Data Analysis for Environmental Science and Management

Bruce Kendall and Chris Costello
Donald Bren School of Environmental Science and Management
University of California, Santa Barbara

Spring 2006

Contents

Chapter 1

Introduction: What is Statistics?

Statistics is a topic that everybody loves to hate. However painful it is to learn, however, it is essential for transforming *data* into *information*. Modern environmental science and environmental management would be impossible without it. Not only do environmental practitioners need to know how to analyze data, they also need to be able to discern the quality of existing analyses (i.e., determine when someone is "lying with statistics" or is just so sloppy that the analysis contains no real information). The goal of this book is to help teach those skills.

> Jargon Alert: Statistics goes by a number of names. Economists and other social scientists often refer to it as *econometrics*. Biologists sometimes refer to *biometrics* or (more commonly) *biometry*. The adjectival forms (equivalent to "statistical") are *econometric* and *biometric*, respectively.

Statistics supports both the scientific objective of *advancing understanding* and the management goal of *making decisions*. You can use statistics to:

Describe patterns in individual variables and in the relationships between variables;

Test theories, both quantitative and qualitative, about underlying processes;

Make predictions about how the system under consideration might perform under other circumstances.

To do these, we perform three tasks that are (usually implicitly) part of every statistical analysis:

Parameter estimation of quantities such as the mean of a single variable or the slope of the relationship between two variables;

Hypothesis testing to determine whether our estimated parameters differ from pre-specified value of interest (often zero);

Model selection to determine which explanatory variables should be included (and in what form), and what probability distributions should be used to describe the unexplained "residual variation."

1.1 Advancing understanding

As an example, consider gasoline consumption. Economic theory suggests that as the price of a good increases, the demand for that good should decline. The question is, does this principle apply to gasoline? It might not: gasoline may be what economists call an "inelastic" good — that is, we need to have it, no matter what the price is. To address this question, we could collect data on per-capita gas consumption and inflation-adjusted gas prices, and see whether there is a negative relationship between them. Of course, the economic theory is "all else equal" — so if our data come from different times or locations, then we have to take into account factors such as overall affluence, transportation infrastructure (urban vs. rural areas, for example, or New York vs. Los Angeles), and awareness of the environmental and social externalities associated with gasoline consumption (for example, consumption patterns might change after the Arab oil embargo, or after global warming became a widely accepted and understood phenomenon). Thus, each *observation* in our dataset be not merely the gas price and consumption in a particular location and time, but associated measurements of factors such as income and attitudes (some of which might vary spatially [transportation infrastructure], others temporally [attitudes], and still others will depend on both [income]). We would then construct a model in which the *dependent variable*, per-capita gasoline consumption, was a function of all of the *independent variables*, including price but also all of the other factors

That might influence gasoline consumption. We would then use statistical procedures (probably some form of regression) to estimate the parameters of this model from our the data. We then examine the estimated parameter associated with price to answer questions such as:

- Is the parameter estimate negative?
- How confident are we that the parameter estimate is different from zero (no effect)?
- What is the magnitude of the parameter estimate? How precise is that estimate?

The first two questions address the conceptual question of whether gasoline is an elastic good. The last questions allow us to *predict* the effects of increasing gas prices, for example through a tax.

> JARGON ALERT: The dependent variable in an analysis is also called the *response variable*, and the independent variables are called *explanatory variables*. Variables are also referred to as *factors*, as in "What factors did you include in your model?"

1.2 Making decisions

Ingesting lead is harmful to the health of small children. In response to concerns by homeowners in a particular development, the EPA developed a protocol for rapidly assessing lead levels by analyzing floor dust, collected using baby wipes from a number of locations throughout the residence. The decision to be made was whether to do more extensive testing to determine the source of the dust. If the true median dust loading in a residence was greater than 50 $\mu g/ft^2$, then the planning team required followup testing. Otherwise, they decided that a dust lead hazard was not present and discontinued testing.

The *statistical* decision rule (which based on an estimate of the median lead level coming from a finite number of observations) has to take into account two possible types of error: classifying the house as contaminated when the true median is below the threshold, and classifying the house as

uncontaminated when the the true median is above the threshold. These error probabilities increase as the variability among observations increases. These error probabilities can be reduced by increasing the number of observations; but in this case the budget precluded collecting more than 50 observations. Based on a preliminary estimate of the amount of variation among observations, a decision rule was developed using a t-test applied to the log-transformed value of the lead levels, that would result in a 5% chance of misclassifying the house as uncontaminated when the true median lead level was exactly 50 μg/ft^2, and a 20% chance of misclassifying the house as contaminated when the true lead level was as low as 35 μg/ft^2. The rule is strongly precautionary, with a higher chance of incorrectly classifying a house as contaminated; this was chosen because of the severity of the health risks.

1.3 A systematic approach to statistical analysis

Throughout the book, we will apply the following approach to analysis:

1. Clearly formulate the problem, question, or decision that you are facing
    - What are the *quantities* that you need to estimate?
2. Write down a *statistical model* that relates the quantities of interest to the data you will collect (or have collected)
    - This model will include a *random* component that represents natural variability or sampling error
3. Estimate the *parameters* of the statistical model
    - In addition to the estimate of the *most likely* value, quantify your *uncertainty* in that estimate
4. Use the results to address your problem, question, or decision
    - Your report should include a quantification of the probability that your answer is incorrect

The nuts and bolts of statistics are in items 2 and 3; much of the creativity and careful thinking goes into items 1 and 4.

Table 1.1: Symbols used for common parameters that describe populations and can be estimated from data. These follow general statistical practice.

Parameter	Parameter estimate	Meaning
α	a, $\hat{\alpha}$	Parameters of an ANOVA
β	b, $\hat{\beta}$	Parameters of a regression
ϵ	e, $\hat{\epsilon}$	Residuals
μ	$\bar{x}$, $\hat{\mu}$	Mean
σ	s, $\hat{\sigma}$	Standard deviation

1.4 Notation and definitions

We generally eschew statistical theory in this book. Nevertheless, formulas abound, although they are less frequent than in many other texts. Thus we need some definitions and notations; deviations from what is laid out here will be explicitly defined in context.

A *sample* is a collection of *observations* of one or more variables. The number of observations is n, and the observations are indexed by the subscript i. Thus a sample of a single variable x would be the set $\{x_1, x_2, \ldots, x_i, \ldots, x_n\}$.

When referring to parameters of statistical models, we distinguish between the true (generally unknown) values that apply to the whole *statistical population* (the universe of all possible observations), and the values that we estimate from our sample. We call the former "parameters," and indicate them with Greek letters; we call the latter "parameter estimates," and denote them either with the corresponding Roman letter (except for the mean) or with the Greek letter with a "hat" over it. Common examples are shown in Table 1.1. Note that 'α' and 'β' are also used to refer to type-I and type-II error rates.

> JARGON ALERT: In colloquial use, "sample" is often used to mean "observation" (e.g., "take a water sample"; "How many samples did you take?")

Finally, we will often use the sigma-notation for summation. Thus

$$\sum_{i=1}^{n} x_i$$

is the same as $x_1 + x_2 + \cdots + x_n$.

1.5 Computer software

The ideal statistical software would be (1) comprehensive (it does all the procedures you need it to do), (2) robust (it always gives the right answer), (3) easy to use, and (4) inexpensive (number 2 sounds like a no-brainer, but many of the formulas that were developed in the days of doing statistical calculations by hand can be subject to severe problems due to round-off error when applied to large datasets in computers). However, there is no software that satisfies all of these.

One solution is the open-source software R (www.r-project.org), which does everything and does it well (most statisticians develop new computation tools for this program first), and the price is right — free. However, it is an unfriendly interface for the beginning or casual user: you have to enter commands in a scripting window, and a complex analysis is like writing a program.

At the other extreme are commercial programs which are largely comprehensive and robust, and have a GUI interface that is more-or-less easy to use. However, unless your institution has a site license, they are expensive, on the order of $1000 per computer. These include SAS, JMP, S-plus, SPSS, and Stata. JMP is probably the easiest to use, but it is the least comprehensive. Nevertheless, it should be adequate for most applications.

Finally, there is Microsoft Excel. Excel's Analysis Toolpak has a number of basic procedures, such as t-tests, ANOVA, and OLS linear regression. However, many of Excel's statistical calculations are unreliable. Early versions of the software could not even correctly calculate a variance if the values involved were large (because of severe round-off error), although this appears to have been corrected in Excel 2003. More perniciously, the underlying probability distributions are known to have bad values in the tails, and we have seen linear regressions where the standard errors are reported (incorrectly) to be effectively zero — like the formerly problematic variance correction,

this results from using non-robust formulas in the calculations. A number of add-ons have also been developed, ranging in price from free to a few hundred dollars (there is a list, with links, at `www.mathtools.net/Excel/Statistics/`). There is generally a positive correlation between the price and the number of procedures it provides, and most have probably not been well tested for robustness. A free add-on that is quite useful is PopTools, available at `www.cse.csiro.au/poptools/`. This is designed primarily for demographic population modeling, but it contains robust replacements for all of Excel's statistical functions, and adds a variety of useful tools, including the ability to do bootstrapping.

1.6 Further Reading

Two of the foremost biometry textbooks are:

- Sokal, R. R., and F. J. Rohlf. 1994. *Biometry*, 3rd ed. W. H. Freeman.
- Zar, J. H. 1998. *Biostatistical Analysis*, 4th ed. Prentice Hall.

A couple of good econometric textbooks are:

- Green, W. H. 2000. *Econometric Analysis*, 4th ed. Prentice Hall.
- Woolridge, J. M. 2003. *Introductory Econometrics*, 2nd ed. Thomson.

All of these texts will provide much greater theoretical depth than we cover in this reader. The biometry texts focus strongly on hypothesis testing, experimental design, and analysis of variance, whereas the econometric texts focus almost exclusively on regression.

Chapter 2

Visualizing and Summarizing Data

2.1 Motivation

A quantitative approach to environmental problem solving can be broken down into a number of steps. We assume that you start with a *qualitative problem*, such as a decision or determination of causality that needs to be made.

1. Formulate quantitative hypotheses and questions that will help you address your general question or issue.

2. Determine the variables to observe.

3. Collect and record the data observations.

4. Study graphics and summaries of the collected data to discover and remove mistakes and to reveal low-dimensional relationships between variables.

5. Choose a model describing the important relationships seen or hypothesized in the data.

6. Fit the model using the appropriate modelling technique.

7. Examine the fit using model summaries and diagnostic plots.

8. Repeat steps 5-7 until you are satisfied with the model.

9. Interpret the model results in light of your general question.

Items 1 and 2 are largely matters of logic; we will illustrate these frequently in lecture and problem sets. We will not address the practical issues of data collection, although we will discuss the statistical issues associated with designing a sampling scheme. Items 5-9 form the bulk of the class; today we quickly review summary statistics and graphical methods for understanding your data.

2.2 Examples & Questions

1. A site has suffered the release of a toxic chemical (TcCB) into the soil, and the company responsible has undertaken cleanup activities. How should we decide whether the cleanup has been adequate?

2. We are trying to design a fuel-efficient car, and want to know what characteristics of current cars are associated with high fuel efficiency.

2.3 Evidence or Data

1. In addition to samples of TcCB concentration (measured in ppb) in the soils at the cleanup site, we also have samples of concentrations at an uncontaminated "reference" site with similar soil characteristics. The concentrations of TcCB at the reference site are not zero, and we need to determine what the normal levels of this chemical are.

2. We have data on 60 models of cars, with measurements of fuel efficiency (mpg), fuel consumption (gallons per 100 miles), weight (lbs), engine displacement (cubic inches), and car class (small, sporty, compact, medium, large, and vans).

2.4 Technique: single data series

2.4.1 Summary statistics

Central tendency

There are three commonly used measures of central tendency, or "average".

- The *arithmetic mean*,
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{2.1}$$
is the common-sense notion of average.

- The *mode* is the most likely value for an observation — the peak of the distribution.

- The *median* is the value exactly half of the observations are greater than and half are less than — the "midpoint" of the data.

If the data are normally distributed, the mode, the median and the mean will be about the same, but if the data are skewed then the median and mode will give more "typical" values than will the mean. The mean is also much more sensitive than the other two to rare extreme observations.

Variability

There are five commonly used measures of variability in data.

- The simplest is the *range*, the difference between the minimum and maximum values. However, because the extreme values are usually rare, another sample of the same population may give a very different range. Thus this is not particularly useful.

- The *interquartile range* is the distance between the 25^{th} percentile (the value below which 25% of the observation fall) and the 75^{th} percentile (the value below which 75% of the observation fall). This measures the range of the middle half of the data, and is quite robust.

- The *variance* is the mean squared distance between an observation and the sample mean:

$$\text{var}(x) = s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{2.2}$$

  If the data are normally distributed, this provides a complete description of the variability. It is also used in a large number of standard statistical techniques. Note that it has units that are the square of the units of the original observation.

- The *standard deviation* is the square root of the variance:

$$\text{sd}(x) = s_x = \sqrt{\text{var}(x)}. \tag{2.3}$$

  This has the same units as the observations; it is proportional to the "width" of the distribution.

- For many types of observations, the variance and standard deviation get larger as the mean gets larger. Also, they will differ depending on the units used in the observations. The *Coefficient of variation* rescales the standard deviation by the mean:

$$CV(x) = \frac{s_x}{\bar{x}}. \tag{2.4}$$

  This measures variation on a scale that is proportional to the mean, and has no units; it is a useful index for comparing variability across disparate datasets that may have been measured on different scales.

Shape

The mean and variance are known as the first and second "moments" of a distribution. Data that are not normally distributed also have higher moments; the most important are skew and kurtosis.

- The *skew* is a measure of how asymmetric the data are, with most of the observations piled up at one end and a few extreme observations at the other. If the long tail extends out to the right, the data are called *right* or *positively* skewed, and the opposite for tails to the left.

- The *kurtosis* measures the extent that the data are concentrated in the center and tails, as opposed to intermediate distances from the mean. Data with positive kurtosis have lots of points near the mean and long tails of extreme values; this is known as a *leptokurtotic* distribution. Data with negative kurtosis have points spread relatively evenly along an interval around the mean, and abrupt upper and lower bounds; this is known as a *platykurtotic* distribution.

Large values of skew or kurtosis mean that analyses that assume normally distributed data may not perform well. Often this can be redressed by transforming the data (see below).

2.4.2 Accuracy: Precision & bias

Accuracy has two aspects. *Precision* measures the repeatability of the estimate of a statistic. For example, if we went out and got a new set of observations, how similar would the resulting mean be to the original estimate? *Bias* represents a systematic over- or under-estimate.

The sample mean is an unbiased estimate of the population mean. Its precision can be estimated from the *standard error*:

$$\text{SE} = \frac{s_x}{\sqrt{n}}. \tag{2.5}$$

If the data are reasonably close to normally distributed and n is more than about 20, then the *95% confidence interval* of the mean ranges from $\bar{x} - 1.96\text{SE}$ to $\bar{x} + 1.96\text{SE}$. Notice precision can be increased by increasing the sample size (n).

Many statistics have a known bias, which can be corrected for. For example, the factor $(n-1)$ in the denominator of the variance calculation is a bias correction factor (if we simply took the arithmetic mean of the squared deviations — dividing by n — then we would underestimate the variance).

Bias can also arise in the data themselves — one of the central problems of sampling design is to avoid this bias.

JARGON ALERT: Precision is also known as *efficiency*.

2.4.3 Visualizing data

There are several effective ways to display a data series, each with different strengths and weaknesses (fig. 2.1).

- A *strip plot* shows every data value on the horizontal axis, adding a vertical "jitter" to ensure that nearby points are distinguishable. This is a graphical equivalent of the data table: it shows all the raw data, but no summaries.

- A *histogram* places the observations into intervals (also called bins or classes) and then displays the number of observations in each bin. This allows you to easily see where the observations are concentrated, and whether the upper and lower observations end abruptly or tail off gradually. However, the visual impact of a histogram can be changed radically by changing the width of the intervals, and there is no universal rule for the optimum interval size.

- If we could take larger and larger sample sizes from the population and make the intervals of the histogram smaller and smaller, we would end up with a picture of the true probability distribution of the data. A *density plot* approximates this with the existing data. For each point on the horizontal axis, the local density of observations within a nearby neighborhood is calculated, often weighted by how far away the observations are. Neighborhood size affects how smooth the density plot is, but intermediate values allow you to see fairly fine structure without getting a bump for every observation.

- A *boxplot* (also called a box-and-whisker plot) is a graphical representation of several important summary statistics. It consists of:

  - A "box" defined by the 25^{th} and 75^{th} percentiles.
  - A line or point on the box at the median.
  - A line or point on the box at the mean (not all programs do this).
  - A line ("whisker") drawn from the 25^{th} percentile to the smallest observation within one "step" (see below) of the 25^{th} percentile.
  - A line drawn from the 75^{th} percentile to the largest observation within one step of the 75^{th} percentile.

  - Observations more than one step away from the box are plotted with a symbol or line.

  A "step" is 1.5 times the difference between the 75^{th} and 25^{th} percentiles. If the observations come from a normal distribution then you should see outside values only 0.7% of the time; thus these values can be viewed as "extreme". It is a good idea to check these points to make sure that they are not data-entry errors, especially if they differ from the rest of the data by a factor of ten.

- A *Q-Q plot* ("quantile-quantile") compares the observations to a theoretical probability distribution, usually the normal. It plots the quantiles of the data against the quantiles of the distribution; if the data fit the distribution exactly, then the plot would be a straight line. We should not be surprised by deviations in the tails of the data, but deviations in the center of the data mean that the distribution is not a good description of the data. The overall shape of the curve, if it differs from a straight line, tells us something about the higher moments:

  - if it curves up, the data have *negative skew*;
  - if it curves down, the data have *positive skew*;
  - if it is S-shaped, the data are *platykurtotic* (evenly spread);
  - if it is shaped like a backwards S, the data are *leptokurtotic* (concentrated near the mean).

- The least informative plot is the *mean and error bars*. This shows a symbol at the mean of the data and lines encompassing the confidence interval of the mean. Be sure that your figure legend says that this is what the error bars represent, for some authors use the error bars to show the standard error of the mean or the standard deviation of the data (these are less easy to quickly interpret). The only real reason for using this type of plot is if you are comparing a lot of samples, so that a boxplot or strip plot would be too busy.

2.4.4 Data transformation

Many of the standard statistical techniques assume that the data are approximately normally distributed. Even if your data are not, it is often possible

Table 2.1: The values of the Box-Cox parameter (λ) that correspond to common data transformations.

λ	transform
1	none
0.5	square root
0	log
-0.5	inverse square root ($1/\sqrt{x}$)
-1	inverse ($1/x$)

to *transform* the data so that they are, by applying a function to each data value. You then do analysis on the transformed data. However, you can *not* back-transform the results.

The two transformations that are used most commonly are the log transform and the square root transform (here and throughout, we use "log" to refer to the natural logarithm). The log transform is particularly useful for comparing data that are on different scales. These are both special cases of a general family of transformations called the *Box–Cox* transform:

$$z = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log(x) & \lambda = 0. \end{cases} \tag{2.6}$$

The parameter λ controls what the transformation is; values that correspond to "standard" transforms are shown in Table 2.1. The "best" transform is the one that has the lowest variance for the transformed data. λ need not be confined to the five standard values.

2.5 Application of Technique: single data series

The summary statistics of TcCB concentrations at the reference and cleanup sites are shown in Table 2.2, and various graphical representations of the Reference site data are in Figure 2.1. The mean and variance are substantially

Table 2.2: Summary statistics for TcCB in the reference and cleanup sites. Notice that the CV, skew, and kurtosis are dimensionless.

Statistic	Reference	Cleanup
Mean (ppb)	0.60	3.92
Mode[a] (ppb)	0.58	0.25
Median (ppb)	0.54	0.43
Range (ppb)	1.11	168.55
Interquartile range (ppb)	0.44	0.87
Variance (ppb^2)	0.08	400.62
Standard deviation (ppb)	0.28	20.02
Coefficient of variation	0.47	5.11
Skew	0.90	7.72
Kurtosis	0.13	62.67

[a] Estimated from density plot.

higher in the cleanup site, but the mode and median are actually lower. Both samples are skewed.

The best Box-Cox transform for the Reference data is the log transform, so we use this on both samples. This allows us to effectively plot both samples on the same scale (Figures 2.2 and 2.3). From these plots it becomes clear that most of observations at the cleanup site are within the range of variation of the reference site (some are even lower!). But there are few observations with very high concentrations, suggesting that there are spots that were either missed by the cleanup or had extremely high initial concentrations. The conclusions that could be drawn from examining these data are:

1. The cleanup has not yet reached reference levels;

2. Rather than continuing cleanup efforts on the entire site, it would be more effective to identify the remaining hot spots and focus efforts on them.

