetic.

Table 1.9 shows the descriptive statistics for the continuous variables of interest, split by hypertension. We see that individuals who show hypertension have much smaller values for the time of hypertension than individuals who do not show hypertension. This makes sense, as individuals who never show evidence of hypertension have time recorded as the final time of contact with the participant. Total cholesterol shows much greater

https://doi.org/10.1017/9781316544778.002 Published online by Cambridge University Press

Table 1.7 *Framingham Heart Study descriptive statistics for categorical variables.*

Variable	Levels	Number	Percent
Hypertension	Yes	8642	74.3%
	No	2985	25.7%
Sex	Female	6605	56.8%
	Male	5022	43.2%
Diabetes	Yes	530	4.6%
	No	11097	95.4%

Table 1.8 *Framingham Heart Study contingency table of categorical predictor counts by hypertension.*

		Hypertension	No hypertension
Total		8642	2985
Sex	Female	4956	1649
	Male	3686	1336
Diabetes	Yes	467	63
	No	8175	2922

Table 1.9 *Framingham Heart Study descriptive statistics continuous variables, by hypertension.*

Variable	Minimum	Median	Mean	Maximum	Variance
	Hypertension				
Time of hypertension	0	826	2116	8764	6360355
Total cholesterol	107.0	241.0	243.8	696.0	2087.197
Age	33.00	56.00	55.91	81.00	90.570
Cigarettes	0.00	0.00	7.674	90.00	146.881
	No hypertension				
Time of hypertension	45	8766	7891	8766	3543448
Total cholesterol	117.0	229.0	233.4	430.0	1894.903
Age	32.00	51.00	51.56	80.00	80.045
Cigarettes	0.00	3.00	9.924	80.00	149.561

variation for participants who show hypertension, as evidenced by the smaller minimum and larger maximum than individuals without hypertension. Individuals without hypertension generally show greater cigarette use, as evidenced by the larger mean and median than those with hypertension.

Figure 1.4 shows plots of hypertension ("1" represents "yes" and "0" represents "no") versus age and total cholesterol. Because the outcome of interest, hypertension, is binary, it is difficult to identify a pattern as with a typical scatter plot. Therefore the loess smoothed curve is superimposed to show the general increasing relationship between each predictor

https://doi.org/10.1017/9781316544778.002 Published online by Cambridge University Press

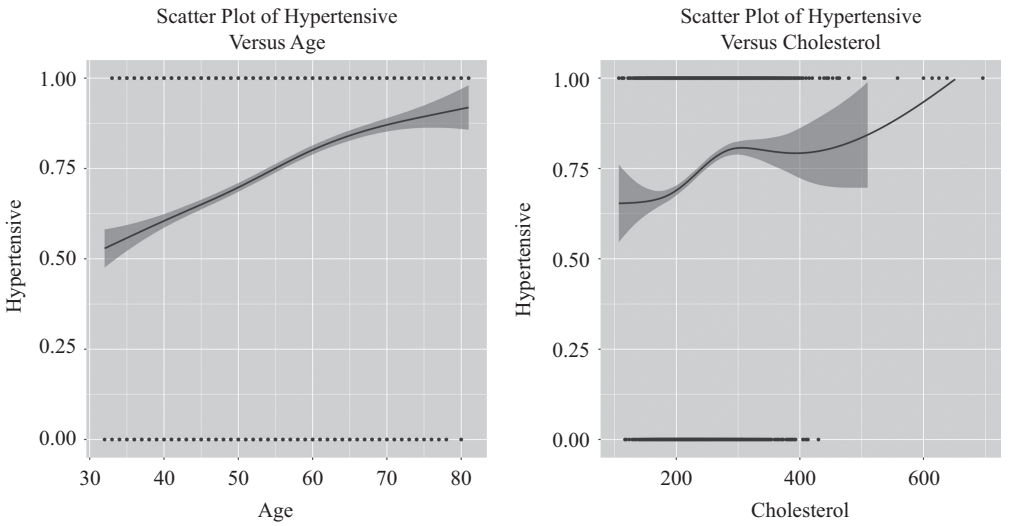

Figure 1.4 Framingham Heart Study plots of hypertension versus age (left panel) and total cholesterol (right panel), with smooth loess curves.