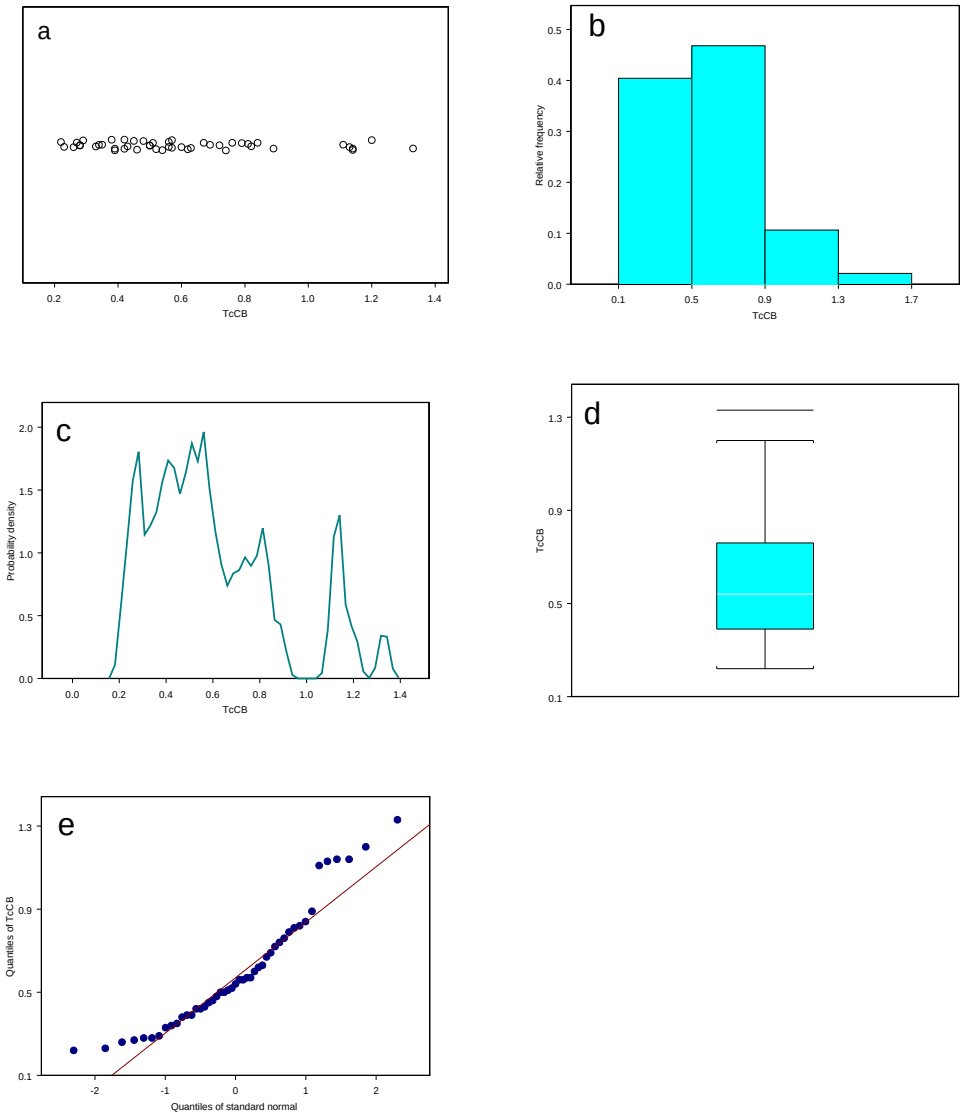


Figure 2.1: Summary plots of TcCB concentrations at the reference site. (a) Strip plot. (b) Histogram. (c) Density plot. (d) Box plot. (e) QQ plot.

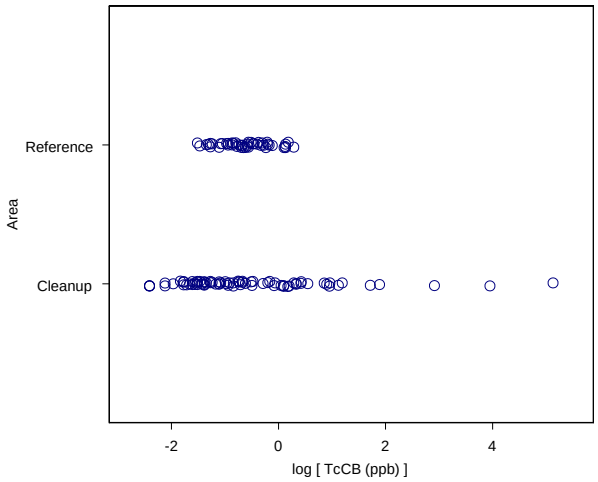


Figure 2.2: Strip plots of the log of TcCB concentrations at the reference and cleanup sites

2.6 Technique: relating data series

Often each observation will contain measurements of two or more variables (e.g., depth to water table and concentration of a chemical). Much of this class will be about discovering relationships among variables, but there are some simple calculations that show the extent to which the variables co-vary in a linear fashion.

The *covariance* of two variables, x and y, is

$$\text{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}). \tag{2.7}$$

If x and y tend to move up and down together, then the covariance will be positive. If they tend to move in opposite directions they will have a negative covariance. If they are unrelated then the covariance will be zero. The covariance may also be zero if they are related in a strongly nonlinear way.

Notice the similarity between the equation for the covariance and that for

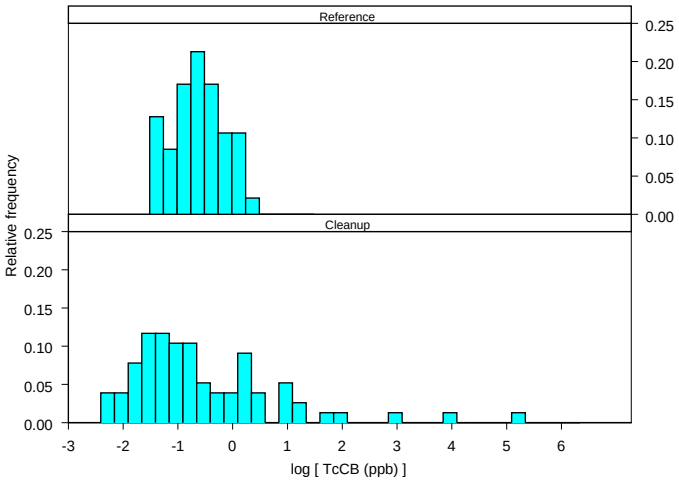


Figure 2.3: Histograms of the log of TcCB concentrations at the reference and cleanup sites

the variance. In fact, the covariance of x with itself is simply the variance of x.

The magnitude of $\text{cov}(x, y)$ is affected by two things: the extent to which the variables move together, and the variances of x and y. Often we want to focus on the former, which is measured by the *correlation coefficient*:

$$r = \frac{\text{cov}(x, y)}{\text{sd}(x)\text{sd}(y)}. \tag{2.8}$$

r ranges from -1 (when x and y move up and down in perfect opposition) to 1 (when x and y move up and down in perfect synchrony). The correlation of a variable with itself is always 1. The square of the correlation coefficient (r^2) measures how much of the variation in x can be predicted by knowing y (and vice-versa).

If there are more than 2 variables, we can produce a table showing the covariance or correlation between all possible pairwise combinations. This is known as the *covariance matrix* or *correlation matrix*. The matrix is symmetric ($\text{cov}(x, y) = \text{cov}(y, x)$), so often the upper right half of the matrix is omitted.

The best way to visualize the relationship between two variables is the *scatterplot*. With multiple variables, most statistical packages can automatically plot all pairwise combinations.

2.7 Application of Technique: relating data series

The covariance matrix of the automobile fuel efficiency data is:

	Weight	Disp.	Mileage	Fuel
Weight	245883	21573	-2014.5	323.68
Disp.	21573	2933	-179.9	29.29
Mileage	-2014	-180	23.0	-3.56
Fuel	324	29	-3.6	0.57

Weight has by far the greatest covariance with mileage; but weight also has by far the largest variance, so it is unclear how strong the relationship is. For this, we turn to the correlation matrix:

	Weight	Disp.	Mileage	Fuel
Weight	1.00	0.80	-0.85	0.86
Disp.	0.80	1.00	-0.69	0.71
Mileage	-0.85	-0.69	1.00	-0.98
Fuel	0.86	0.71	-0.98	1.00

Now we see that it is the fuel consumption that has the strongest relationship with mileage. This is not a useful result, however, as mileage and fuel consumption are just inverses of each other. Of the remaining variables, weight has a much stronger correlation with mileage than does engine displacement. Indeed, weight explains about 70% of the variation in mileage ($r^2 = 0.72$).

Before drawing conclusions, we need to ensure that the relationships between the variables are approximately linear. The scatterplots confirm this (Figure 2.4). Thus, all else being equal, it seems that the best way to increase fuel efficiency is to reduce vehicle weight.

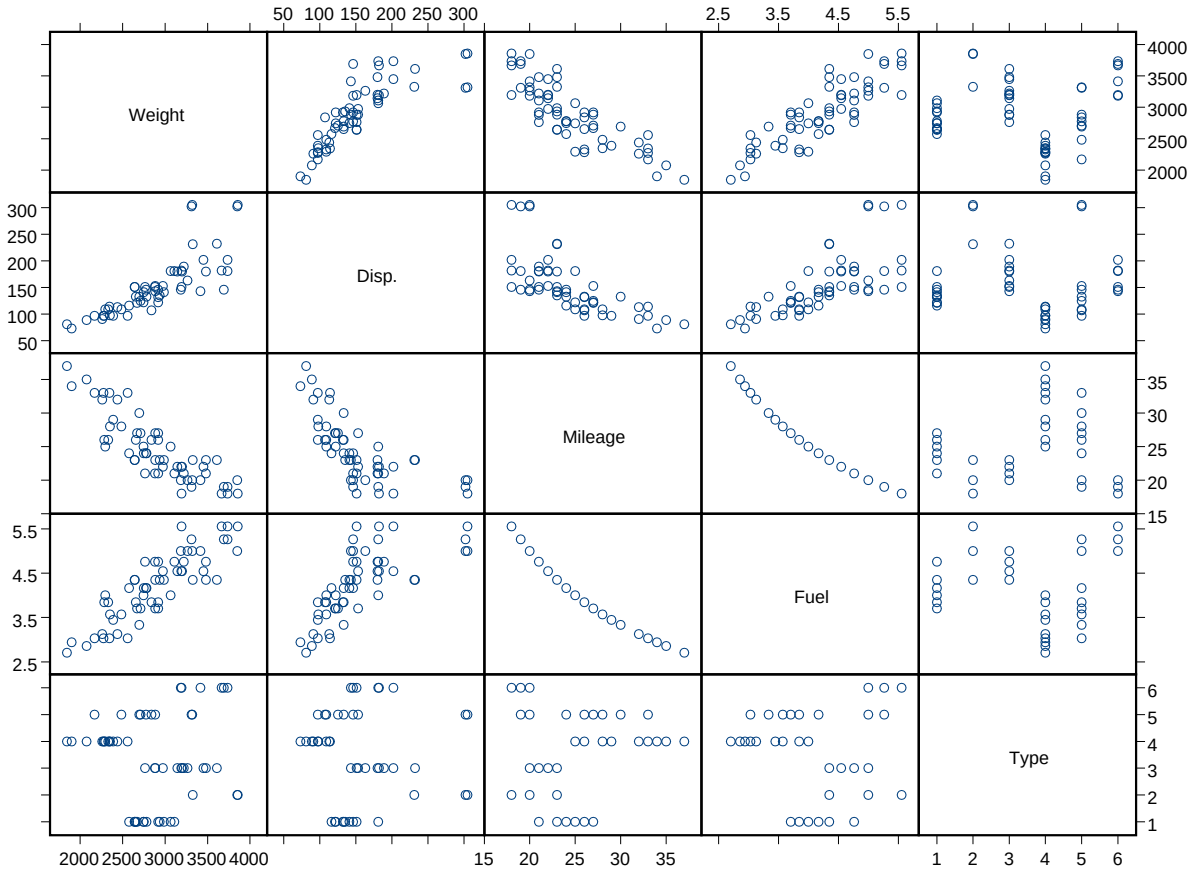


Figure 2.4: Scatterplots of the fuel data. In each panel, the horizontal axis is the variable named above or below it, and the vertical axis is the variable named to the right or left.

2.8 Presenting statistical results and graphics

You should never simply cut and paste the output of the statistics program, for two reasons. First, there is generally more output than is needed to support the conclusion you are trying to draw. Second, most programs print out far more significant figures than are justified by the data.

Many statistical results can be succinctly presented in the text; we will show you how to do this as we go along. For example, the standard way of presenting means is:

> The mean soil concentrations of TcCB ($\bar{x} \pm \mathrm{SE}$) were 0.60 ± 0.04 ($n = 47$) at the reference site and 3.92 ± 2.28 ($n = 77$) at the cleanup site.

ALWAYS include the sample size or degrees of freedom in any description of statistical results. Notice that the standard error and the sample size provide enough information to calculate the variance, so you don't need to report this separately. Whenever you use the "±" notation, you need to make clear what you are referring to.

More extensive results can be placed in tables, but structure the table to showcase the information *you* want to convey, not simply the way the output comes. Columns of tables should be clearly labelled, and the table should have a caption that explains what it is. If the column titles need detailed explanation, then use footnotes. Use a horizontal line to separate the column titles from the data. Use other horizontal lines sparingly, generally only to separate major conceptual sections, or if they are needed to guide the eye across large areas of white space.

Figures should have clearly labelled axes, and if there are multiple symbols these should be labelled in a legend on the graph itself or explained in the figure caption. The figure caption should also provide any relevant additional information about the variables, and should open with a brief statement about what the figure shows. The reader should be able to understand the figure from the figure and its caption, without digging through the rest of the text.

By convention, the table legend goes above the table, while the figure legend goes below the figure.

Chapter 3

Probability theory

3.1 Motivation

Although we have some idea of how environmental systems work, we recognize that there exists some random, or *stochastic*, component to the processes we study. The randomness may arise from uncertainty about how a process works, from errors in our measurements, from random exogenous environmental shocks. **Probability theory** is a formal framework in which to study and better understand randomness; it therefore forms the basis for statistical analysis.

3.2 Example: Detecting a Leak

Storage containers for hazardous materials are not guaranteed to maintain structural integrity forever. The probability of leakage of a candidate storage design has been estimated on the basis of laboratory experiments. As part of the implementation of the storage plan, a monitoring program will be initiated to assess whether any leakage has occurred. This assessment will be based on results from water and soil samples near the site. However, the instruments used to measure these samples are imperfect, that is, they occasionally record false positives or false negatives. We wish to answer 2 questions.

1. Suppose the test is positive (soil measurements say that a leak has occurred), what is the probability that a leak, in fact, has occurred at the site?

2. Suppose the test comes up negative (i.e. no leak has been detected), what is the probability that a leak actually occurred?

3.3 Evidence or Data

As we will see in the next section, all we require to answer these questions are some summary statistics. Suppose that 10% of 40-year-old tanks leak, our test method successfully detects a leak 95% of the time, and that the test returns a false positive 10% of the time. This can be summarized as:

- Pr(Leak occurs) = 0.10
- Pr(Leak detected given leak occurs) = 0.95
- Pr(Leak detected given no leak occurs) = 0.10 (false positive)

These imply:

- Pr(No leak occurs) = 0.90
- Pr(No leak detected given leak occurs) = 0.05 (false negative)
- Pr(No leak detected given no leak occurs) = 0.90

At first glance, we might think that the answer to the questions above are simply 0.95 and 0.05, from above. These answers are incorrect. Loosely speaking, they ignore the information that leaks occur 10% of the time. Later we'll see how to incorporate this information

This is just one of many types of problems in basic probability theory. Since probability theory is such a fundamental concept, we'll develop the basic theory from first principles, eventually acquiring the tools required to answer the leak detection question.

3.4 Techniques & Tools

We need to develop some basic definitions in order to organize a discussion about probability[1].

[1]Many of these notes are adapted from: Larson. 1982. Introduction to Probability Theory and Statistical Inference. Wiley.

3.4.1 Sample Space

Definition 3.4.1
The sample space S is the set of all possible outcomes that might be observed.

You should always describe the sample space for experiments. Here are some examples:

- Determine the species of a randomly selected mosquito from the four known mosquito species living in Goleta Slough. $S = \{1, 2, 3, 4\}$.
- Measure the concentration of a toxic substance in the flesh of 100 fish. Record the two highest concentrations. $S = \{(c_1, c_2) : 0 \leq c_1 \leq 1, 0 \leq c_2 \leq c_1\}$, where 1 is the maximum possible concentration.

Definition 3.4.2
An event A is a subset of a sample space S: $A \subset S$.

- Recreational fishers can choose to fish at any one of 4 sites (denoted I, II, III, and IV) along a river. We are interested in the site chosen by the most fishers in the month of June. The sample space, that is the set of all possible outcomes of this experiment is $S = \{I, II, III, IV\}$. A possible outcome of this experiment is that site III is the site chosen by the most fishers in the month of June. So, we can let $E3$ be the event that site III is the site chosen by the most fishers in the month of June.

This definition just formalizes the notion that after the experiment is over, some outcome (i.e. some element of S) will be observed.

3.4.2 Some Set Operations

In order to calculate statistics of interest, we often have to manipulate sets of items. Two definitions, the union of two (or more) sets and the intersection of two (or more) sets, are required.

Definition 3.4.3
The union of two sets, A and B, written $A \cup B$, is the set that consists of all the elements that belong to A or to B or to both.

It may help to think of the union symbol ($\cup$) as the word "or".

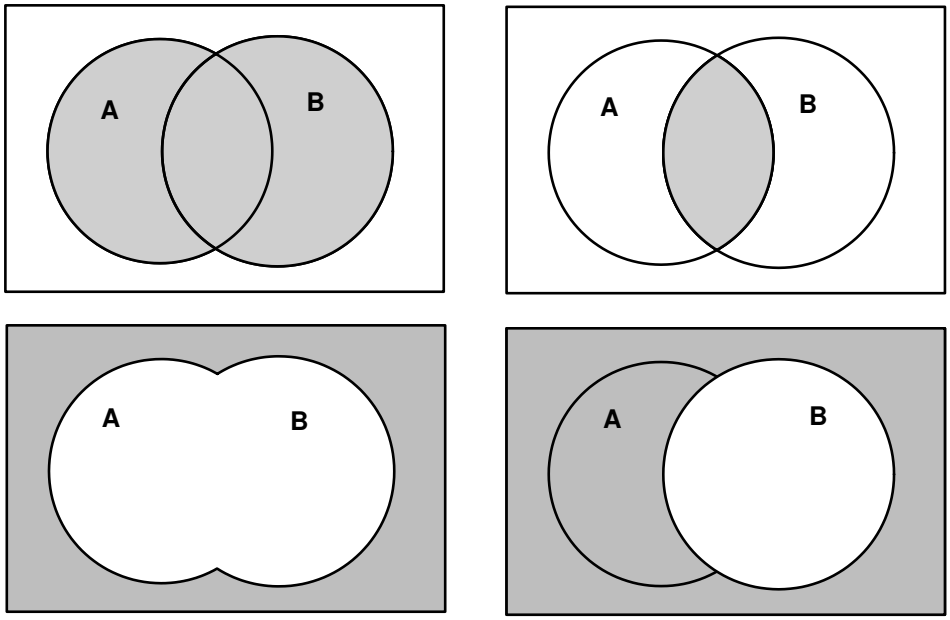


Figure 3.1: Combinations of two sets. Clockwise from upper left shaded area shows: $A \cup B$, $A \cap B$, $\bar{B}$, and $\bar{A} \cap \bar{B}$.

Definition 3.4.4
The intersection of two sets, written $A \cap B$, is the set that consists of all elements that belong to both A and to B.

It may help to think of the intersection symbol ($\cap$) as the word "and". Figure 3.1 gives some examples. The shaded area of the panels represent the cases: $A \cup B$, $A \cap B$, $\bar{A} \cap \bar{B}$, and $\bar{B}$, respectively. The bar over a set refers to the complement of that set.

3.4.3 Probability Axioms

Many different outcomes can result from an experiment. The probability of an event is another way of describing the "chance" or the "likelihood" of that event occurring. Probabilities are dimensionless real numbers associated with events. In words, the probability axioms are:

1. The relative frequency of an event that is certain to occur must be 1, because it will occur 100% of the time.

2. The relative frequency of events is never negative; so probabilities are never negative.

3. If two events cannot occur simultaneously (i.e. they are mutually exclusive), the probability of occurrence of the union of those events is the sum of the probabilities of the two events[2].

Mathematically, these definitions are written as follows:

1. $P(S) = 1$,

2. $P(A) \geq 0$ for all $A \subset S$,

3. $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$.

Almost all results in probability theory are derived from these simple conditions. We won't derive or prove them here, but we will frequently introduce new ones when they are required to solve a problem.

[2]Mathematically, the probability the union of two mutually exclusive events is the sum of their probabilities. For example, with the roll of one die, the probability of rolling a 1 or a 2 is $\frac{1}{6} + \frac{1}{6} = \frac{1}{3}$.

Definition 3.4.5
The complement of A, denoted $\bar{A}$, can be thought of as "not A" and has probability: $P(\bar{A}) = 1 - P(A)$.

In the leak detection example above, the sample space for the outcome of an experiment testing for a leak is: $S = \{\text{no leak detected, leak detected}\}$. Let A be the event that "a leak occurs". Then $P(A) = 0.10$. We now know that $\bar{A}$ is the probability that a leak does not occur, $P(\bar{A}) = 0.90$. Some important results follow:

- $P(\emptyset) = 0$ for any S,
- $P(\bar{A} \cap B) = P(B) - P(A \cap B)$,
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$,
- If $A_1, A_2, ..., A_k$ are mutually exclusive events, then $P(A_1 \cup A_2 \cup ... \cup A_k) = P(A_1) + P(A_2) + ... + P(A_k)$.

3.4.4 Counting Techniques

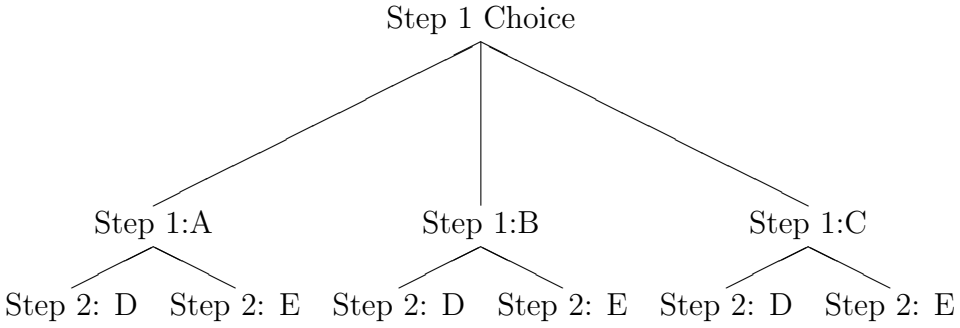
For a finite sample space, if all events are equally likely, then the probability of an event, A, is just the ratio of the number of elements in A to the number of elements in S. Often, counting the number of elements is not a trivial task.

Definition 3.4.6
The multiplication principle states that: If a first operation can be performed in any of n_1 ways, and a second operation can be performed in any of n_2 ways, then both operations can be performed in $n_1 \cdot n_2$ ways.

As an example, suppose we are interested in the output of a particular toxic chemical from a manufacturing facility. The manufacturing process takes two steps: In the first step, one of 3 toxic chemicals is used. In the second step one of 2 (different) chemicals is used. There are a total of 6 possible chemical combinations that can be output. A tree diagram is a graphical representation of the multiplication principle. The example above is graphed as a tree diagram below:

More generally, we have the following definition.

Definition 3.4.7
An arrangement of n symbols in a definite order is called a permutation of the n symbols.



Using the multiplication principle, it should be clear that the total number of permutations of n symbols is $n(n-1)(n-2)...(2)(1)$, which is denoted $n!$, and is read "n factorial"[3]. Studying food webs requires a solid understanding of counting techniques. For example, suppose 5 species (called S1, S2, .., S5) are to be arranged in every possible order (one such order is S1, S2, S3, S4, S5). How many such orders exist? [The trick to remembering this rule is the following: the first position can be any one of the 5 possible species. The second position can be any of the 4 remaining species. The third position can be any of the 3 remaining species...So the answer is $5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$.

Even more generally, suppose we want to group things r elements at a time (instead of n at a time as in the factorial case). Then we have the following definition:

Definition 3.4.8
The number of r-tuples we can make ($r \leq n$), using n different symbols (each only once) is called the number of permutations of n things r at a time, and is denoted by ${}_nP_r$.

We can derive the following condition:

$$ {}_nP_r = \frac{n!}{(n-r)!} \tag{3.1}$$

Extending the food web example, how many permutations of 5 species are there taken 3 species at a time? The answer is $\frac{5!}{(2)!} = 5 \cdot 4 \cdot 3 = 60$.

Suppose that instead of the number of r-tuples, we want to know the number of subsets of size r, then we use the following definition:

[3]Note that, by definition, $0! = 1$.

Definition 3.4.9
The number of distinct subsets, each of size r, that can be constructed from a set with n elements is called the number of combinations of things r at a time; it is given by $\binom{n}{r}$.

And the number is computed as follows:

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} \tag{3.2}$$

Note that the number of combinations is smaller than the number of permutations. Two admissible *permutations* of species are $\{S1, S4, S5\}$ and $\{S4, S1, S5\}$, since with permutations, the order is what matters. The number of combinations of 5 species taken 3 at a time is just $\binom{5}{3} = \frac{5!}{3!(2)!} = 10$. This makes sense because there are exactly 3! different ways to arrange each permutation of three species. To avoid double counting, we only need one of them.

3.4.5 Conditional Probability

Often we know that some event has taken place, and we would like to know the probability of some other event taking place given this information. For example if we are interested in the probability of an oil tanker running aground, we may also wish to condition on whether the captain was drunk (Exxon Valdez should ring a bell). In this case, you might let A be the event that an oil tanker runs aground and B be the event that the captain is drunk.

We have the following definition:

Definition 3.4.10
The conditional probability of event A occurring, given that event B has occurred, is written

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{3.3}$$

where $P(B) > 0$.

Note that the conditional probability for A may be $>$, $<$, or $=$ the unconditional probability of A, $P(A)$. Based on previous experience, the probability of an oil tanker captain being drunk is .0023, and the probability of an oil tanker captain being drunk and running his ship aground is .0013. So the probability of an oil tanker running aground, given that the captain is drunk is simply $P(A|B) = \frac{.0013}{.0023} = .565$ or about 57%.

3.4.6 Independent Events

Often the appropriateness of a statistical technique hinges on whether two (or more) events are independent. The formal definition is:

Definition 3.4.11
Two events, A and B, are independent if and only if $P(A \cap B) = P(A)P(B)$.

That is, two events are independent if the occurrence of event B provides no information about the occurrence of event A. Furthermore, if the occurrence of the event B provides no information about the occurrence of the event A, then A and B are independent (that's the "if and only if" part of the definition). Consider the following example regarding the likelihood of a randomly selected individual contracting cancer:

	Contracts Cancer	Doesn't Contract Cancer
Exposed	0.50	0.20
Not Exposed	0.10	0.20

Where "Exposed" means the individual was exposed to a particular compound in question. Let A be the event that the individual has been exposed, and let B be the event that the individual contracts cancer. Then we have:

$$P(A \cap B) = 0.5 \tag{3.4}$$
$$P(A \cap \bar{B}) = 0.2 \tag{3.5}$$
$$P(\bar{A} \cap B) = 0.1 \tag{3.6}$$
$$P(A) = P(A \cap B) + P(A \cap \bar{B}) = 0.7 \tag{3.7}$$
$$P(B) = P(A \cap B) + P(\bar{A} \cap B) = 0.6 \tag{3.8}$$

To determine whether the events A and B are independent, note that $P(A \cap B) = 0.5 \neq (.7)(.6)$, so they are not independent.

3.4.7 Bayes' Theorem

An extremely useful theorem in statistics is Bayes' Theorem, after the Reverend T. Bayes (1764). First, we need to define a partition:

Definition 3.4.12
A partition, $E_1, E_2, ..., E_n$, of the whole sample space, S, cuts the whole sample space into n mutually exclusive pieces. $E_1, E_2, ..., E_n$ is a partition if $E_i \cap E_j = \emptyset$ for all $i \neq j$ and if $E_1 \cup E_2 \cup ... \cup E_n = S$.

Bayes' Theorem can be stated as follows: Let $E_1, ..., E_n$ be a partition of S. Then for any event $A \subset S$:

$$P(E_i|A) = \frac{P(E_i)P(A|E_i)}{\sum_{j=1}^{n} P(E_j)P(A|E_j)}, \quad i = 1, 2, ..., n \tag{3.9}$$

The motivating question for these notes is an example of Bayes' Theorem. See the answer below.

3.5 Application of Technique

At this point, we are prepared to answer our original question. First, we establish our notation to solve this problem. Define the following events:

- $L =$ the event that a leak occurs (so $\bar{L}$ is the event that a leak does not occur).
- $D =$ the event that a leak is detected (so $\bar{D}$ is the event that no leak is detected).

In the problem we were given the following data:

$$P(L) = 0.10 \tag{3.10}$$
$$P(D|L) = 0.95 \tag{3.11}$$
$$P(D|\bar{L}) = 0.10 \tag{3.12}$$
$$P(\bar{L}) = 0.90 \tag{3.13}$$
$$P(\bar{D}|L) = 0.05 \tag{3.14}$$
$$P(\bar{D}|\bar{L}) = 0.90 \tag{3.15}$$

Our instruments detect a leak, what is the probability that a leak has actually occurred? To answer this question, we apply Bayes' Theorem (equation 3.9), as follows:

$$P(L|D) = \frac{P(D|L)P(L)}{P(D|L)P(L) + P(D|\bar{L})P(\bar{L})} \tag{3.16}$$
$$= \frac{(0.95)(0.1)}{(0.95)(0.1) + (0.1)(0.9)} = 0.5135 \tag{3.17}$$

That is, the probability that a leak has occurred given that a leak is detected is only about 51% (despite the fact that the probability that a leak is detected given that a leak has occurred is 95%). This counter-intuitive result occurs because actual leaks are rare, so there is a high chance that a given "detection" is actually a false positive.

You may wish to confirm the following results:

- If no leak is detected, there is a 1% chance that one has occurred ($P(L|\bar{D}) = 0.0061$)
- If a leak is detected, there is a 49% chance that a leak has occurred ($P(\bar{L}|D) = 0.4865$)
- If no leak is detected, there is a 99% chance that there isn't one ($P(\bar{L}|\bar{D}) = 0.9939$)

Chapter 4

Random Variables and Distributions

4.1 Motivation

These notes define a random variable, and describe several useful distributions (both discrete and continuous). The concept of a random variable is central in analyzing data, and a good understanding of the commonly used distributions will aid your understanding of many techniques we'll develop in this course. In this chapter we also attempt to provide useful interpretations of the distributions for use in modeling work.

4.2 Example & Question

Here's one example (adapted from Millard and Neerchal): A hazardous waste facility monitors the groundwater adjacent to the facility to ensure no chemicals are seeping into the groundwater. Several wells have been drilled, and once a month groundwater is pumped from each well and tested for certain chemicals. The facility has a permit specifying an Alternate Concentration Limit (ACL) of 30 ppb for aldicarb. This limit cannot be exceeded more than 5% of the time.

If the natural distribution of aldicarb at a particular well can be modeled as a normal distribution with a mean of 20 and a standard deviation of 4, how often will this well exceed the ACL? Restated, the question is: What is the probability that the aldicarb level at that well will exceed 30 ppb on any

given sampling occasion, even if no aldicarb is discharged from the facility?

4.3 Evidence or Data

The required data are given above; actually, the information given above is in the form of *statistics*, the raw data have not been provided. Suppose, instead of providing you with statistics (the parameters of the normal distribution), you had access to the raw data. First, we need to define a random variable:

Definition 4.3.1
A random variable is the value of the next observation in an experiment. More formally, a random variable X is a real-valued function of the elements of a sample space S.

Put simply, a random variable is a variable that can take on different values, each with potentially different probabilities. Tomorrow's high temperature at Santa Barbara airport is a random variable. At the end of the day tomorrow, we will know the high temperature. The actual high temperature is called a realization of a random variable. Note that we usually denote random variables with capital letters (e.g. X), while realizations of a random variable, or observed numbers, are denoted by lower-case letters (e.g. x).

Back to the example, those data might be displayed as a histogram:

Definition 4.3.2
A histogram is a plot of the frequency of observation of a random variable over many discrete intervals.

Suppose you are presented with data of the aldicarb readings in naturally occurring soils. These data are plotted in figure 4.1 as a scatter plot and as a histogram with "bins" of integers from 10 to 32. In contrast, we are presented with a theoretical distribution (Normal distribution with particular parameters).

4.4 Technique

We defined random variable above, but it is useful to further break the definition into several categories.

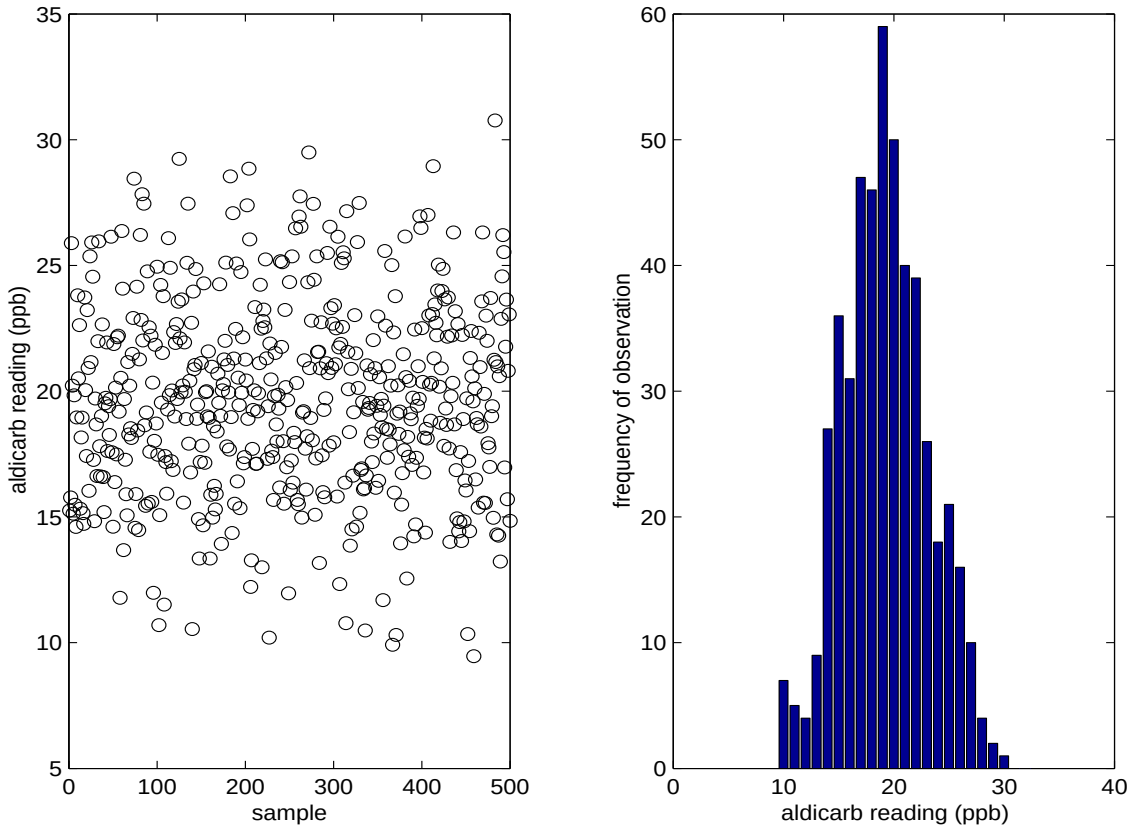


Figure 4.1: (a) Measurements of aldicarb in 500 groundwater samples. (b) The data of part a shown as a histogram.

Definition 4.4.1
There are two forms of numeric random variables:

- *A discrete random variable is a random variable that can take on only integer values (or values that can be converted to integers) (for example, the number of offspring of a grizzly bear sow).*
- *A continuous random variable is a random variable that can be any real number (for example, the concentration of ozone in Santa Barbara County on June 13).*

Random variables can also be non-numeric:

- *A categorical random variable (also called nominal) is a random variable that takes on a discrete qualitative value (for example, the sample space could be {red, green, blue} or {car, truck, minivan, SUV}.*
- *An ordinal random variable (or rank) is a categorical random variable that preserves some information about a (otherwise unmeasured) quantitative variable. For example, in the sample space {strongly disagree, disagree, agree, strongly agree}, the answers are in increasing order of agreement, but we have no measurement of the "distance" between them.*

The concept of a histogram is particularly intuitive for discrete random variables. Take the example above. For example, suppose 10 sows were were selected, and they had the following numbers of offspring: 1, 2, 1, 2, 2, 2, 2, 3, 1, and 2. A histogram of these data gives the relative frequency of 1 offspring, 2 offspring, and 3 offspring. These frequencies are: .3, .6, and .1, respectively. For discrete random variables that can take on many different values (such as the number of eggs laid by salmon), and for continuous random variables, it is useful to break the "bins" into discrete values that span the range of possibilities (see example above).

This brings us to the concept of a probability distribution, defined below:

Definition 4.4.2
For a discrete random variable, a probability distribution can be thought of as the histogram of outcomes if you could take a very large (approaching infinite) number of samples from the population. For a continuous random variable, the probability distribution is what the histogram would look like with an infinite number of samples where the histogram "bins" get narrower and narrower.

Definition 4.4.3
The probability density function (or pdf) of a random variable is the relative frequency of occurrence of that random variable. It is usually denoted $f(x)$. The pdf of a discrete r.v. is also called the probability mass function.

The "area" under the pdf is exactly one.

Empirical probability density functions are typically described graphically, or with summary statistics. Theoretical pdf's are usually described by a mathematical equation, with one or two parameters. In the next section, we describe many of the most commonly used probability density functions.

The height of a pdf at a particular value gives the probability of occurrence of that event. More formally, the random variable has probability $f(x)\Delta x$ of being in the interval $[x - \frac{\Delta x}{2}, x + \frac{\Delta x}{2}]$, where Δx is small. For continuous random variables, this probability goes to zero as we let Δx go to zero. This problem is circumvented when we use the following concept:

Definition 4.4.4
The cumulative distribution function (or cdf) of a random variable, X, is a function, $F(X)$ that gives the probability that the random variable is less than or equal to some number, x.

If the pdf is given by the function $f(x)$, then we can write the cdf of a continuous random variable mathematically as follows:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt \tag{4.1}$$

The cdf for a discrete random variable is written:

$$F(x) = P(X \leq x) = \sum_{i \leq x} f(i) \tag{4.2}$$

4.4.1 Useful Theoretical Distributions

Knowing the characteristics of commonly-used distributions is very important. We'll start with the distributions of discrete random variables, and then turn to those for continuous random variables. In each case, we'll provide an interpretation of the random variable, a mathematical description of the pdf. The mean and variance of each distribution are shown in table 4.1.

Discrete Random Variables

Distribution	Mean	Variance
Binomial	np	$np(1-p)$
Poisson	λ	λ
Geometric	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Uniform	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Normal	μ	σ^2
Gamma	$\frac{n}{a}$	$\frac{n}{a^2}$
Exponential	μ	μ^2

Table 4.1: Means and variances of some common theoretical distributions. Parameters are defined in the text.

1. Bernoulli Trial. A Bernoulli "trial" is an experiment with two possible outcomes, success or failure. We can write the density as follows:

$$f(x) = \begin{cases} p & \text{if} \quad x = 1 \\ 1-p & \text{if} \quad x = 0, \end{cases} \tag{4.3}$$

   where p is the "probability of success." As a simple example, the gender of each offspring of the grizzly bears above may be thought of as a Bernoulli trial. Each offspring has probability of, say, 0.51 of being a female (arbitrarily labeled a "success", and given a value of 1) and probability 0.49 of being a failure (male; a value of 0). Note that, although we have coded it with digits, bear gender is *not* a numeric variable!

2. Binomial Distribution. Closely related to the Bernoulli Distribution, the Binomial Distribution models the number of successes in n independent trials. The probability density function gives the probability of exactly x successes in n trials, given a probability p of "success". The density is:

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, ..., n, \tag{4.4}$$

   where $\binom{n}{x}$, also called the *binomial coefficient*, is defined in equation (3.2). Out of a random sample of 50 grizzly bear cubs, what is the probability that only 15 will be female? Restated, we want to know

the probability of exactly 15 successes in 50 trials, given that $p = 0.51$. We have $f(15) = \binom{50}{15}(0.51)^{15}(1-0.51)^{35} = \frac{50!}{15!35!}(0.51)^{15}(1-0.51)^{35} = 0.001327$ or about one tenth of one percent.

3. Multinomial Distribution. The multinomial distribution extends the binomial to the case when more than two outcomes are possible. If there are M possible outcomes, $\{X_1, X_2, \ldots, X_M\}$, each occurring with probability p_i, and there are n trials, the multinomial pdf gives the probability that trial n_1 trials have outcome X_1, n_2 trials have outcome X_2, ..., n_M trials have outcome X_M, where $\sum_{i=1}^{M} n_i = n$ and $\sum_{i=1}^{M} p_i = 1$. The probability density function is

$$f(n_1, n_2, \ldots, n_M) = \frac{n!}{n_1! n_2! \cdots n_M!} p_1^{n_1} p_2^{n_2} \cdots p_M^{n_M}. \tag{4.5}$$

4. Poisson Distribution. A Poisson random variable can take on integer values, and is often used to model a counting process. Qualitatively, the conditions required for a Poisson distribution to hold are:

   (a) The number of occurrences within one time period (or area) is independent of the number of occurrences in a non-overlapping time period (or area).

   (b) The expected number of occurrences within one time period (or area) is constant over all time periods (or areas).

   (c) The expected number of occurrences within one time period (or area) decreases to 0 as the length of the time period (or size of the area) decreases to 0.

   The density is:

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, ... \tag{4.6}$$

   which gives the probability that exactly x events will occur in a given time interval (or area). The parameter λ gives the average (or expected) number of events that will occur during that time interval (or area).

   A Poisson random variable can be used, for example, to model the number of exotic species that enter a country over time interval. Suppose, on average, 10 exotic species arrive per year: $\lambda = 10$, and we wish to know the probability that exactly 12 species will arrive. We

have $f(12) = \frac{10^{12}e^{-10}}{12!} = 0.095$ or about 10%. Poisson random variables are heavily used in queueing theory, which has some very interesting ecological applications.

The sum of two or more Poisson random variables is itself a Poisson random variable, even if each of the λ's is different. The parameter for the new distribution is given by the sum of the individual λ's.

5. Geometric Distribution. The geometric random variable is also closely related to independent Bernoulli trials in the following way: The geometric random variable is the number of trials needed to get the first "success". If the probability of success of any given trial is p, then the probability that the first success is achieved on trial x is given by:

$$f(x) = p(1-p)^{x-1} \tag{4.7}$$

This formula should make sense because the trials are independent and there are $x-1$ failures, followed by one success. In the grizzly bear example, what is the probability that we will have to sample 4 males before we observe the first female? The answer is: $f(5) = 0.51(0.49)^4 = 0.029$ or about 3%.

Continuous Random Variables

1. Uniform Distribution. The uniform distribution is both a discrete and a continuous distribution. It is usually what lay people mean when they say "random", that is, it is the only distribution for which all possible outcomes are equally likely. The pdf gives the relative frequency of observing the value x, and is given as follows:

$$f(x) = \frac{1}{b-a}, \quad a \leq x \leq b \tag{4.8}$$

2. Normal Distribution. This is probably the most widely used distribution, and also goes by the name "Gaussian Distribution". The distribution is symmetric, so that the probability of occurrence of an event u units below the mean is equal to the probability of occurrence of an event u units above the mean. It is bell-shaped, so events near the mean are very likely, while events far from the mean are very unlikely. Another interesting property is that it is unbounded, so the probability

of observing an arbitrarily large or small number is non-zero. The pdf is given as follows:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right), \quad -\infty < x < \infty \tag{4.9}$$

where the parameters σ and μ have the very important interpretations of the standard deviation and mean of the distribution, respectively. We usually denote a normal random variable with mean μ and standard deviation σ as $N(\mu, \sigma)$. Some extremely important properties of the normal random variables follow:

- The sum of several Normal Random Variables is itself a Normal Random Variable.
- The average of several Normal Random Variables is itself a Normal Random Variable.
- The Central Limit Theorem states that when you add or average many independent random variables, regardless of their distribution, the resulting distribution is approximately normal (as the number of r.v.s becomes infinite, the approximation becomes exact). Consider the following experiment: A random variable is constructed as the sum of 20 uniformly distributed random variables, on the interval [0,1]. The experiment proceeds as follows:
  (a) 10 realizations from the uniform distribution are drawn, and added together,
  (b) This process is repeated N times,
  (c) A histogram of the N realizations of the newly constructed random variable is generated. This histogram should look, more or less, like a normal distribution. The more samples we take, the closer it should look.