and the prevalence of hypertension. In this case the loess curve can be thought of as a smoothed estimate of the proportion of individuals showing evidence of hypertension for various values of the predictors. It appears the proportion of individuals with hypertension increases with both age and cholesterol, although the nature of the relationship does not appear to be linear for cholesterol. The exploratory statistics presented suggest the relationship between hypertension and predictors such as age and cholesterol should not be investigated using normal linear regression methods.

1.8.3 Fire-Climate Interactions in the American West Data

The models of Fire-Climate Interactions in the American West use the number of annual tree ring fire-scar markers for the years beginning 1130 through 2004. The regions and the sites in which the tress are located are used to construct the models. The tree ring fire-scar markers from individual trees that experienced fires since the year 1130 are used. These marker tree rings were collected from among the 350 sites in each of the four regions of the fire data, but we are concerned here only with the sites listed below (Trouet et al., 2010).

- year: the year indicated by sampled tree rings.
- region: four regions are used in this text. They are the Intermountain West (IW), Northern California (NC), the Pacific Northwest (PNW), and the Southwest (SW).
- site: within region are sample sites. They are:
  - IW: Ashenfelder, Cheesman Lake, Manitou, and Old Tree
  - NC: sites not used
  - PNW: Frosty, Nile Creek, South Deep, and Twenty Mile
  - SW: Blacks, Round Mountain, and Cerro / Hoya / Marchanita

https://doi.org/10.1017/9781316544778.002 Published online by Cambridge University Press

Table 1.10 *Fire-Climate continuous variables data summary.*

Variable	Minimum	Median	Mean	Maximum	Variance
Decade	1 130	1 604	1 610	2 000	54 825.246
Year	1 130	1 608	1 604	2 004	54 810.000
Rings/year	0	0	0.756	19	3.061
Rings/decade	0	28	30.55	87	614.120

Table 1.11 *Fire data site by region contingency table (the NC sites are not used).*

	IW	NC	PNW	SW	Total
Total	875	841	661	771	3 148
Ashenfelder	88		0	0	
Blacks	0		0	78	
Cerro Hoya Marchanita	0		0	78	
Cheesman Lake	88		0	0	
Frosty	0		67	0	
Manitou	88		0	0	
NileCreek	0		67	0	
OldTree	88		0	0	
Round Mountain	0		0	78	
Twenty Mile	0		67	0	

The fire-scarred tree ring samples were taken from the American West, stratified by region and further stratified by site. Models may pool over site or over region, or models may account for the stratifications through longitudinal analysis. This handbook accrues the yearly counts into decadal summaries such as means or counts prior to applying these data to various model types.

The fire data continuous variable summary is given in Table 1.10. The row "decade" summarizes the number of decades from between 1130 and 2004. The column labeled "region" gives the numbers of tree ring samples in each of the four given regions. "Year" is the number of years between 1130 and 2004. The counts of tree ring samples by year are given in "rings/year," and the number of tree ring samples aggregated by decade is in column "rings/decade." Table 1.11 has the counts of rings for site by region. The Northern California region's sites are not used, and hence the associated site counts for region NC are not given.

Figure 1.5 is a matrix plot of the decade, decadal counts, and the log-transformed decadal counts which allows us to examine these variables' histograms, pairwise scatter plots (with loess smoothers), and the pairwise linear correlations. The plot matrix diagonal gives each variable's histogram. The first row, first column histogram is of decade, which is of minor interest.