  This dramatic and surprising result is depicted in figure 4.2 for $N = 100$ realizations, and for $N = 100000$ realizations.

The normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$ is called the Standard Normal and is denoted $N(0, 1)$. Any normally distributed random variable with mean μ and standard deviation σ

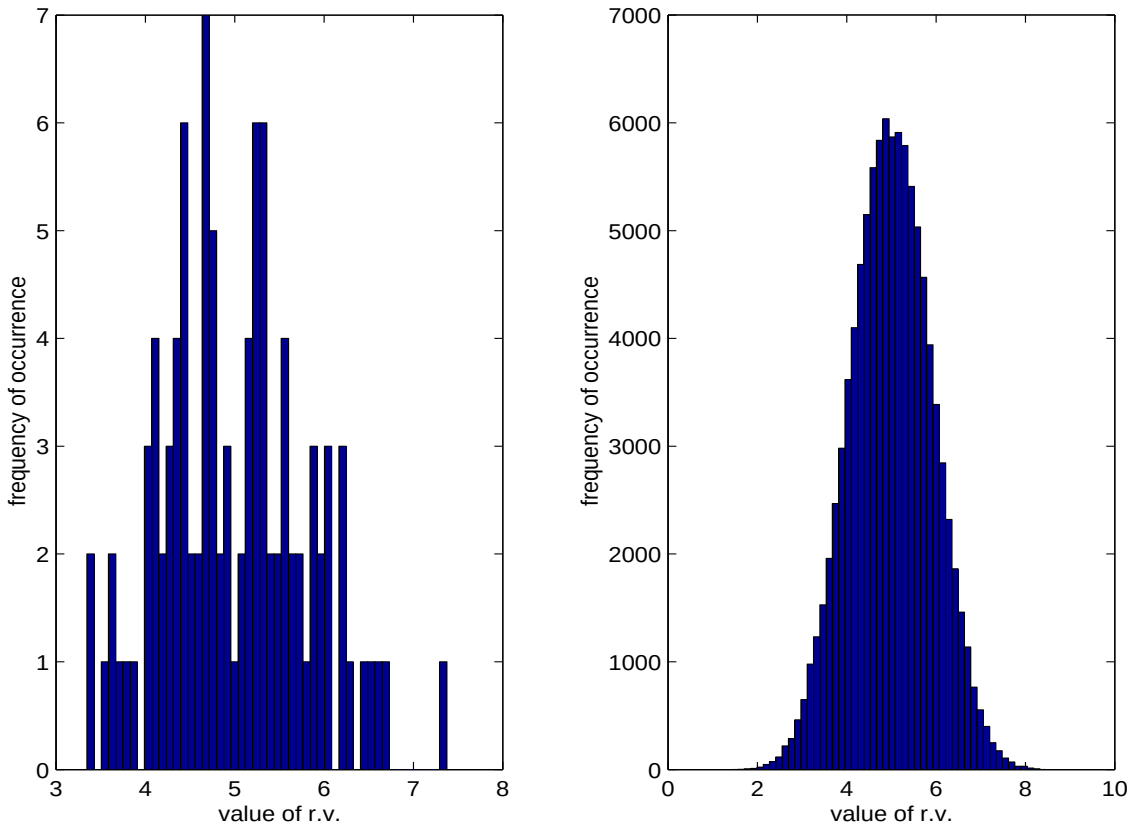


Figure 4.2: The mean of 10 random variables, each drawn from a uniform distribution on [0, 10]. (a) 100 realizations. (b) 100,000 realizations.

can be transformed back to the standard normal distribution with the following transformation, called the "Z transformation":

$$Z = \frac{X - \mu}{\sigma} \tag{4.10}$$

That is, Z is a $N(0, 1)$ random variable. You may be asking yourself how this information is useful. With a continuous random variable, we cannot ask questions such as "What is the probability that the realization of the random variable will be x?" because the sample space has an infinite number of elements. We can, however, ask questions like "What is the probability that the realization of the random variable will be above (or below) x?" Similarly, we can ask, "What is the probability that the realization of the random variable lies between x_1 and x_2?" Since the pdf (and hence the cdf) for the normal distribution is so hairy, most statistics texts print tables for the standard normal in the back.

For example, suppose a random variable X is distributed according to the standard normal distribution. We want to know the probability that the random variable takes on a value of ≥ 1.56. From the table on page 297 in Manly, the probability of the random variable having a value between 0 and 1.56 is 0.441. We know the probability of having a value of less than 0 is 0.50. Therefore, the probability of having a value of ≥ 1.56 is 1-.5-.441=0.059, or about 6%.

3. Log-normal Distribution. A log-normal random variable is a variable whose logarithm is normally distributed. That is, if $Y = \log(X)$, where Y is a normally distributed r.v., then X is a log-normally distributed r.v. Rearranging this equation, $X = e^Y$, where Y is normally distributed. Since Y can range from $(-\infty, \infty)$, X can range from $(e^{-\infty}, e^{\infty})$, or from $(0, \infty)$. The log-normal distribution has two parameters, μ and σ, which refer to the mean and standard deviation of the associated normal distribution, respectively (these are *not* the mean and variance of the log-normal distribution itself). Since log-normal random variables are bounded by 0 and are closely related to the normal distribution, they are often used to describe the errors or shocks in environmental data. The pdf for the log-normal distribution

is given below:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\log(x)-\mu}{\sigma}\right)^2\right), \quad x > 0 \tag{4.11}$$

One application is in population modeling, where, under certain assumptions, random environmental variability can be treated as a log-normal random variable. Suppose[1] a population of initial size N_0 will grow, and in particular, the population at some later date, t, call it N_t, to be the product of N_0 and the daily survival probabilities $s_0, s_1, s_2, ...$, where s_i is the probability that an individual survives from day i to day $i+1$. The population at time t is: $N_t = N_0 s_0 s_1 ... s_{t-1}$. Taking the natural log of both sides gives: $\log(N_t) = \log(N_0) + \log(s_0) + \log(s_1) + ... + \log(s_{t-1})$. It is likely that the daily survival probabilities are random variables. Remember that the sum of a bunch of random variables approaches the normal distribution (by the Central Limit Theorem), so we can construct a random variable $Y = \log(s_0)+\log(s_1)+...+\log(s_{t-1})$, that is normally distributed. So we have $log(N_t) = log(N_0) + Y$. Rearranging gives: $N(t) = \exp(\log(N_0) + Y) = e^{\log(N_0)} e^Y = N_0 e^Y$. Letting the random variable $X = e^Y$, where Y is normally distributed, gives $N(t) = N_0 X$, where X is log-normally distributed. This example shows that the log-normal distribution is appropriate even with limited knowledge about the original distribution of the random survival probability distributions.

4. Gamma Distribution. This distribution can take only non-negative values, and is often used in hydrology. Although there is a three parameter version, we'll stick to the two parameter version of this distribution here. The pdf for a gamma random variable, X, is:

$$f(x) = \frac{a^n}{\Gamma(n)} e^{-ax} x^{n-1} \tag{4.12}$$

where the mean is $\frac{n}{a}$ and $\Gamma(n)$ is a function that can be though of as a normalizing constant (if n is an integer, $\Gamma(n) = (n-1)!$). See any statistics text for more detailed coverage. See The Ecological Detective for a lengthy ecological application of the gamma distribution.

[1] From The Ecological Detective.

5. Extreme Value Distribution. Very useful in hydrology, ecology, and other environmental sciences, the extreme value distribution is used in modeling maxima and minima of a process or series of measurements. Examples include the maximum daily concentration of some pollutant, the minimum annual temperature, and the like. Although different extreme value distributions exist, perhaps the most common is the Gumbel Distribution, which is described here. The pdf for the Gumbel Distribution is given below:

$$f(x) = \frac{1}{\theta}\exp(-(x-\eta)/\theta)\exp(-e^{-(x-\eta)/\theta}), \tag{4.13}$$

$$\infty < x < \infty, \infty < \eta < \infty, \theta > 0 \tag{4.14}$$

The basic idea behind the extreme value distribution is that this is the limiting distribution (as the number of samples approaches infinity), of the maximum (or minimum) value of many draws from a continuous distribution. For example, consider the following example: Suppose temperatures on a South Pacific island, are normally distributed with mean 65°, and standard deviation 10°; $N(65, 10)$. Weather doesn't change much from season to season, so we'll treat each day as a random draw. A temperature reading is taken every day. What is the probability that the maximum temperature next year will exceed 103°? Here's an algorithm for thinking about this question:

(a) Take 365 draws from a $N(65, 10)$ distribution.

(b) Record the maximum temperature for that year.

(c) Repeat this process 10000 times.

(d) Plot a histogram of the 10000 maxima.

(e) The area of the histogram above 103° is the answer to the question. See the figure.

We need not have conducted this experiment if we just realized that the resulting distribution was an extreme value distribution.

6. Exponential Distribution. This distribution, is a continuous distribution that is often used to model inter-arrival times of events. The pdf is:

$$f(x) = \frac{1}{\mu}exp(\frac{-x}{\mu}), \quad x \geq 0 \tag{4.15}$$

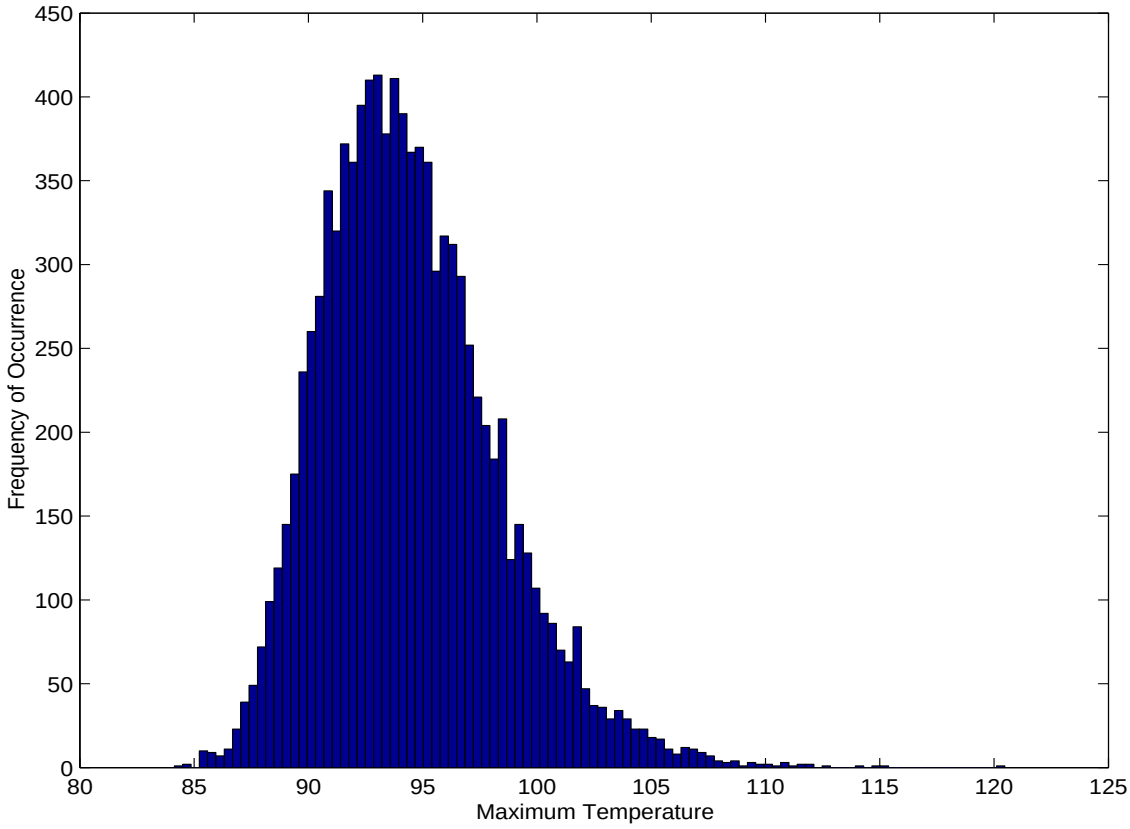


Figure 4.3: Distribution of maximum daily temperatures within a year, where daily temperature is drawn from $N(65, 10)$.

The single parameter, μ is the mean of the distribution. The exponential distribution has the unique property (among continuous r.v.'s) of being "memoryless"). That means that the probability that an event takes place in the next Δt time units is independent of the current time. This often has very important implications in modeling environmental phenomena.

This distribution is closely related to the Poisson distribution. Our Poisson random variable example had to do with the number of exotic species that arrived into a county during a given year. Suppose instead of the number that arrive by a certain time we are interested in the time between arrivals. The Poisson parameter λ gave the expected number of arrivals per year, therefore $\frac{1}{\lambda}$ gives the expected time between arrivals, which is the parameter of the exponential distribution (μ). Suppose the interarrival times of exotic species are, on average, 37 days; so species arrive about every 37 days. An exotic species arrived today. What is the probability that a new exotic will arrive in the next 10 days? Your book doesn't have a table for the exponential distribution, but you can look one up on the web, use S-Plus, or find a book that does. Conceptually, the answer is the area under the exponential pdf below 10, which is 0.237 or about 24%.

7. Chi-Squared Distribution. Like the log-normal, this distribution is very closely linked to the normal distribution. Suppose some control, X, is taken, and the response, Z is measured. The measured response is not precisely X, but it is $Z = X + Y$, where Y is normally distributed, $N(0, 1)$. The difference between the predicted response (X) and the observed response (Z) is $(Z - X)^2 = Y^2$. Here Y^2 is a chi-squared random variable. If we had n independent variables $(X_1, X_2, ..., X_n)$, and n responses $(Z_1, Z_2, .., Z_n)$, then the sum of the squared deviations is also a χ^2 random variable with n degrees of freedom. It is given the symbol χ^2_n. The pdf for the χ^2 distribution is quite complicated, and is not reproduced in these notes.

enditemize

4.5 Application of Technique

By now it should be fairly straightforward to answer our question above. The question boils down to: **What is the probability that we can observe a violation (≥ 30 ppb) from natural sources alone?** From looking at the data, it appears that only one or two out of 500 observations from the data exceed the limit of 30 ppb, so it is tempting to suggest that about 2 out of every 500 observations will exceed the threshold. However, we are given the true distribution: $N(20, 4)$. We can calculate the required probability in several equivalent ways:

1. We can find the area under the $N(20, 4)$ distribution above 30.
2. Equivalently, we can use the standard normal distribution, and find the area under the standard normal above the value $\frac{30-20}{4} = 2.5$.
3. Or, we can use the cdf for $N(20, 4)$ or the cdf for the standard normal.

The second approach is probably the most straightforward. Simply look up the table in Manly for $z = 2.5$. The given value is 0.494 which is the probability of z being between 0 and 2.5. We want the probability that $z > 2.5$. We know the $P(z < 0) = .5$, so the answer is 1-.5-.494=.006 or about 6 in every thousand. Figure 4.4 shows the pdf and cdf for both $N(20, 4)$ and $N(0, 1)$. You should conceptually be able to answer this question (i.e. by pointing to the answer) using any one of these graphs.

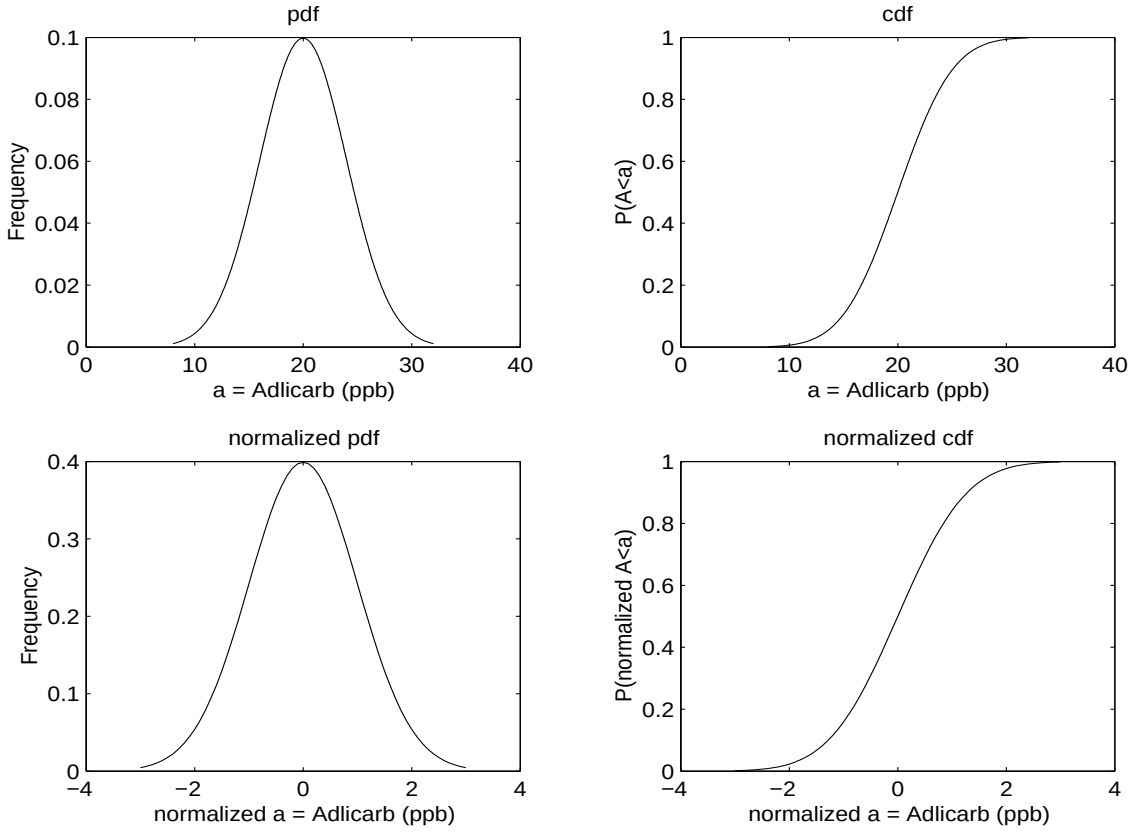


Figure 4.4: Pdf and cdf of the the actual and normalized aldicarb concentration.

Chapter 5

Hypothesis Testing

5.1 Motivation

Often we are interested in whether a parameter is different from a specified value, or whether the parameters from two separate samples differ from one another. The statistical procedure for this is the *hypothesis test*. This is set up to *reject* hypotheses, not accept them, so we set up a null hypothesis (typically that there is no effect, or that regulatory actions have not been triggered) and determine how likely we would be to generate our data if the null hypothesis were true. We then decide whether we can reject the null hypothesis; if so, then we conclude that the alternate hypothesis (which is the complement of the null hypothesis) is true. The general procedure is as follows:

1. Decide on a *null hypothesis* to be tested.

2. Decide whether the alternative to the null hypothesis is that there is *any* difference (*two–sided*) or whether the difference has to be in a particular direction (*one–sided*).

3. Choose a suitable test statistic which measures the extent to which the data are consistent with the null hypothesis.

4. Determine the distribution the test statistic would take on if the null hypothesis were true (e.g., the t-distribution).

5. Calculate the test statistic, S, for the observed data.

6. Calculate the probability P (also called the P–value) of obtaining a value as extreme as, or more extreme than, S if the null hypothesis were true.

7. If P is small enough, conclude that the null hypothesis is not true.

5.2 Examples & Questions

Does the mileage of "small" cars (e.g., Ford Escort, Honda Civic) differ from the mileage of "compact" cars (e.g., Ford Tempo, Honda Accord)?

5.3 Evidence or Data

The dataset "exfuel" contains gas mileage for 13 small and 15 compact cars. We looked at these data when examining correlations.

5.4 Technique

5.4.1 The null hypothesis

The null hypothesis, written H_0, is a statement about a true underlying parameter. It is usually a statement of no pattern or of meeting some criterion:

- A particular regression parameter is zero; that is, the independent variable has no effect on the dependent variable (H_0: $\beta_1 = 0$).

- Two samples, X and Y, have the same mean (H_0: $\mu_X = \mu_Y$).

- The mean of a sample of contaminant measurements does not exceed some critical threshold θ (H_0: $\mu \leq \theta$).

Typically we are not interested in the null hypothesis itself, but in the alternative hypothesis, which is its complement. Thus in practice we often formulate the alternative hypothesis first.

5.4.2 The alternate hypothesis

The alternate hypothesis, usually denoted H_A (sometimes H_1) typically says that there is some effect, or that the regulatory threshold has been exceeded. The null hypothesis and the alternate hypothesis *must be complements*: if one is true, the other must be false, and vice versa. The alternate hypotheses that go with the null hypotheses from the previous section are:

- H_A: $\beta_1 \neq 0$.
- H_A: $\mu_X \neq \mu_Y$.
- H_A: $\mu > \theta$.

One–sided vs. two–sided

The first two examples above are two–sided tests — you are interested in whether there is an effect, regardless of the direction of the effect. The third is a one–sided test — you are only interested in whether the sample mean exceeds the threshold (if it's less than the threshold, that's great... but it doesn't trigger regulatory action). It is easier to detect a one–tailed effect, for you are only looking at one tail of the distribution (this will become clearer in the examples). However, you should only invoke a one–tailed test *before* you analyze the data, and should be based either on only having an interest in one direction or because the scientific processes give you a reasonable expectation that the effect should be in a particular direction. Some programs (including Excel) print results for both one–sided and two–sided tests — do not succumb to the temptation to decide which one to use after seeing the result.

5.4.3 The t–test

It is easiest to explain the next few steps with reference to a particular test. We will use the t–test, which applies to all three of the examples given in the earlier sections. We will discuss other tests in the next lecture.

The one–sample t–test

In the third case we have a single sample, and we want to compare its mean to a specified value, θ. The test statistic is the difference between θ and $\bar{X}$,

scaled by the standard error of $\bar{X}$:

$$t_X = \frac{\bar{X} - \theta}{\mathrm{SE}(\bar{X})}. \tag{5.1}$$

If the true mean of the population were θ, then t_X would be t–distributed with degrees of freedom (df, sometimes written ν) $n - 1$, where n is the number of data points in the sample. Note that t_X can have either sign, depending on whether $\bar{X}$ is less than or greater than θ. We then compare t_X to the appropriate t–distribution to obtain the P–value. For the one–sided case here,

$$P = \Pr[t_{[n-1]} > t_X]. \tag{5.2}$$

If we wanted to test in the opposite direction, then

$$P = \Pr[-t_{[n-1]} < t_X]. \tag{5.3}$$

Finally, if we were doing a two–sided test, we would test against the absolute value of t_X (since the t–distribution is symmetric):

$$P = 2\Pr[t_{[n-1]} > |t_X|]. \tag{5.4}$$

The two–sample t–test

You want to test whether the means of two samples X and Y of different populations are different. Once again we will calculate a t-statistic, but the formula differs depending on whether the population variances are equal. If they are, then

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\left[\frac{(n_X-1)\mathrm{var}(X)+(n_Y-1)\mathrm{var}(Y)}{n_X+n_Y-2}\right]\left(\frac{n_X+n_Y}{n_X n_Y}\right)}}, \tag{5.5}$$

where n_X and n_Y are the sample sizes of X and Y. This is compared with a t–distribution with $(n_X + n_Y - 2)$ degrees of freedom. If n_X and n_Y are large this can be approximated by

$$t \approx \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\mathrm{var}(X)}{n_X} + \frac{\mathrm{var}(Y)}{n_Y}}}. \tag{5.6}$$

If instead of testing the null hypothesis that the means are equal you want to test H_0: $\mu_X - \mu_Y = \theta$, simply replace the numerator of eqs. (5.5) and (5.6) with $(\bar{X} - \bar{Y} - \theta)$.

If the variances of the populations are unequal, then you calculate the t-statistic the same way, but there is a more complex formula for the degrees of freedom:

$$df = \frac{(s_{\bar{X}}^2 + s_{\bar{Y}}^2)^2}{\frac{s_{\bar{X}}^4}{n_X - 1} + \frac{s_{\bar{Y}}^4}{n_Y - 1}}, \tag{5.7}$$

where $s_{\bar{X}}$ is the standard error of $\bar{X}$. This is called the "Welch modified two-sample t–test."

Is there a statistical hypothesis test for whether the variances are equal? You bet! If the underlying population variance are the same, then the quantity $\mathrm{var}(X)/\mathrm{var}(Y)$ follows an F-distribution. The F-distribution has a separate number of degrees of freedom for each sample; they are $df_X = n_X - 1$ and $df_Y = n_Y - 1$.

Assumptions of the t–test

The t–test assumes that the observations in each sample are normally distributed. When using the t–test to look at parameters of a regression, the residuals should be normally distributed, as well as satisfying all the other assumptions of OLS regression.

5.4.4 Interpret the P–value

The P–value is the probability of observing your data if the null hypothesis is true. If you reject the null hypothesis, P tells you the probability that this is an error. *P is not the probability that the null hypothesis is true*: in reality it is either true or it is not.

The smaller P is, the more confidently you can reject the null hypothesis. In science, $P < 0.05$ is often used as a standard for "statistical significance," but this is an arbitrary choice. Other commonly used thresholds are 0.1, 0.01, and 0.001. Your choice should depend in large part on the costs of making a mistake (see next section). In any case, you should always report the P–value, so that the reader can draw his or her own conclusions.

In the real world, P is an approximation: there are always some minor violations of the test assumptions, and unless your sample size is huge then

you would get a different P–value if you repeated the sample. Thus you should not bother with too many significant digits when reporting P; a single non-zero digit (0.06, 0.00007, etc.) is sufficient.

5.4.5 Type–I and type–II error

When we decide whether to reject the null hypothesis, there are two kinds of error we might make. A *type–I error* occurs if the null hypothesis really is true but we reject it. A *type–II error* occurs if the null hypothesis really is false but we fail to reject it. The maximum acceptable error probabilities are denoted α and β. α is the critical value of P below which we reject the null hypothesis; β is controlled indirectly by α, the sample size, and the sample variance. The actual formula for β is test-specific.

> JARGON ALERT: Type–I error is also referred to as *false rejection error*; type–II error is also referred to as *false acceptance error*.

The quantity $(1-\beta)$ is known as the *power* of the test. The larger the true effect, the easier it is to detect it, so power has to be defined in the context of an effect size. For example, we might be willing to tolerate a 10% chance of accepting the null hypothesis that the contaminant level is not above the threshold if the true level of contamination is one unit above the threshold.

For a given sample size, there is a trade-off between the two types of errors: if we decrease α then β increases, and vice versa. If we want to decrease β while holding α constant, then we need to increase the sample size. Another way to increase the power is to reduce the sample variance. While we don't have control over the amount of natural variability, recall from the beginning of the course that we can often adjust the the sampling scheme to reduce the variance among observations.

The two kinds of error have different costs associated with them. For example, if we were testing whether a chemical concentration exceeded a contamination threshold, then a type I error would mean that we would force the company to clean up the site even though it is not contaminated; this would result in unnecessary expenditure. A type II error means that we let the company off the hook even though the site is contaminated; a non-monetary cost is that the environment is polluted, and monetary costs might include increases in health-care costs for people associated with the

```
Test Name:                  Welch Modified Two-Sample t-Test
Estimated Parameter(s):     mean of x = 31
                            mean of y = 24.13333
Data:                       x: Small in DS2 , and y: Compact in DS2
Test Statistic:             t = 5.905054
Test Statistic Parameter:   df = 16.98065
P-value:                    0.00001738092
95 % Confidence Interval:   LCL = 4.413064
                            UCL = 9.32027
```

Table 5.1: Splus output for a two–sample t–test applied to the mileage data.

site or publicly funded cleanup when the contamination is discovered in the future. Ideally, one would set α and β so that the expected costs of the two types of errors would be the same: $\alpha C_{\mathrm{I}} = \beta C_{\mathrm{II}}$, where C_{I} and C_{II} are the costs of a type I and type II error, respectively.

5.5 Application of Technique

5.5.1 Gas mileage

The means of the gas mileage for small and compact cars are 31 and 24 mpg, respectively. The variances are 14.5 and 3.5. The ratio of the variances is $14.5/3.5 = 4.1$. Comparing this with tabulated values of the F-distribution with 12 and 14 degrees of freedom shows that P is between 0.01 and 0.02. Thus it is highly unlikely that the variances truly are equal, and we use a t-test with unequal variances.

The output is shown in table 5.1. You would not show all this information in a report. Instead, you would report this in text as follows: The mean mileage of small cars differs from that of compact cars (Welch modified two–sample t–test: $t = 5.91$, $df = 17.0$, $P = 0.00002$).

The 95% confidence interval that Splus reports is for the difference in the means.

Chapter 6

ANOVA

6.1 Motivation

The t–test allows us to look for differences in the means of two groups. What if there are more than two? ANOVA is a generalization of the t–test that allows us to look simultaneously for differences in the means of several groups that may differ in more than one way.

6.2 Examples & Questions

We want to know whether trout can develop "resistance" to metals toxicity — that is, if they are first exposed to a low concentration of metals in the water, can they subsequently survive longer in a high concentration?

6.3 Evidence or Data

Data were collected on about 60 fish from three different groups: hatchery brown trout (*Salmo trutta*), hatchery rainbow trout (*Onchorynchus mykiss*), and brown trout from the Clark Fork River in Montana. Half of each group (the controls) was kept in clean water for three weeks; the other half were kept in water with low concentrations of a metals mixture that contaminates the Clark Fork River. All fish survived this period. Then all fish were placed in water with the metals contamination at the maximum concentration observed in the river, and time to death (in hours) was recorded.

6.4 Technique

6.4.1 Single-factor ANOVA

Suppose that you have observations of a dependent variable Y, and each observation is associated with an independent categorical variable X that has i different "levels" (possible values). For example, X could be control/treatment, student/faculty/staff, green/blue/red/yellow, etc. The ANOVA equation that represents this situation is:

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \tag{6.1}$$

where y_{ij} is the j'th observation of $\{y_i\}$, the group of obervations that has the ith level of X, μ is the overall mean of Y, and α_i is the deviation from that mean that comes from being in the ith level of X. The sum of all the coefficients α is zero.

Given a collection of data, our estimates of these parameters are

$$\hat{\mu} = \bar{y} \tag{6.2}$$

$$\hat{\alpha_i} = a_i = \bar{y_i} - \bar{y} \tag{6.3}$$

In the context of hypothesis testing, the null hypothesis is that X has no effect on Y. If there are I levels of X, this null hypothesis is

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_I. \tag{6.4}$$

Notice that if there are only two levels, then this is simply a t-test.

We now need to look at three components of the sums-of-squares. Let a_i be our estimate of α_i and $\bar{y}$ be our estimate of μ. The sum of squares among the coefficients a is the *among-group* sum of squares:

$$SSG = k \sum_{i=1}^{I} a_i^2. \tag{6.5}$$

The squared deviations between the data and the relevant group mean is the *within-group* sum of squares, and is equivalent to the residuals in a regression:

$$SSE = \sum_{i=1}^{I} \sum_{j=1}^{k_i} (y_{ij} - (a_i + \bar{y}))^2, \tag{6.6}$$

where k_i is the number of observations in group i. The *total* sum of squares is the sum of these:

$$SST = \sum_{i=1}^{I} \sum_{j=1}^{k_i} (y_{ij} - \bar{y})^2 \tag{6.7}$$

$$= SSG + SSE. \tag{6.8}$$

The degrees of freedom associated with SSG is $I - 1$, and if there are n total observations, the degrees of freedom associated with SSE is $n - I$, and with SST is $n - 1$. These are used to create the associated *mean squares*:

$$MSG = SSG/(I - 1) \tag{6.9}$$

$$MSE = SSE/(n - I) \tag{6.10}$$

$$MST = SST/(n - 1) \tag{6.11}$$

The F-test for model significance

Notice that MST is an estimate of the population variance of Y, σ_Y^2. If the null hypothesis is true, then MSE and MSG are *also* estimates of σ_Y^2. To see this, consider the special situation where there are equal numbers of observations in each group (all the k's are equal) — this is known as a *balanced design*. Then

$$MSE = \frac{1}{I(k-1)} \sum_{i=1}^{I} \sum_{j=1}^{k} (y_{ij} - (a_i + \bar{y}))^2 \tag{6.12}$$

$$= \frac{1}{I} \sum_{i=1}^{I} \left[\frac{1}{k-1} \sum_{j=1}^{k} (y_{ij} - (a_i + \bar{y}))^2 \right]. \tag{6.13}$$

Since $(a_i + \bar{y})$ is the sample mean within group i, each term within the brackets is an estimate of σ_Y^2 (with sample size k), and MSE is the average of these estimates, so itself is an estimate of the variance. Furthermore,

$$MSG = k \left[\frac{1}{I-1} \sum_{i=1}^{I} (a_i - \bar{y})^2 \right]. \tag{6.14}$$

The term inside the brackets is the variance of I independent estimates of the population mean, each based on k observations. The "variance" of an estimate is the same thing as the square of its standard error. Thus multiplying this by k will give us an estimate of the population variance.

Thus if the null hypothesis is true, then MSG and MSE are independent estimates of the *same* population variance, and so the ratio MSG/MSE will be F-distributed with $I-1$ and $n-I$ degrees of freedom. P is the area of the tail outside the value given by the ratio, and if it small, we can reject the null hypothesis that X has no effect on Y.

It is harder to visualize when the sample sizes in the various groups is unequal, but the same property holds there as well. It also holds for OLS regression. Thus the P associated with the F-test at the bottom of regression or ANOVA output is associated with the null hypothesis that the *model as a whole* does not explain any of the variation in the independent variable.

Multiple comparisons

If you reject the null hypotheses, all you can say is that the mean of Y is not the same across the various levels of X. Often you will also want to know which groups differ from each other. You might think you could do t-tests on all possible pairwise combinations ... and you would be right. But how do you interpret the P-values?

Suppose you have n levels in X. Then you have $m = n(n-1)/2$ unique pairwise comparisons. If you plan to say that two means are different if $P < \alpha$ from a t-test, then even if no two pairs of means are different, then you will have at least one $P < \alpha$ with probability $1-(1-\alpha)^m$. This rapidly increases as the number of levels increases (figure 6.1).

There are a number of ways of dealing with this, but the most straightforward is to use the *Bonferroni correction* to your critical α value. If you want an experiment-wise type I error of no more than α, then the critical value for each test (the comparison-wise type I error) should be set to

$$\alpha' = 1-(1-\alpha)^{1/m}. \tag{6.15}$$

As always, you should report the actual P-values.

6.4.2 Multi-factor ANOVA

ANOVA can also be employed when there are multiple independent variables. For example, with two independent variables, the model is

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, \tag{6.16}$$

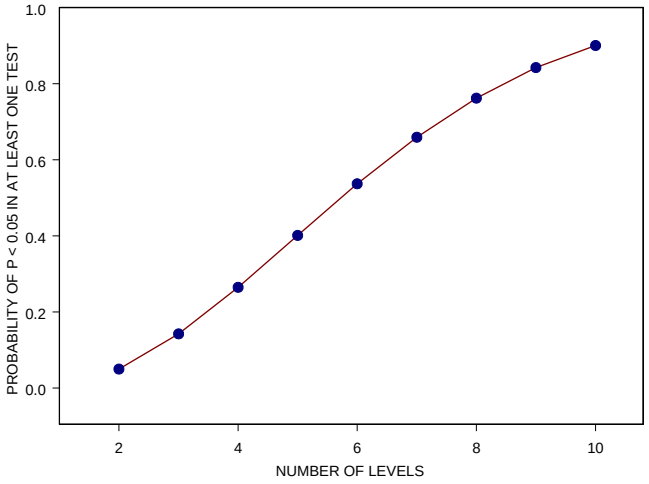


Figure 6.1: The probability of obtaining $P < 0.05$ in at least one test when performing multiple comparisons tests among n levels even if none of the means actually differ.

Table 6.1: ANOVA table for a two-factor model. There are I levels of factor A, J levels of factor B, and m observations for each combination of factor levels.

Source of variation	Sum of squares	*df*	Mean square	*F*
Factor A	SSA	$I-1$	$\text{MSA} = \text{SSA}/(I-1)$	MSA/MSE
Factor B	SSB	$J-1$	$\text{MSB} = \text{SSB}/(J-1)$	MSB/MSE
Interaction	SSAB	$(I-1)(J-1)$	$\text{MSAB} = \text{SSAB}/[(I-1)(J-1)]$	MSAB/MSE
Error	SSE	$IJ(m-1)$	$\text{MSE} = \text{SSE}/[IJ(m-1)]$	

where $(\alpha\beta)_{ij}$ is the *interaction term*, representing the additional effect of being in both level i of the first variable and level j of the second variable.

Once again we construct sums of squares and mean squares (table 6.1). There is now an F statistic (and an associated P-value) for each of the factors alone (often referred to as "main effects") and for the interaction term (in order to estimate the interaction term you need more than one observation in each (i, j) combination).

6.4.3 Assumptions of ANOVA

The assumptions of ANOVA are identical to those for OLS regression described in chapter 8. In particular, the population distributions should be normal, and have equal variances across all of the levels (homoscedasticity). If these are violated then the hypothesis tests may be incorrect. However, the results are robust to fairly substantial violations of these assumptions. Situations that are likely to cause problems are strong kurtosis combined with small sample size, and differences in variance when the sample size differs among groups.

6.5 Application of Technique

The two factors that we want to analyze are the species/source of the fish and whether they had the acclimation treatment. Unfortunately the variances of the the different groups vary tremendously, from 20 to 1200. After log-transforming the data the variances are much more consistent among groups.

Applying the S-plus ANOVA routine produces the output in table 6.2. The most important part is the table showing the sums of squares; a cleaned up version of this (suitable for a publication) is in table 6.3. These results show that the acclimation treatment did increase the survival time of trout; however the source of the fish also affected survival time. Furthermore, the significant interaction term indicates that the effect of acclimation differed among the three fish.

Notice that only some of the parameter estimates are reported. The others can be recovered by remembering that the coefficients within a level sum to zero. Thus if α_1 and α_2 refer to the treatment and control manipulations, respectively, and if β_1, β_2, and β_3 refer to the Clark Fork brown trout, the hatchery brown trout, and the hatchery rainbow trout respectively, then we

Table 6.2: S-plus output for the fish survival analysis.

```
*** Analysis of Variance Model ***

Short Output: Call:
   aov(formula = Log.Survival ~ Treatment + Fish + Treatment:Fish, data = Trout,
    na.action = na.exclude)

Terms:
                Treatment     Fish Treatment:Fish Residuals
 Sum of Squares  14.73256  4.33096        1.52487  23.43014
Deg. of Freedom         1        2              2       175

Residual standard error: 0.365905 Estimated effects may be
unbalanced

                Df Sum of Sq  Mean Sq  F Value       Pr(F)
     Treatment   1  14.73256 14.73256 110.0377 0.000000000
          Fish   2   4.33096  2.16548  16.1740 0.000000359
Treatment:Fish   2   1.52487  0.76244   5.6946 0.004017894
     Residuals 175  23.43014  0.13389

Estimated Coefficients:
 (Intercept) Treatment      Fish1     Fish2 TreatmentFish1 TreatmentFish2
    3.509212 0.2811782 0.07003326 0.0999444    -0.07261913     0.04968512
```

Table 6.3: Effects of fish (hatchery brown trout, hatchery rainbow trout, and Clark Fork River brown trout) and acclimation treatment (treatment/control) on the mean log survival time of trout exposed to toxic metals.

	df	Sum of Sq	Mean Sq	*F*	*P*
Treatment	1	14.73	14.73	110.04	$<< 0.0001$
Fish	2	4.33	2.17	16.17	$<< 0.0001$
Treatment×Fish	2	1.52	0.76	5.69	0.004
Residuals	175	23.43	0.13		

Table 6.4: Parameter estimates for the ANOVA results shown in table 6.3

Parameter	Calculated as	Value
$\hat{\mu}$		3.509
a_1		0.281
a_2	$-a_1$	−0.281
b_1		0.070
b_2		0.100
b_3	$-(b_1 + b_2)$	−0.170
$(ab)_{11}$		−0.073
$(ab)_{12}$		0.050
$(ab)_{13}$	$-((ab)_{11} + (ab)_{12})$	0.023
$(ab)_{21}$	$-(ab)_{11}$	0.073
$(ab)_{22}$	$-(ab)_{12}$	−0.050
$(ab)_{23}$	$-(ab)_{13}$	−0.032

can read the estimates of μ, α_1, β_1, β_2, $(\alpha\beta)_{11}$, and $(\alpha\beta)_{12}$ directly from the output. The others have to be calculated as shown in table 6.4.

The results of a single factor ANOVA can be reported in the text: "the means were significantly different (single-factor ANOVA: F = X, df = Y, P = Z)." However, the results of a multi-factor ANOVA are shown in a table structured like table 6.3.

Chapter 7

Environmental Sampling

7.1 Motivation

When designing a new study to address an environmental problem, you will first have to have identified what statistical quantity is of interest (e.g., the slope of the relationship between automobile weight and mileage, or the difference in mean chemical concentration between a cleanup and a reference site). Given this objective, you now want to collect data that will give you an estimate of the parameter that is precise enough to answer the question at hand. If the estimate is imprecise, you may not be able to say that there is any effect at all. Conversely, the higher your precision, the smaller the effect that you can detect.

7.2 Example & Question

In the TcCB example from chapter 2, we compared the cleanup site to a single reference site. However, there might be variation among potential reference sites. Today's problem is to design a sampling scheme to estimate the mean concentrations at 10 different reference sites. Anticipating that the differences may be small, we want the 95% confidence interval at each site to be no more than 0.1 ppb; that is, the standard error of the mean should be about 0.025.

7.3 Evidence or Data

We have the 47 observations of TcCB concentration at a single reference site.

7.4 Technique

A statistical *population* is the entire collection of potential measurements about which we want to make a statement (which may be infinitely large). This is the first step at which refining of the overall question must be done. For example, if the question is "What is the concentration of dissolved oxygen in this stream?" we might refine it to "What is the average concentration of dissolved oxygen in a particular section of the stream at a depth of 0.5 m over a particular 3–day period?" The population is the set of all possible measures of dissolved oxygen in that section of stream at 0.5 m depth over that 3–day period.

A *sample* is some subset of the population; if it contains all the elements of the population, it is called a *census*. Each *sample unit* generates a single observation.

The first step is to determine the desired precision. This may be set by an agency; if not, decide what the smallest effect size that is "meaningful", and set the precision so that you have a reasonable chance of detecting such an effect if it exists.

We have already seen that the precision of an estimate of the mean increases with sample size and decreases with the sample variance (the formulas for standard errors of other statistics are more complex than that of the mean, but all follow this general principle). In order to assess the precision of the estimate at all, the observations must be a *random sample* of the population. Thus, having defined the population, there are three questions to be addressed:

1. What should I use as my sample unit?

2. How many sample units should I examine?

3. How should I select my sample units?

7.4.1 What kind of sample units?

The objective here is to design the sample units so as to reduce the variance among observations. If there is some sort of small-scale heterogeneity in the environment that is incidental to the question at hand, then sample units can be made large enough so that the heterogeneity is encompassed *within* each observation, rather than being expressed as variation *among* observations. For example, we might expect that the dissolved oxygen levels in the stream might be very different in riffles and pools. If so, then point samples will have a large variance reflecting this heterogeneity. In contrast, if the sample unit were large enough that it would typically encompass both a riffle and a pool (say, by taking the average of 10 point measurements over a 20 m stretch of stream), then the variability among observations will be much lower.

An important way to enlarge your sample unit is *composite sampling*, which is especially valuable when the cost of lab analysis is high relative to the cost of collecting the specimen in the field. To do this you collect several physical samples from a previously defined area, mix them up, and analyze the mix as a single observation. Thus you are physically averaging several spots within your sample unit.

However, larger sample units can be more expensive to measure, and more prone to *observation error* (imagine counting all oak trees in a hectare vs. in a square kilometer). Given a fixed budget, as the size of the sample unit increases, the number of observations will typically need to decrease.