The second row, second column panel is the decadal counts in which we see that smaller counts dominate. Smaller counts also dominate in the third row, third column panel of the matrix which is the histogram of the log of the decadal counts. Note that for purposes of

https://doi.org/10.1017/9781316544778.002 Published online by Cambridge University Press

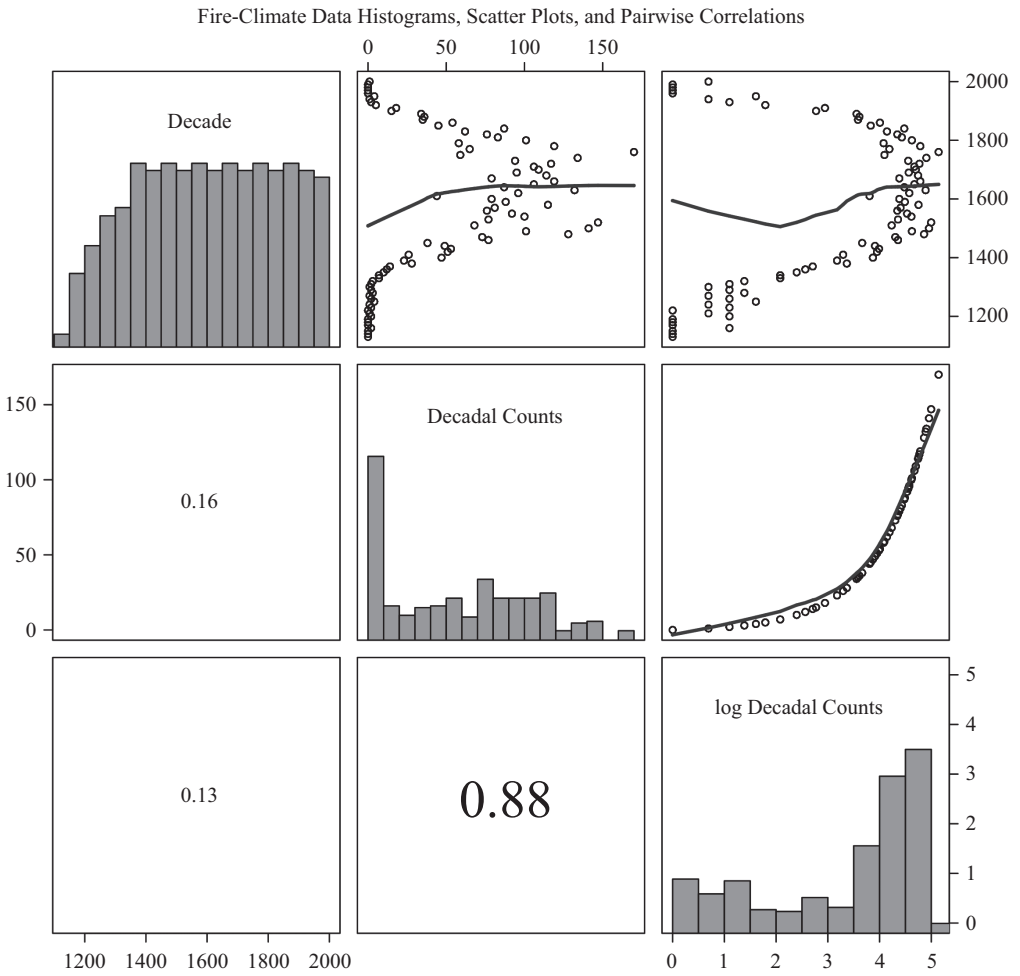

Figure 1.5 Matrix plot of the histograms, scatter plots, and linear correlation statistics of the three pairings of decade, decadal counts, and log-transformed decadal counts.

generating the log plot a value of one is added to each count as the raw counts include the value zero. This histogram shows a shift of dominance from the low counts of the untransformed decadal counts, to the higher values of the transformed data. Even though the transformation has shifted the mode of the distribution, it is clearly not a Gaussian distribution.

The linear correlation value for the pairing of decade and the decadal counts is the second row, second column panel of Figure 1.5. The small correlation, $r = 0.16$, suggests no linear relationship. The linear correlation value of decade and the log of the decadal counts is given in the third row, first column. The small correlation, $r = 0.13$, suggests no linear relationship.