7.4.2 How many sample units?

As many as possible! Not really, but it is rare that you will have enough time and money to collect as many observations as you would like. There are diminishing returns, however, as precision varies with the square root of sample size. Thus if you do have an unlimited budget, then you should collect just enough data to achieve the desired level of precision.

However, since the precision also depends on the sample variance, how can we predict precision ahead of time? The solution is to collect a *preliminary sample* of relatively small size from which to get a preliminary estimate of the sample variance. As few as 10 or 15 observations may give you a reasonable estimate, although if the environment is heterogeneous or the data are strongly skewed then it will be an underestimate (only three of the 77 observations of TcCB at the cleanup site had values greater than 10; a

sample size of 15 would have a good chance of not including any of them). Then use this variance estimate to calculate the sample size needed for the desired precision. For example, if you want the standard error of the mean to be a and preliminary variance estimate is b, then then we need to solve the equation

$$a = \frac{\sqrt{b}}{\sqrt{n}} \tag{7.1}$$

for n, giving

$$n = \frac{b}{a^2}. \tag{7.2}$$

Once you begin the full-scale sampling program, then you can use the new data to revise your variance estimates and adaptively increase the sample size, if necessary.

Other useful formulas, again assuming that you have an initial estimate b of the sample variance:

1. If you want the 95% confidence interval of the mean to be $\bar{x} \pm \delta$, then the required sample size is approximately

$$n \approx \frac{4b}{\delta^2}. \tag{7.3}$$

   The justification for this should be clear to you!

2. Suppose you are taking two samples of size n from two populations that may have different means, but have the same variances. To have a 95% chance of rejecting the null hypothesis that the population means are the same when in fact they differ by δ, you need a sample size of approximately

$$n \approx \frac{8b}{\delta^2}. \tag{7.4}$$

Notice that since the precision depends on the square root of sample size, there are diminishing returns (recall that each observation costs time and money). Thus if you can put a value on precision you can calculate the "optimal" sampling effort.

7.4.3 Selecting sample units

All statistical analyses assume that the sample units have been selected at *random* — each sample unit in the population has an equal chance of being selected, and the prior selection of one unit doesn't affect the chance that another is selected. In practice this entails either

1. enumerating all the units in the population, and using random numbers to select the ones to measure; or
2. if the population is infinite (all spatial locations, for example), then use random numbers to select the coordinates of the sample locations.

Stratified random sampling

Another way to deal with environmental heterogeneity is *stratified random sampling*. If the environment is into a number of easily identified types — called *strata* (for example, riffles and pools) — then you simply sample randomly within each stratum, calculate each stratum's statistic, and then combine them, taking into account the relative amount of each stratum in the environment. For example, suppose that there are K strata, with the j'th stratum covering fraction N_j of the population; you estimate the mean of stratum j to be $\bar{x}_j$. Then the mean of the entire sample is

$$\bar{x} = \sum_{j=1}^{K} N_j \bar{x}_j. \tag{7.5}$$

Similarly, the sample variance is the weighted mean of the individual stratum variances; if the strata really do affect the variable being measured, then the stratum variances can be much smaller than would be found if the entire population were sampled randomly. The standard error of the mean is the square root of

$$\mathrm{var}(\bar{x}) = \sum_{j=1}^{K} N_j^2 \mathrm{var}(\bar{x}_j). \tag{7.6}$$

There are two additional benefits to stratified random sampling. It allows you to ensure that you include rare strata that contribute disproportionably to the statistic, but that may well be missed by a random sample of reasonable size. An example of such a rare-but-important stratum might be zones of dead plants marking where major chemical leaks occurred.

Second, if the strata vary not only in their means but in their variances, you can adjust the sampling effort in each stratum, focusing most effort on the highly variable strata.

Systematic sampling

Systematic sampling, also called *uniform*, is selecting units on a regular grid (in space) or selecting every k'th individual from a finite population. This is a good approach as long as the sampling frequency doesn't match the frequency of some repeating pattern in the environment. For example, if you sample on a 1-mile grid in Iowa, you have a chance of all of your observations falling on roads, which would give you a very biased view of the Midwestern landscape!

As long as you do not have this kind of bias, uniform sampling turns out to have a higher precision than a corresponding random sample. However, it is not easy to correct for this, so the standard practice is to use the formulas for random samples, and treat them as conservative estimates.

It is usually easier to locate a systematically chosen site than a randomly chosen one; for some kinds of data, this is a substantial fraction of the time cost of sampling.

Haphazard sampling

A commonly used alternative is *haphazard* or *judgement* sampling. The former is capricious selection by the investigator, without conscious bias; the latter is an attempt to select "representative" or politically desirable units. Both are very much *not* random; even in haphazard sampling, there tend to be unconscious biases toward certain site characteristics or uniform sampling. Another form of haphazard sampling is *convenience* sampling, in which one selects the units that are easy to get to.

> ***There is no way to estimate the precision of haphazardly sampled data.***

7.5 Application of Technique

We use the existing data as our preliminary sample. This assumes that all of the sites have similar sample variances; the only way to determine this for

sure would be to take preliminary samples at each of the other sites.

Recall that the variance in TcCB concentrations is 0.08. Thus to get a standard error of 0.025 we need $n = 0.08/(0.025)^2 = 128$. This is the sample size per site, so over all 10 sites we need 1280 soil samples.

Suppose that it costs \$1 to collect a soil core, and \$10 to have it analyzed at the lab. This gives a cost of \$1408 per site. Now suppose that we composite the soil cores in groups of 8, and that we make sure that they are really well mixed (this costs \$5 per group). With sixteen groups per site, this now costs $\$128 + 16 * \$15 = \$368$ per site. If the composites are perfectly mixed, the precision of the estimated mean is the same as if we had sent all 128 cores to the lab.

In principle we could composite all 128 cores and just bear the cost of a single lab analysis. This might be required if the lab work was extremely expensive. However, the more physical samples you composite, the harder it is to ensure that they are well mixed. Compositing also loses information about the variability among physical samples, which would be important if you are interested in anything more than the mean. It also requires that you *assume* that the sample variance is the same in all of your sites as it was in the preliminary sample, rather than estimating it directly from each site.

7.6 Further Reading

Two excellent sources of sample design in an ecological context are:

- Elzinga, C.L., D.W. Salzer, J.W. Willoughby, and J.P. Gibbs. 2002. *Monitoring plant and animal populations.* Blackwell Science, Malden, NY.

- Sutherland, W. J., ed. 1996. *Ecological Census Techniques: A Handbook.* Cambridge University Press, Cambridge, UK.

Chapter 8

Linear Regression

8.1 Motivation

Linear regression is probably the most widely used, and useful, statistical technique for solving environmental problems. Linear regression models are extremely powerful, and have the power to empirically tease out very complicated relationships between variables. Generally speaking, the technique is useful, among other applications, in helping explain observations of a dependent variable, usually denoted y, with observed values of one or more independent variables, usually denoted x_1, x_2, ... A key feature of all regression models is the error term, which is included to capture sources of error that are not captured by other variables. Linear regression models have been heavily studied, and are very well-understood. They are only appropriate under certain assumptions, and they are often misused, even in published journal articles. These notes are intended to provide you with a broad overview of linear regression, but are not intended to exhaust all details.

> JARGON ALERT: The dependent variable is also referred to as the *explained* variable, the *response* variable, the *predicted* variable, or the *regressand*. The independent variable is also referred to as the *explanatory* variable, the *control* variable, the *predictor* variable, or the *regressor*.

The basic principle of linear regression can be illustrated with a very sim-

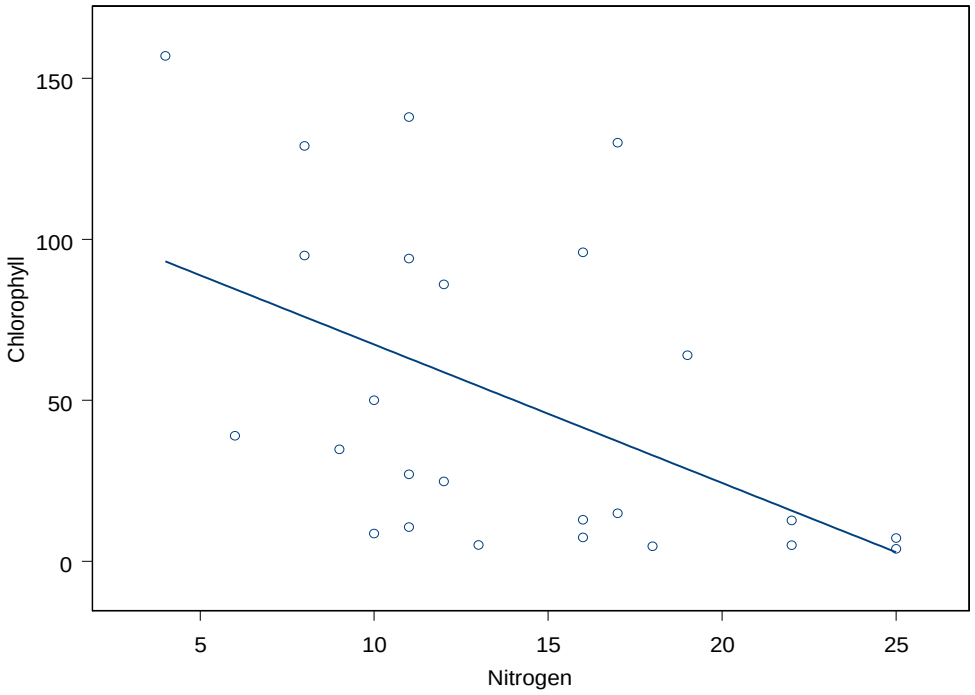


Figure 8.1: Chlorophyll-A vs. nitrogen concentration, with a fitted regression line.

ple example. [Example 3.1 in Manly]. Chlorophyll-a (denoted C) is a widely used indicator of lake water quality. High concentrations of Chlorophyll-a are associated with eutrophication, which is affected by the level of nitrogen in the water. We are specifically interested in the effect an increase in nitrogen would have on the Chlorophyll-a in a lake. Using Manly's data (in the course web - called Chlorophyll.xls), the plot of Chlorophyll-a vs. Nitrogen, with a fitted linear regression line, is given in figure 8.1. But something is wrong here, the figure suggests that more nitrogen leads to lower Chlorophyll-a, which runs counter to our intuition. Perhaps we are omitting an important variable that might help explain the observed level of Chlorophyll-a in a lake.

In fact, high phosphorus (P) *and* high nitrogen (N) levels are associated

with high Chlorophyll-a. Therefore, both variables must be included in the regression, even if we are only interested in the effect of N on Chlorophyll-a. Manly uses the following linear model to represent the relationship between C, P, and N:

$$C_i = \beta_0 + \beta_1 P_i + \beta_2 N_i + \epsilon_i \tag{8.1}$$

where C is chlorophyll-a, N is nitrogen, P is phosphorous, β_0, β_1, and β_2 are parameters that are unknown (to be estimated from the data), and ϵ is an error term. The error term will pick up an variation in the data that is unexplained by P and N. Before we estimate the parameters of this equation, let's think a little about it; here are some observations and questions:

- Suppose $P = 0$ and $N = 0$, what would we expect the level of C to be?
- What should be the sign of β_1 and β_2?
- Is our linear specification appropriate?
- What justification do we have for an additive error term?

Given that we believe the linear model we wrote above, how can we estimate the unknown parameters, β_0, β_1, and β_2? That's the purpose of this chapter. We want to choose the β's to create the "best fit" of this model to our data. The estimated parameters that accomplish this are: $\hat{\beta}_0 = -9.386$, $\hat{\beta}_1 = 0.333$, and $\hat{\beta}_2 = 1.200$. The $\hat{}$ indicates that the thing is an estimated parameter. We haven't said anything about how these numbers were chosen; that is the topic of the next chapter.

Suppose someone is thinking of developing a golf course near a lake, which is expected to increase the concentration of nitrogen in the lake by a small margin. By how much can we expect Chlorophyll-a to increase per unit increase in nitrogen? The answer, is simply our estimate of β_2, which is 1.2; a one unit increase in nitrogen leads to about a 1.2 unit increase in Chlorophyll-a.

8.2 Example & Question

U.S. Gasoline Market: It seems plausible that the gas price has some effect on gas consumption. From a policy perspective, the price elasticity of demand could be extremely important, when, say, the government is contemplating increasing the gasoline tax. Our question is: how responsive is

gas consumption to gas price? We might also be interested in how income or other characteristics affect gas consumption.

8.3 Evidence or Data

The first panel in figure 8.2 shows the per capita gas consumption in the U.S. through time. But this doesn't give us any information about how responsive gas consumption is to gas price. Your first instinct might be the following: Since we wish to know how gas price affects gas consumption, let's just use our data (from the Economic Report of the President, and posted in the course shared file called gasmarket.xls) and plot gas price vs. gas consumption [1]. The problem with this approach is that we have excluded many other variables that are likely to affect the consumption of gasoline. In fact, we are likely to get very misleading results using this approach (see figure that shows a positive relationship between gas price and gas consumption - oops!).

What other variables are likely to affect gas consumption (besides price)? Income, price of new cars, and price of used cars are probably appropriate. Figure 8.3 gives plots of all four independent variables (price, income, price of new cars, price of used cars) through time. The remainder of these notes is devoted to discussing, in a general way, how these types of data can be combined to answer questions such as the one posed above.

Note that the data presented here differ in units from the data described in class. The qualitative patterns are the same, but the estimated coefficients differ.

8.4 Technique

In this section, we discuss the Ordinary Least Squares (OLS) estimator within the context of the Classical Linear Regression Model. But first, let's define what we mean by an "estimator":

Definition 8.4.1
An estimator is a rule or strategy for using data to estimate an unknown

[1] The astute student will note that we have conformed to the way demand curves are typically drawn. This orientation of the axes is misleading, since gas price actually affects gas consumption, not the other way around.

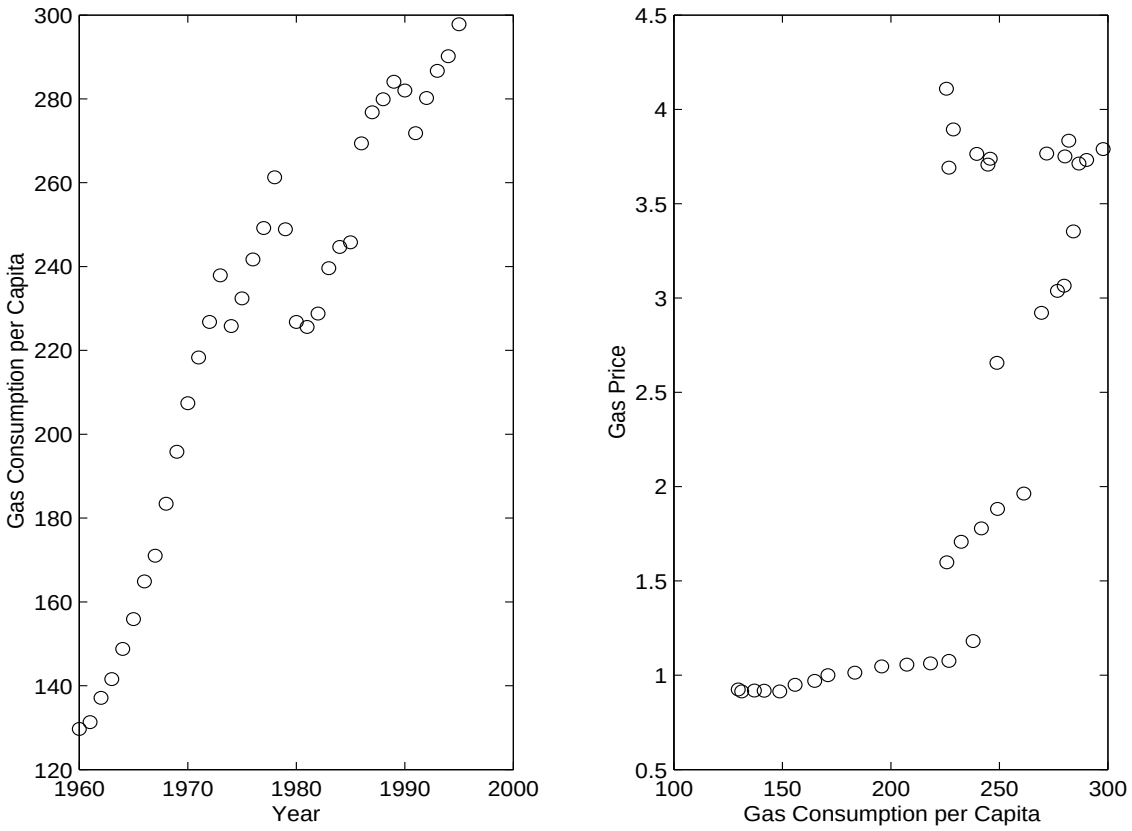


Figure 8.2: Gas consumption through time (left) and as related to gas price (right).

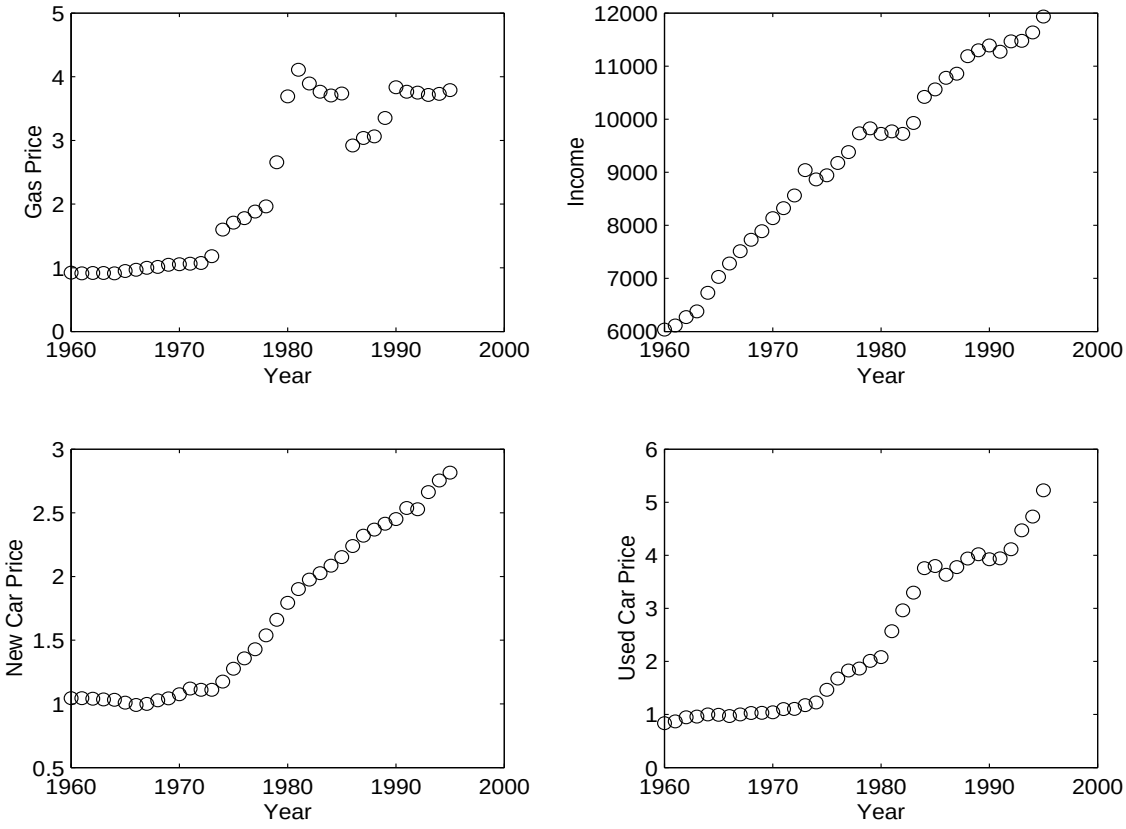


Figure 8.3: Gas price, income, price of new cars, and price of old cars throuh time.

parameter, and is defined before the data are drawn[2].

It should be clear that some estimators are better than others. We won't go into a great deal of detail here about different estimators, but you should know what an estimator is, and that it is just one way to estimate an unknown parameter; other ways exist.

Whenever we want to use data to parameterize a model (such as the Chlorophyll-a model or the U.S. gasoline market model), we must choose an estimator with which to choose the "best" parameters for that model. It turns out that one such estimator, the OLS estimator, is widely (though not universally) applicable. Most of the discussion in this section is devoted to a discussion of the assumptions under which the OLS estimator is valid.

8.4.1 What is "Linear" Regression?

What do we mean when we say "linear regression". Regression refers to the fact that although observed data are variable, they tend to "regress" towards their mean. Linear refers to the type of equation we use in our models. To use the OLS estimator, the model must be linear in parameters. The standard regression equation, $Y = \alpha + \beta X + \epsilon$ is linear in the parameters α and β. The regression equation $Y = \gamma \frac{X^2}{log(X)} + \theta Z^3 + \epsilon$ is also linear in parameters (γ and θ), and could be estimated using the OLS estimator. What we mean by "could be estimated by OLS" is that we can use the technique of ordinary least squares regression to find the "best" parameters to achieve a certain criterion.

On the other hand, suppose we are modeling fish stock dynamics, and we wish to use the "Ricker Model". The Ricker Model is given below:

$$R_{t+1} = S_t e^{\phi(1-S_t)} \epsilon_t \tag{8.2}$$

where R is the number of fish "recruits", S is the number of spawners and ϕ is a parameter to be estimated. This model is NOT linear in the parameter, ϕ, so it cannot be estimated using the OLS estimator. However, suppose we take the log of both sides:

$$log(R_{t+1}) = log(S_t) + \phi(1 - S_t) + log(\epsilon) \tag{8.3}$$

Now we have a model that is linear in parameters, so we can estimate the model using OLS.

[2]Adapted from Greene. Econometric Analysis

8.4.2 Assumptions for CLRM

In order to use the OLS estimator, the following five basic assumptions must hold:

1. The dependent variable (usually denoted Y) can be expressed as a function of a specific set of independent variables, where the function is linear in unknown coefficients or parameters, and an additive error (or disturbance) term. The coefficients are assumed to be constants but are unknown. Violations of this assumption are called "specification errors", some of which are listed below:

    - Wrong set of regressors - omitting relevant independent variables or including variables that do not belong.
    - Nonlinearity - when the relationship is not linear in parameters
    - Changing parameters - when the parameters do not remain constant during the period in which the data were collected.