The upper triangular matrix of cells of Figure 1.5 are scatter plots of the pairings. We are not interested in the pairing of decadal counts and log of decadal counts except, if there is interest in the appearance of the log transformation. We see that neither decadal count (first row, second column) nor log decadal count (first row, third column) is a linear function

https://doi.org/10.1017/9781316544778.002 Published online by Cambridge University Press

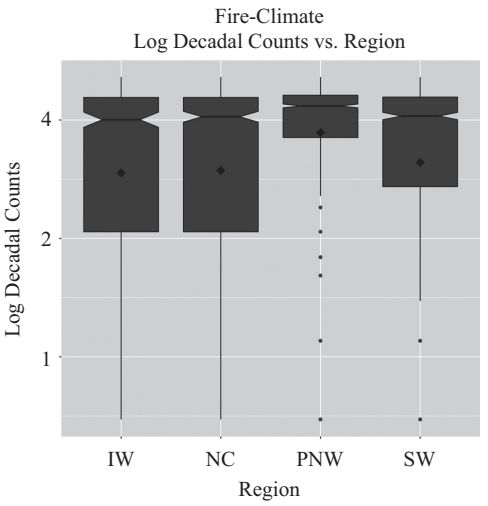

Figure 1.6 Box-whisker plot of the log of decadal count by region.

of decade. This nonlinear behavior invalidates the correlation statistics, as does the lack of normality in the histograms, and we must make an accounting of these nonlinearities when constructing models.

The fire data summary, Table 1.10, shows that region is a categorical variable with four levels, and for each level, the table gives the number of tree ring samples. The table shows that no region has zero counts, allowing each region to be included in models. We examine the log decadal counts by region using a box-whisker plot in Figure 1.6. Note the skewness in each of the plots, which is a reflection of the combined log counts skewed histogram we saw above. This suggests a Gaussian model is likely an inappropriate model type for these data.

1.8.4 English Wikipedia Clickstream Data

Wikipedia (Wulczyn and Taraborelli, 2015) makes clickstream data available from its request logs, and we use the February, 2015 English subset of these data. The link pairings of website referrals and site requests are examined by link type and previous site. The data are requests for articles in the main namespace of the desktop version of English Wikipedia. Pairings of the referring and requested sites with fewer than 10 observations were removed from the data set by Wikipedia analysts.

The data consist of six variables and we use the following three, whose definitions are from Wulczyn and Taraborelli (2015):

- n: the number of pairings of the referring and requested sites indicating the number of times web paths are used.
- prev_title: the mapping of a referring URL to any of the following values: Bing, empty, Google, Main_Page, other, Wikipedia, or Yahoo.
- type: indicates if the pairings of the referring and requested sites are:

https://doi.org/10.1017/9781316544778.002 Published online by Cambridge University Press

Table 1.12 *English Wikipedia Clickstream continuous and categorical variables descriptive statistics. The number of pairings is the continuous (counts) variable. Link type and previous title are the categorical variables.*

Variable	Minimum	Median	Mean	Maximum	Variance
Number of pairings	10	95.05	23	95 738	778 093.6

Variable	Levels	Number	Percent
Link type	Link	37 545	57.35%
	Other	27 602	42.16%
	Redlink	317	0.48%
Previous title	Bing	1 593	2.43%
	Empty	6 733	10.29%
	Google	8 078	12.34%
	Main_Page	509	0.78%
	Other	43 359	66.23%
	Wikipedia	3 962	6.05%
	Yahoo	1 231	1.88%

- link: the referrer and request are both articles and the referrer links to the request.
- redlink: the referrer is an article and links to the request, but the request is not in the production enwiki.page table.
- other: the referrer and request are both articles but the referrer does not link to the request.

A summary of the clickstream data is given in Table 1.12. The summary includes of the number of pairings of the referring and requested sites. The table gives the counts of the three levels of link type ("link type") and the counts of the seven previous titles' ("previous titles") levels. The table shows that no category has zero counts, allowing each category to be included in models.

The combination of the pairings of link types and previous titles is given in Table 1.13. This two-way table is important to determine if any combination of the category levels has zero or small numbers relative to the expected cell count. Small counts tend to dominate from expected cell counts, thus possibly biasing tests of significance. The table shows zeros and counts of 1 and 2 for several combinations of the levels, and hence, interacting these two categories as a predictor in a model is likely to fail.

The numbers of pairings will be partitioned in Chapter 5 for modeling as the probability of these pairings. In Chapter 6, the counts of pairings will be modeled.

Figure 1.7 shows the histograms of the counts of referring and requested pairings. The left-hand panel is the raw counts, and we see that smaller counts dominate even though it was necessary to exclude all counts above 1 000 to obtain a displayable plot. The right-hand panel of the figure is a histogram of the natural log of these same counts, and it is clear that the data are truncated on the left. This is verified by the minimum count in Table 1.12 at 10. Each plot has an overlay of a smoothed density curve.

The relationship between the categories in Table 1.12 and the log counts is shown in Figure 1.8. The left-hand panel of the figure is a box-whisker plot of the pairings types

https://doi.org/10.1017/9781316544778.002 Published online by Cambridge University Press

Table 1.13 *English Wikipedia Clickstream link type by previous title contingency table.*

Previous title		Link type			
		Link	Other	Redlink	Total
	Total	37 545	27 602	317	65 464
Bing	Count	1	1 592	0	1 593
	Percent	30%	1 013%	0.00%	
Empty	Count	2	6 731	0	6 733
	Percent	0.00%	10.28%	0.00%	
Google	Count	2	8 075	1	8 078
	Percent	0.00%	12.34%	< 0.01%	
Main_Page	Count	2	507	0	509
	Percent	0.00%	0.77%	0.00%	
Other	Count	37 537	5 506	316	43 359
	Percent	57.34%	8.41%	0.48%	
Wikipedia	Count	1	3 960	0	3 961
	Percent	0.00%	6.05%	0.00%	
Yahoo	Count	0	1 231	0	1 231
	Percent	0.00%	1.88%	0.00%	

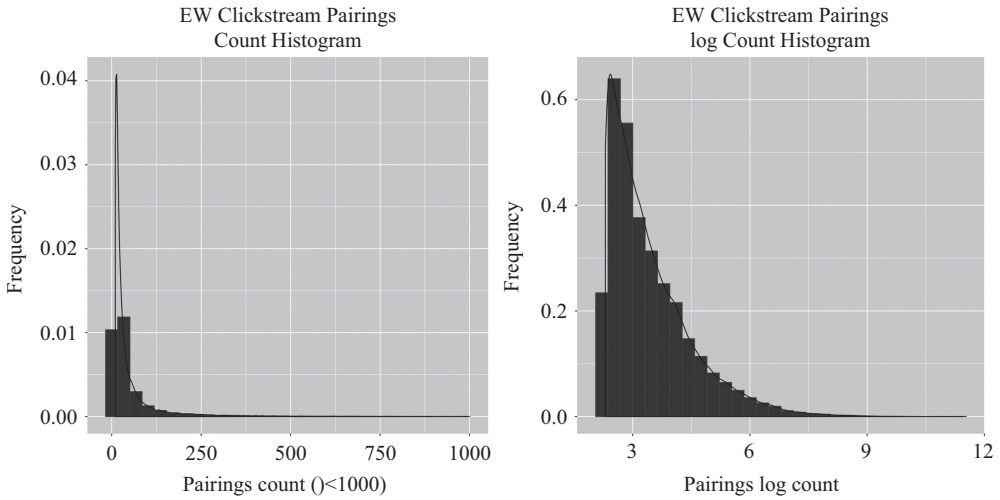

Figure 1.7 A histogram of the English Wikipedia Clickstream raw counts of the referrer and requested sites pairings counts is given in the left panel. Note that in order to produce a viewable plot, counts greater than 1 000 were excluded. The right-hand panel is a histogram of the log of the pairings counts. Both plots have smoothed density overlays.

https://doi.org/10.1017/9781316544778.002 Published online by Cambridge University Press