2. The expected value of the disturbance term is zero; i.e. the mean of the distribution from which the disturbance term is drawn is zero. A violation of this assumption introduces bias in the intercept of the regression equation.

3. The disturbance terms (there is one for every row of data) all have the same variance and are not correlated with one another. This assumption is often violated with one of the following problems:

    - Heterskedasticity - when the disturbances do not all have the same variance (often the case in cross-sectional data); constant variance is called "homoskedasticity",
    - Autocorrelated Errors - when the disturbances are correlated with one another (often the case in time-series data)

4. For standard statistical inference (the next lecture) we usually assume $e_i \sim N(0, \sigma^2)$.

5. It is possible to repeat the sample with the same independent variables. Some common violations are:

- Errors in variables - when there is measurement error in the independent variables.
- Autoregression - when a lagged value of the dependent variable is an independent variable
- Simultaneous equations - when several dependent variables are determined jointly by several relationships.

6. The number of observations is greater than the number of independent variables, and that there are no exact linear relationships between the independent variables.

8.4.3 Properties of Estimators

There are certain properties that we want our estimators to embody. The two main properties are unbiasedness and efficiency. The bias of an estimate is the distance that estimate is from the true value of the thing it is estimating. For example, suppose you are considering purchasing an instrument that measures the salt concentration in water samples. You can think of the instrument as an "estimator" of the salt concentration. Machines typically introduce some small measurement error, and you want to know how accurate is the brand of machine you are considering purchasing. To test the machine's accuracy, you conduct the following experiment:

1. Take in a sample with a known salt concentration of, say, 50 ppm,
2. Use the machine to measure the salt concentration,
3. Repeat the measurement, say 10000 times,
4. Plot a histogram of the measurements.

Figure 8.4 shows the histograms for (1) a biased machine (because the expected value of the reading is not equal to the known true value), and (2) an unbiased machine. The second desirable property, "efficiency", means that an estimator should have the smallest variance. In the example above, the first machine (the biased one) has a smaller variance, and in that respect is a better estimator than the first. The example illustrates that often we will have to make a tradeoff between biasedness and efficiency. Under the

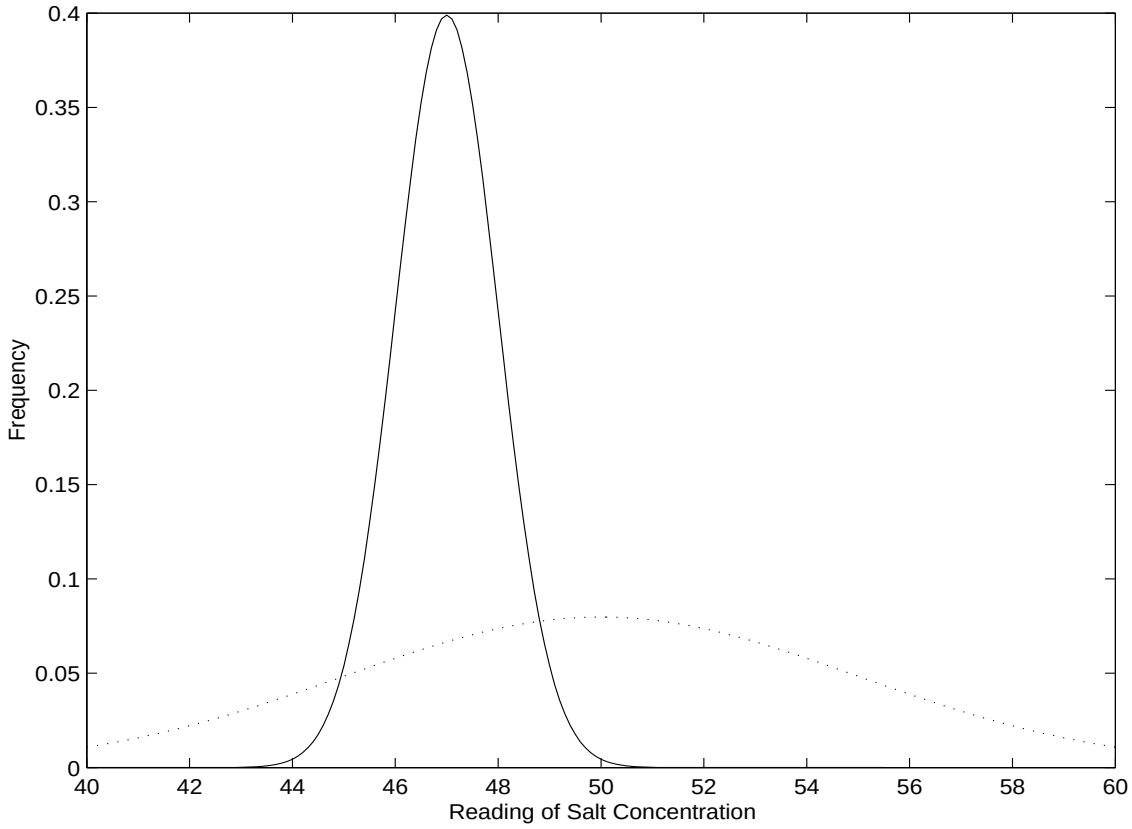


Figure 8.4: A biased estimator of salt concentration (solid line) and an unbiased but less efficient estimator (dotted line).

assumptions outlined above, the OLS estimator is the "best" unbiased estimator, which means that of all the possible unbiased estimators, the OLS estimator has the minimum variance.

8.4.4 Correlation vs. Causation

Explanatory variables may "explain" the dependent variable well, but this does not necessarily imply a causal link. This highlights the problem with "data mining", where many many relationships are tested to see which one has the best "fit". A high fit does not necessarily mean that the essential structure of the relationship between two or more variables has been accurately modeled. This makes it extremely difficult to use the model for any sort of predictions.

The most appropriate way to select a model is to first use basic principles or theory to construct a structural relationship between the variables of interest. For example, just because chicken production and global CO_2 measurements are nearly perfectly correlated over time, doesn't mean that if chicken production increases in the future, so will global CO_2.

True causality is extremely difficult to tease out statistically. One method is called "Granger Causality", which essentially asks whether the explanatory variable happens prior to the dependent variable; if so, a causal link is more defensible. We won't go into a great deal of detail on this subject, but it is important to remember that *correlation does not imply causation.*

8.4.5 Plotting the Residuals

Plotting the residuals is a cheap and effective way to assess model performance and specification. Remember, under our assumptions, the expected value of any one error is zero, and the variance or "spread" of the residuals should remain constant for all residuals. Put simply, the residuals should not exhibit any "pattern". If they do, something is probably wrong, and a great deal of effort should be taken to try to uncover the source of this pattern.

Several plots provided by S-Plus are helpful in assessing the fit and assumptions of a model. Some particularly useful plots are as follows (with examples from the Chlorophyll-a example above):

1. Residuals vs. Fit (figure 8.5). This plot is used to assess whether there is unexplained structure in the model (e.g. whether you have

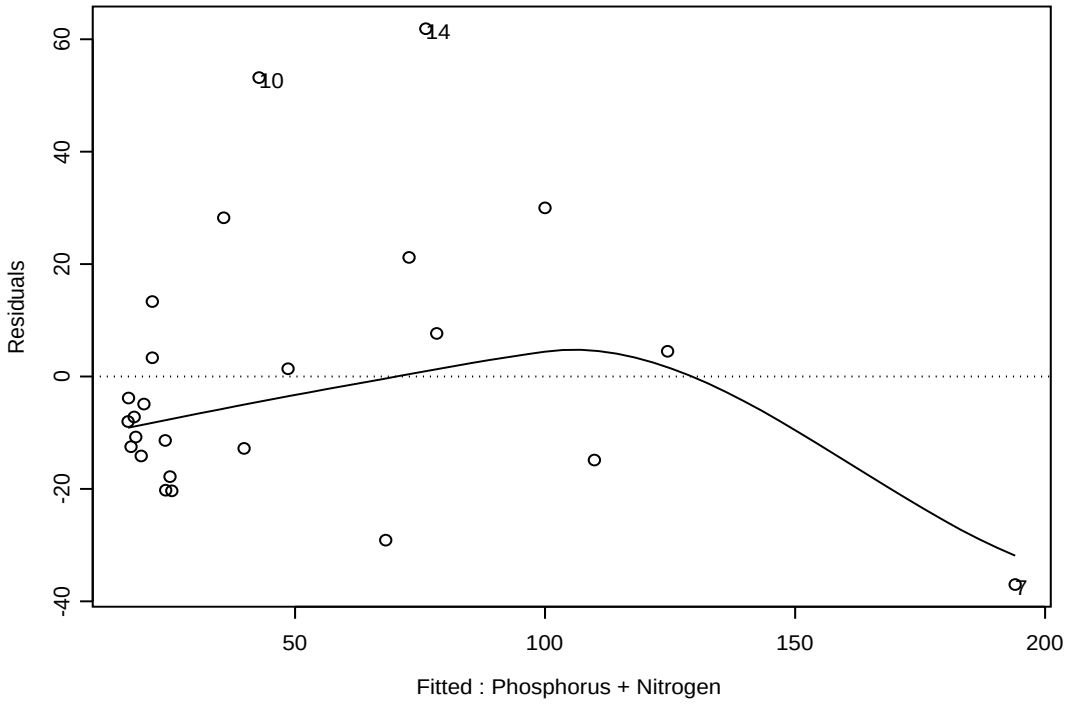


Figure 8.5: Residuals of the chlorophyll-a regression plotted against the predicted values.

misspecified the model or something was measured with error). Under the assumptions of the CLRM, this should appear as random noise. Sadly, it fails this test, since it appears as though small fitted values also tend to have negative residuals.

2. Normal Quantile Plot of Residuals (figure 8.6). Recall our assumption that the errors of a linear regression model are normally distributed. As introduced earlier, the residuals will fall on the diagonal line if they are normally distributed. The normal quantile plot for the chlorophyll example is in figure 8.6, which suggests that the residuals closely follow the normal distribution, except for large values, which are more likely

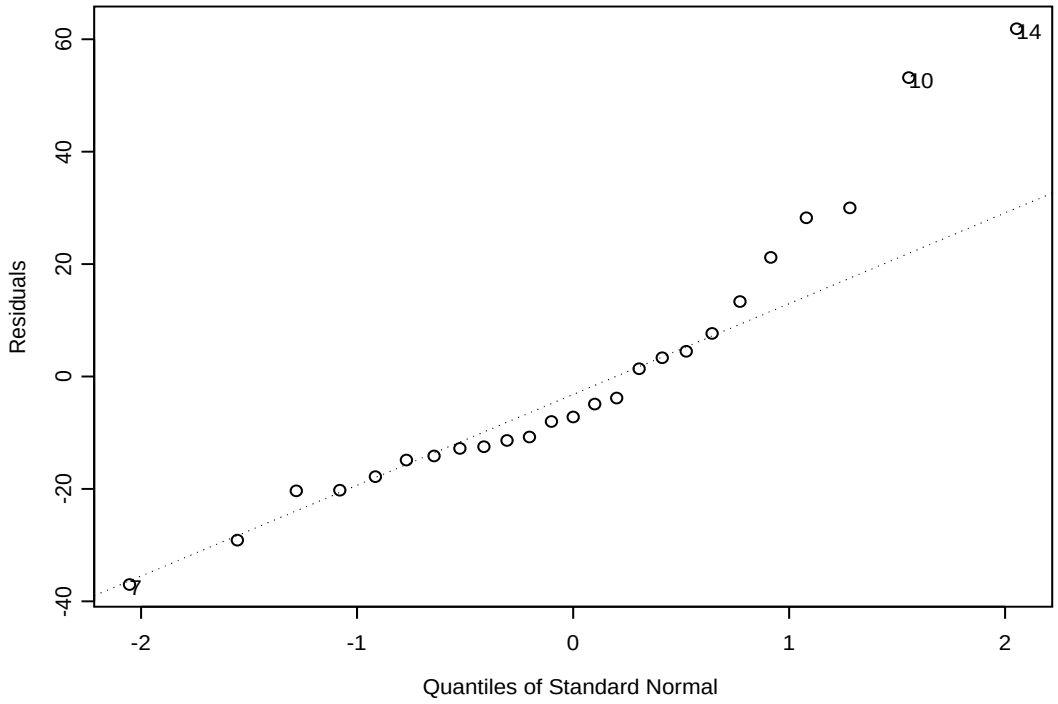


Figure 8.6: Quantile plot of residuals from the chlorophyll-a regression. If the residuals were normally distributed then the points would fall along the dotted line.

for our residuals than they are for the normal distribution. See figure 8.7 for the histogram of our residuals. Many other plots are available, and the S-Plus online help is extremely useful in this regard[3].

8.4.6 Consequences of Violating Assumptions

Violating any of the assumptions of the CLRM causes the estimates, or our confidence in the estimates, to be incorrect. The following is a brief attempt

[3]See Help-Online Manuals-Guide to Statistics vol I. Then look up linear regression.

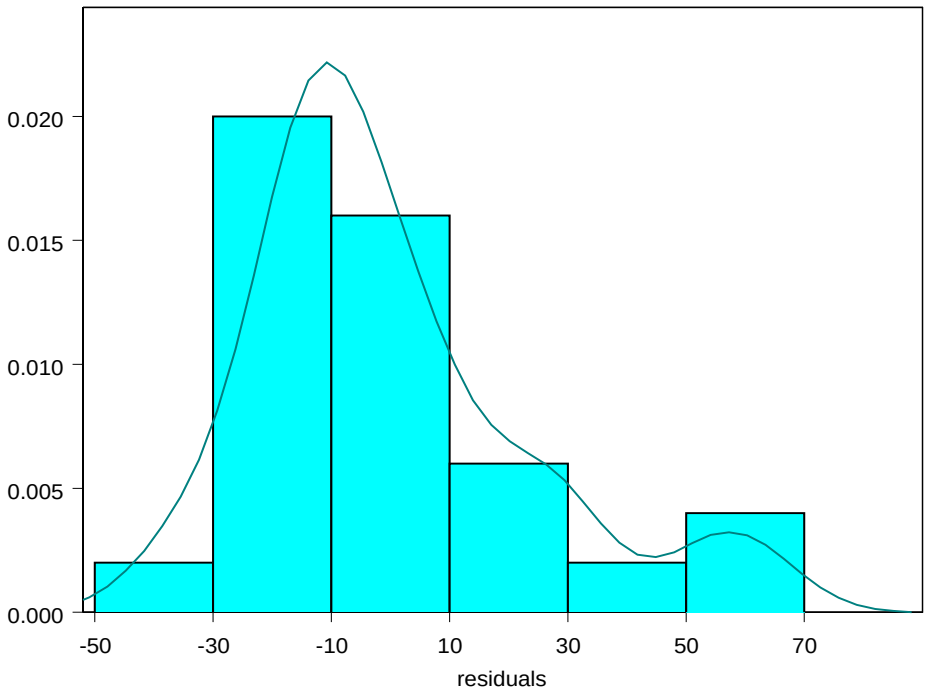


Figure 8.7: Histogram and density plot of the residuals from the chlorophyll-a regression.

to summarize the consequences of common violations, and what can be done about it.

Problem	How Detect?	Consequence	Possible Corrections
Autocorrelation	Durbin-Watson	Unbiased; wrong inf.	GLS
Heteroskedasticity	Plot resids[4]	Unbiased; wrong inf.	GLS
Contemporaneous Corr.[5]	Plot; Hausman Test	Biased	I.V.
Multicolinearity	Correlation table	Usually OK	omit?
Omitted Variables	Theory[6]	Biased	Add variable(s)
Wrong Regressors	Theory	Unbiased; extra noise	Omit Variables
True nonlinear	Plot residuals	Wrong inf.	Non-linear model

8.4.7 A Practical Guide to Model Specification

1. Start with theory. What variables do you need? How should they enter? What is the source of error, and how does it enter? The alternative is "data mining", which refers to a situation when you allow the data to drive the functional form of the model. Data mining is appropriate in some situations, but your first line of inquiry should be into the theoretical underpinnings of the process you are trying to model. Absolutely do not fit a model to data and then do statistical inference on the same data - this is "double dipping". You may be able to split your data sample, fit the model to part and do statistical inference on the other part.

2. Check the assumptions for the CLRM to make sure the OLS estimator is an appropriate estimator for the problem you have specified.

3. Collect and plot your data. Look for outliers, inconsistencies, etc. Get to know your data.

4. Estimate the model, and run F tests (description coming in a later lecture) to test restrictions of the model. At this point, you may want

[4]Formal tests include: Goldfeld-Quandt, Breusch-Pagan, and White.

[5]When e_n is correlated with X_n. Measurement error gives rise to contemporaneous correlation. A form of measurement error exists when $E(e_i) \neq 0$, in which case the intercept term will be biased.

[6]Plotting residuals and getting to know your data may give rise to understanding of what omitted variables cause outlying $\hat{e}_i$'s.

to try a Box-Cox transform (of your variables- thus keeping the model linear in parameters) to ensure the errors are normally distributed.

5. Check the R^2 statistic and the "adjusted" R^2 statistic to get an idea of how much variation your model explains.

6. Plot residuals. If there appears to be a pattern (anything other than a random scatter), you have a misspecification problem (that is, one of the assumptions of the CLRM does not hold).

7. Seek alternative explanations. What else may cause this pattern in the data?

8.5 Application of Technique

So far, we have not discussed how the estimator works, and that remains for the next set of notes. Let's decide which regressors (independent variables) we will use in our regression - we already know, by assumption (1) we must include all relevant regressors (in fact, the plot of gas consumption vs. price provides a perfect example of this - the regression line would slope up, suggesting that people purchased more gas as the price increased). From demand theory in economics, the primary constituents of demand are price, income, and prices of substitutes and compelements. To proceed, let's try to explain gas consumption per capita (G) with (1) gas price (Pg), (2) income (Y), (3) new car prices (Pnc), and (4) used car prices (Puc) in the regression. Now we need to decide on a functional form for our model. The simplest functional form is linear in the regressors:

$$G = \beta_0 + \beta_1 Pg + \beta_2 Y + \beta_3 Pnc + \beta_4 Puc + \epsilon \tag{8.4}$$

Another common specification in economics is the log-log specification:

$$log(G) = \beta_0 + \beta_1 \log(Pg) + \beta_2 \log(Y) + \beta_3 \log(Pnc) + \beta_4 \log(Puc) + \epsilon \tag{8.5}$$

One reason the log-log specification is commonly used is that the parameter estimates can be interpreted as elasticities (the estimate of β_1, called $\hat{\beta}_1$, is intepreted as the % change in gas consumption with a 1% change in gas price).

Now we'll estimate both equations above, and provide the parameter estimates and some summary statistics below:

Model	β_0	β_1	β_2	β_3	β_4	R^2	p-val: F
Linear	-0.09 (.08)	-0.04 (.002)	0.0002 (.000)	-0.10 (.11)	-0.04 (.08)	.97	.000
Log-Log	-12.34 (.000)	-0.06 (.08)	1.37 (.000)	-0.13 (.33)	-0.12 (.16)	.96	.000

Now, suppose a parameter estimate turns out to be zero. That suggests that changes in the variable to which the parameter is attached do not have any effect on the dependent variable. For example, suppose our best estimate of β_1 in equation 8.4 was 0 (instead of -0.04). That would suggest that, in fact, price has no effect on gas consumption. Since there is still variability in the data, the value in parenthesis, called the "p-value" for a particular parameter, gives the probability that the parameter is as high (or low) as the parameter estimate simply by chance. Don't worry too much about this now, but just remember that a very low p-value (say ≤ 0.05) means that we have a great deal of confidence that the coefficient is truly different from zero. A high p-value suggests that the true value of the parameter may be zero (and therefore, it's associated variable may have no effect).

The second summary statistic is the R^2 value which gives a measure of the goodness of fit of our model. R^2 values of, say, $\geq .7$ are very high (it ranges from 0 to 1), and usually mean the model fits very well. We'll return to this later.

Finally, the p-value of the F-statistic is analogous to the p-value for each parameter estimate, except that it refers to the model as a whole. Low p-values ($p < .05$ or so) for the F-statistic mean that the parameter estimates, taken as a whole, do provide some explanatory power for the dependent variable.

We'll conclude by answering our gas market question: What is the effect of a price change on the quantity of gasoline purchased. Holding all other things constant (income and car prices), we can say (using model 8.4 above) that a one unit increase in price results in about a .04 unit decrease in the quantity purchased (since the units of gas per pop are thousands of gallons, a \$0.10 increase in the gas price leads to approximately a 4 gallon decrease in gas consumption per capita. Higher income has a positive effect on gas purchases, as expected, and the prices of both new and used cars has a negative effect on gas prices.

8.5.1 Confidence intervals for the predicted marginal effect

On the basis of a multiple linear regression analysis, we concluded that a \$0.10 increase in the gas price corresponds to about a 4 gallon decrease in per capita gas consumption. But how precise is this number? On thing we may be particularly interested in is whether we are even certain that the true value is indeed positive at all (for reasons that only show up in the error term, some people might actually drive more after the gas price increase). The question of precision of our estimate is best answered by calculating a confidence interval for the true value of this response. We'll compute a 90% confidence interval. Recall our linear model specification (where the dependent variable is G, gas consumption per capita):

$$G = \beta_0 + \beta_1 Pg + \beta_2 Y + \beta_3 Pnc + \beta_4 Puc + \epsilon \tag{8.6}$$

where, as reported in a previous set of notes, the estimate of β_1 is -.04237 with a standard error of .00984.