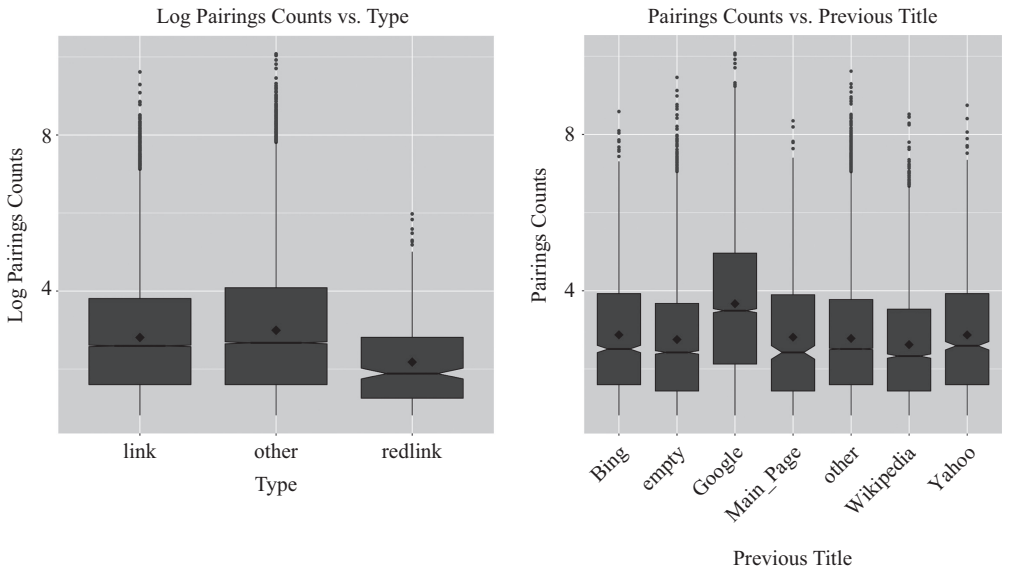

Figure 1.8 English Wikipedia Clickstream link type log counts (left panel) and the log counts of the previous titles (right panel). All the categories show skewed behavior.

(Type) log counts, and the right-hand plot is a box-whisker plot of the pairings of Previous Titles log counts. Note the skew in each of these plots. This suggests a Gaussian model is likely an inappropriate model type.

The EDA for the clickstream data show two important points to consider when choosing and constructing a model. The first is that the number of pairings fails to follow a Gaussian distribution function, and hence, normal linear regression is likely to be an inadequate and inappropriate model. The second point is that the categorical variables type and previous title show several zero counts in the two-way table which means either the interaction of the two variables will fail as a predictor, or remedial measures such as combining levels must be made if an interaction is to be used. Subject-matter expertise should be used prior to implementing remedial measures to assure conformance to the study intention.

1.9 Summary

Data sets usually contain variables with missing values and possible outliers. Rarely are they as well behaved as those in many textbooks, specifically, textbooks on empirical modeling methods. Messy data are seen in some of the tables and plots in this handbook. EDA is not only critical for finding and constructing optimal models based on response distributions, it is mandatory for identifying unbalanced data, missing values, sparse cross-classifications of variables, and the like. Possibly severe consequences can result if EDA is ignored as a precursor to model identification and construction. Throughout this handbook we will make reference to the exploratory analyses presented in this chapter, and we will build upon these basic descriptions as we construct models for specific data situations.

https://doi.org/10.1017/9781316544778.002 Published online by Cambridge University Press

1.10 Further Reading

John Tukey elucidated the essence of EDA in Tukey (1977). The EDA techniques we have presented in this chapter are but a subset of the methods presented by Tukey's text.

Cleveland (1993) demonstrates the visualization of data including user interaction and animation. This book focuses on interfaces around large and complex data sets.

Good introductory texts on basic statistical methods include McKillup (2011), Freedman et al. (2007) and Agresti and Finlay (2008).

https://doi.org/10.1017/9781316544778.002 Published online by Cambridge University Press