If we assume that the errors from our regression are normally distributed, then our confidence interval for any given parameter (β_1 in this case), is based on the t distribution. The critical t statistic for a 90% confidence interval with 32 degrees of freedom (number of data points (37) minus number of parameters(5)) is 1.7. The associated confidence limits are:

$$L_l = -.04237 - (1.7)(.00984) = -.0591 \tag{8.7}$$
$$L_u = -.04237 + (1.7)(.00984) = -.0256 \tag{8.8}$$

In other words, we expect the true response to be contained in the interval [-.0591, -.0256] about 90% of the time. This suggests that there really is an important negative response of consumption to gas price increases; a 10 cent increase in gas price will correspond to something like a 3 to 6 gallon decrease in per capita gas consumption.

Bootstrapping the confidence intervals in S-Plus (using the script "coef(lm(G.POP $\sim$ PG+Y+PNC+PUC))" in the Expression box), gives empirical percentile confidence interval for β_1 of [-.063, -0.026]. The distribution of β_1 from 1000 bootstrapped samples is shown in figure 8.5.1.

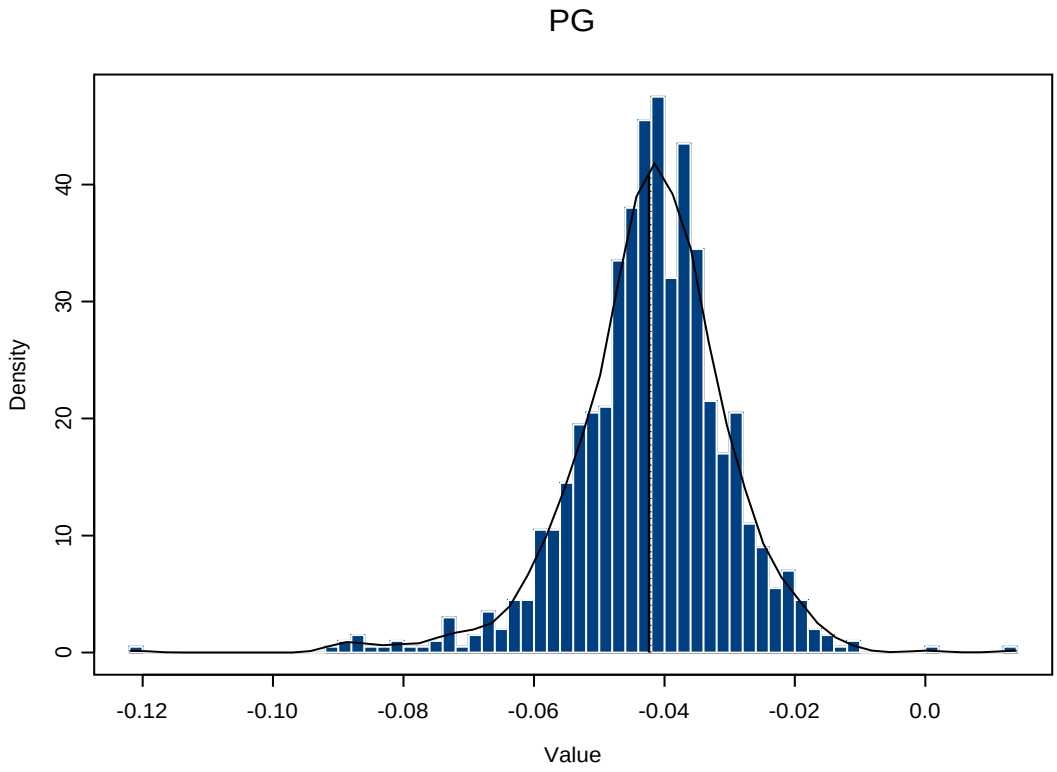

PG
Density
0
10
20
30
40
-0.12
-0.10
-0.08
-0.06
-0.04
-0.02
0.0
Value

Table 8.1: Splus output for the regression of per-capita gas consumption on gas price, income, and new and used car prices.

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0898	0.0508	-1.7687	0.0868
GasPrice	-0.0424	0.0098	-4.3058	0.0002
Income	0.0002	0.0000	23.4189	0.0000
New.Car.Price	-0.1014	0.0617	-1.6429	0.1105
Used.Car.Price	-0.0432	0.0241	-1.7913	0.0830

8.5.2 Significance of regression terms

Recall that the model for gas consumption was

$$G = \beta_0 + \beta_1 P + \beta_2 I + \beta_3 N + \beta_4 U, \tag{8.9}$$

where G is the per-capita gas consumption, P is the price of gas, I is the average income, N is the average price of a new car, and U is the price of a used car. Running the model in Splus returns the output in table 8.1. The column labelled `Pr(>|t|)` gives the P–value for the null hypothesis that the coefficient equals zero. Thus we might conclude that there is very strong evidence that gas price and income affect gas consumption, and rather weak evidence that car prices have an effect.

In a report you want to include much of this information in tabular format. Note that you only need two of SE, t, and P; the usual convention is SE and P. You also want to give the variables natural names, instead of whatever you had in the computer dataset, and you want to make sure that you explain the statistical test used to generate the P-values. Thus you might present something like table 8.2.

Table 8.2: Estimates and standard errors of the coefficients in the gas consumption regression (eq. 8.9). The P–values are for a two–sided t–test against the hypothesis that the coefficient is zero ($df = 31$).

	Estimate	SE	P
Intercept	-0.090	0.051	0.09
Gas Price	-0.042	0.0098	0.0002
Income	0.0002	0.0000	¡0.0001
New Car Price	-0.10	0.062	0.1
Used Car Price	-0.043	0.024	0.08

Chapter 9

OLS Estimation and Dummy Variables

9.1 Motivation

The classical linear regression model introduced in the previous set of notes provides a framework in which to model how one or more independent variables affect a dependent variable. The process by which the independent variables influence the dependent variable should be driven by theory, as should the distribution and entry of the error term. However, the values of unknown parameters of the model should be driven by the data. Until now, we have abstracted away from the manner in which these parameters are chosen. These notes show how the hypothetical linear relationship between variables can be quantified using data.

9.2 Example & Question

Alien (a.k.a. exotic, invasive, and non-indigenous) species cause significant economic and ecological damage worldwide. Ecologists, agronomists, and other scientists have studied this problem in depth, and have generated hypotheses about the variables that are thought to influence the number of alien species in a country. Variables such as the country's area, number of native species, geographic location, GDP, trade activity, percent land in agriculture, and population density are all thought to play a role. We are interested in estimating the importance of each of these variables. In particular, it has

been hypothesized that the more densely populated is a country, the more susceptible the country is to invasions. We'd like to formulate this theory into a testable hypothesis. We would also like to know whether islands are more invaded than continental countries. Finally, for conservation purposes, we want to know the extent to which disturbance, in terms of the percent of land in agriculture, affects how invaded a country is.

9.3 Evidence or Data

The data for answering this question are posted in the course web. To get started, we will first examine only the relationship between a country's population density (POP_t) and the associated ratio of alien to native plant species (A_t). The full data set for these two variables is shown in figure 9.1.

To better illustrate the concept of parameter estimation in the classical linear regression model, we first condense this rather large, complicated, problem into an over-simplified version of the model. In this over-simplified model, the only things affecting the ratio of alien to native plant species in a country are the population density and some random error (from unpredictable sources). Further suppose that our dataset is composed of only three data points: for Cayman Islands, France, and Poland. These data are represented in figure 9.2.

9.4 Technique

First we discuss the oversimplified model, then we'll turn to answer our original question of how population density affects the ratio of alien to native species in a country. For our oversimplified model (three data points and one regressor), we might specify the following statistical model:

$$A_i = \beta_1 + \beta_2 P_i + \epsilon_i \tag{9.1}$$

where A_i is the i^{th} observation of the variable A (ratio of alien to native plant species), β_1 is the yet unknown value of the intercept term, β_2 is the yet unknown value of the coefficient on P, P_i is the i^{th} observation of population density, and ϵ_i is the i^{th} error term. Note that in this example, there are 3 A_i's (known), 3 P_i's (known), 1 β_1 (unknown), 1 β_2 (unknown), and 3 ϵ_i's

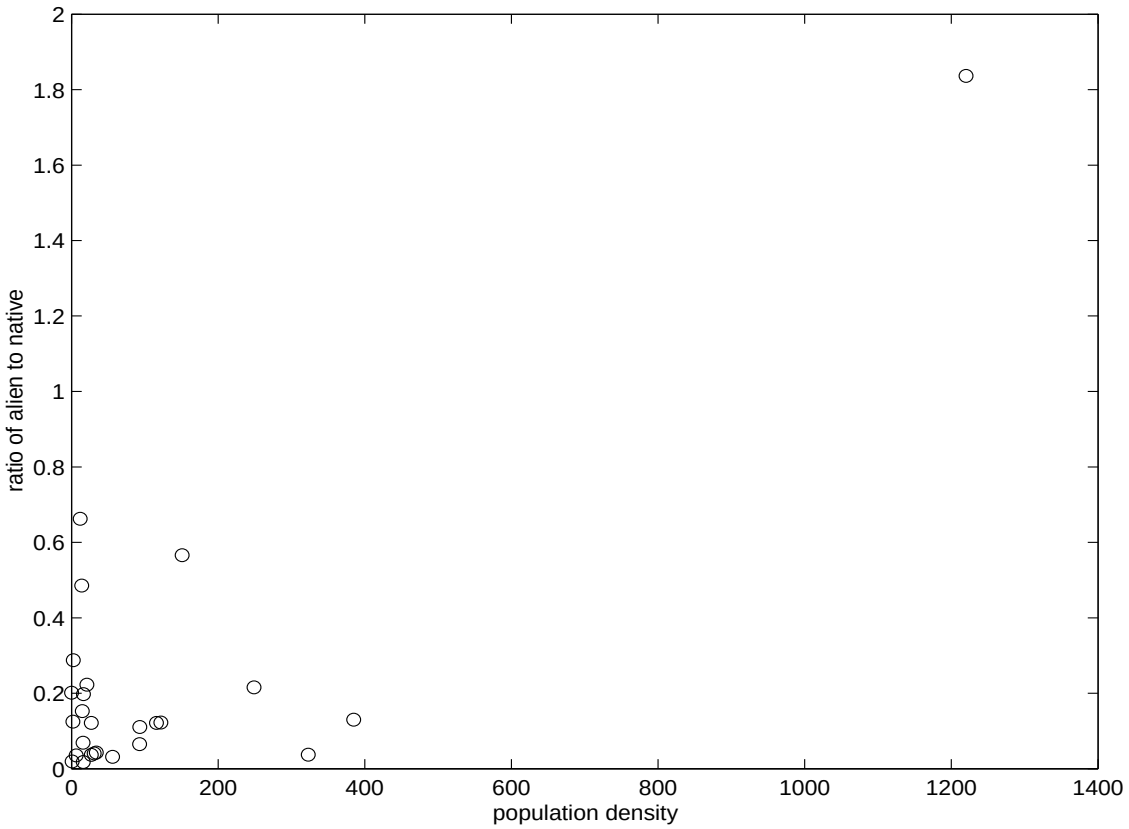


Figure 9.1: Ratio of alien to native species vs. population density, by country.

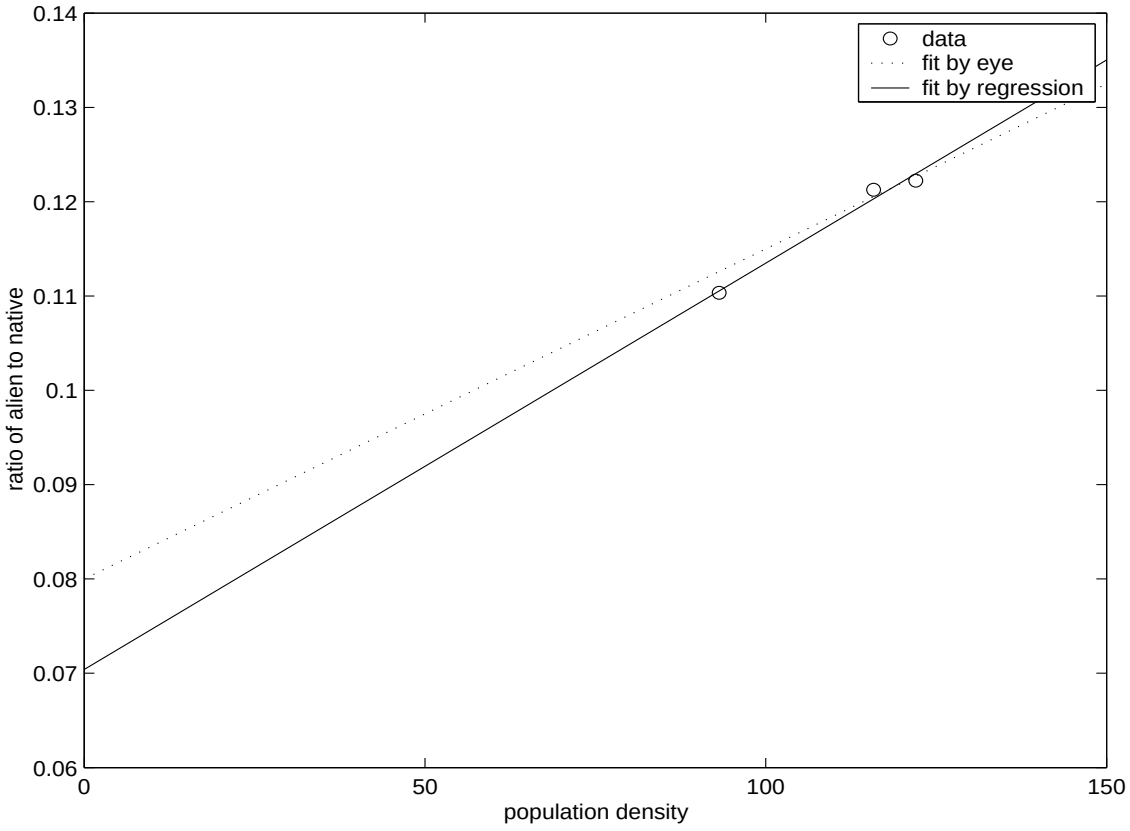


Figure 9.2: Ratio of alien to native species vs. population density, for three countries.

(unknown - they depend on the β's). Our task is to estimate the parameters β_1 and β_2 to give the "best fit" to the data.

The first step is to define what is called the "residual" for each observation. The residual is simply the difference between the obeserved value of the dependent variable and the predicted value of the dependent variable. For example, suppose we just draw a line by hand that seems to fit well (see the line in the figure). The intercept for this line is 0.8, and the slope is 0.00035. Using this regression line, the residuals (denoted e_i), are $e_1 = -0.00073$, $e_2 = 0.0023$, and $e_3 = 0.00048$ (you should make sure you can calculate these residuals). We want to fit a regression line in a more systematic way (rather than estimating it by hand), in order to make the residuals as small as possible. Obviously, a line that fits some observations very well will not fit others well at all; so we need a criterion that takes account of all the residuals simultaneously.

Suppose, for example, we simply wish to minimize the sum of the residuals. This doesn't work because we automatically obtain a sum of residuals of 0 if we let the intercept equal the average of the dependent variable observations and the slope equal to zero. The most common criterion is to minimize the sum of squared residuals. In this case, the only way to achieve a sum squared residual value of 0 is to have the regression line perfectly predict each and every data point. Choosing regression parameters to minimize the sum of squared residuals places a very high penalty on observations farther away from the regression line, and is therefore very sensitive to outliers. One data point can drag an entire regression line in its direction in order to reduce the sum squared residuals.

In our simple three point example, we have the following:

$$A_i = \beta_1 + \beta_2 P_i + \epsilon_i \tag{9.2}$$
$$\epsilon_i = \beta_1 + \beta_2 P_i - A_i \tag{9.3}$$
$$e_i = \hat{\beta}_1 + \hat{\beta}_2 P_i - A_i \tag{9.4}$$

So the sum of squared residuals, call it S, is:

$$S = e_1^2 + e_2^2 + e_3^2 \tag{9.5}$$
$$= (\hat{\beta}_1 + \hat{\beta}_2 115.83 - 0.121)^2 + (\hat{\beta}_1 + \hat{\beta}_2 93.17 - 0.110)^2 \tag{9.6}$$
$$+ (\hat{\beta}_1 + \hat{\beta}_2 122.0 - 0.122)^2 \tag{9.7}$$

To work this by hand, one needs to take the derivative of S with respect to

$\hat{\beta}_1$ and $\hat{\beta}_2$, and set them both equal to zero. This is equivalent to running this simple regression in S-Plus, giving the following output:

```
    *** Linear Model ***

Call: lm(formula = aliennative ~ popdens, data = paramest3,
    na.action = na.exclude)
Residuals:
         1          2          3
 0.0009573 -0.0002049 -0.0007524

Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept)  0.0704  0.0064    11.0228  0.0576
    popdens  0.0004  0.0001     7.4954  0.0844

Residual standard error: 0.001235 on 1 degrees of freedom
Multiple R-Squared: 0.9825
F-statistic: 56.18 on 1 and 1 degrees of freedom, the p-value
    is 0.08444
```

So the parameter estimates that minimize the sum of squared residuals are: $\hat{\beta}_1 = 0.0704$ and $\hat{\beta}_2 = 0.0004$.

Now that the idea of minimizing the sum of squared residuals is fixed in our minds, we turn to an investigation into our original questions (restated here):

1. Is population density an important factor in determining how "invaded" a country is? In which direction?

2. Holding all else constant, are island nations more invaded than continental nations? If so, by how much?

3. We were unable to collect invasive species data for an island nation that is "average" in all other respects. What is the predicted ratio of alien to native plant species?

The regression problem is also solved with a one–sample t–test: simply substitute the estimate of β_1 and the standard error of the estimate (produced

by the regression routine) into the following equation:

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}. \tag{9.8}$$

Everything proceeds the same way from here, except that the degrees of freedom are given by $df = n - k$, where k is the number of parameters in the regression. The P–values that are output from the regression routine are testing against the null hypothesis that the parameter equals zero, but you could test against any null value. For example, in the Ricker fisheries model from your homework, you might want to test whether the estimated coefficient of $\ln S_t$ was different from one (if it is, you may need to reformulate your model).

9.4.1 Dummy Variables

A dummy variable is an independent variable (right hand side variable) that acts as a "flag" to denote a certain characteristic. In this example, we want to allow island nations to behave differently than continental nations. There are several ways that island nations may behave differently than continental nations. To allow for this possibility, we construct a "dummy variable" called ISL, that takes on the value 1 if the country is an island, and takes on the value 0 if the country is not an island. The coefficient on ISL will then be interpreted as the increase in invasiveness of an island nation versus a continental nation, holding all other variables constant.

> JARGON ALERT: A dummy variable may also be called an *indicator variable*.

Dummy variables are often used to denote gender, race, or other such variables. In general, if there are n possible categories between which you want to distinguish, you need $n - 1$ dummy variables. In this example there are 2 categories (island and continental), so we only need one dummy variable. Suppose we wanted to distinguish between continents. There are 6 continents represented in our data set, so we would need 5 dummy variables (call them $D1$, ..., $D5$) taking on the following values. For each country, D1 equals 1 if the country is in North America and 0 otherwise, D2 equals 1 if the country is in South America and 0 otherwise, ..., D5 equals 1 if the country is in Europe and 0 otherwise.

This method of using dummy variables can be interpreted as changing the intercept across categories (that is, island nations have a different intercept than continental nations). But what if we think, for example, that island nations have a fundamentally different response to changes in, say, population density than do continental nations? Then we need to interact our dummy variable with the variable of interest. In this case, we would need to include a variable, call it $ISLP$, that is the product of the dummy variable ISL and the variable P. That way, we can investigate whether island nations respond differently than continental nations.

JARGON ALERT: In biometry, an analysis that incorporates both continuous and categorical variables is often called *ANCOVA* (ANalysis of COVAriance).

9.4.2 Goodness of Fit

The goodness of fit of a linear model is usually assessed with a statistic called the coefficient of determination, or R^2. It is the square of the correlation coefficient between the dependent variable and the OLS predicted value of the dependent variable. For the sake of discussion, let the dependent variable be denoted y. Here are some interesting and important facts about the coefficient of determination (largely adapted from Kennedy "A Guide to Econometrics"):

- The total squared variation of y about its mean, $\sum(y-\bar{y})^2$, is called the sum or squares total (SST: we will revisit this in constructing ANOVA tables). The explained variation, the sum of squared deviations of the estimated y's about their mean, $\sum(\hat{y}-\bar{\hat{y}})^2$ is called SSR (sum squares regression). Then R^2 is SSR/SST, the percentage of the total variation that is explained by the regression.

- R^2 can take on values between 0 and 1. Time series models typically give higher R^2 values than cross sectional models. Generally speaking, .7 or higher would be considered high. A high R^2 value is often called a "good fit".

- Comparing R^2 values for different models is often difficult because the statistic is sensitive to the range of variation of the dependent variable.

- The R^2 and OLS criteria are formally identical: the OLS estimator maximizes R^2.

- Practitioners often place too much importance on R^2 values. In particular, adding variables to the right hand side can only increase the R^2 statistic. For this reason, some people prefer the "adjusted" R^2, which imposes a penalty for adding explanatory variables.

9.5 Application of Technique

These are very important questions in understanding alien species introductions and in designing policies to mitigate damage from invasive species. Our left hand side (dependent) variable is the proportion of alien to native plant species (A). A high value of A means that the country is highly invaded. Of course, this is not an exhaustive statistical analysis, but we can run and analyze a linear regression to get an idea of the answer to the above questions. Here's our statistical model:

$$A_i = \beta_1 + \beta_2 P_i + \beta_3 ISL_i + \beta_4 GDP_i + \beta_5 AGR_i + \epsilon_i \tag{9.9}$$

where A and P are as before, ISL is a dummy variable for island, GDP is gross domestic product, AGR is percentage of land in agricultural production, and ϵ is an error term. Before we run this regression, let's predict some of the signs and magnitude of the coefficients.

- β_1 is the intercept term. It gives the proportion of alien to native species in preindustrial times. For island nations, the intercept is actually $\beta_1 + \beta_3$. For continental nations, the intercept is just β_1 (make sure you understand this). We expect β_1 to be non-negative, but very small. In general, you should include an intercept term in your models, except where theory suggests that it equals zero. This is a case where it might be appropriate to constrain the intercept to zero.

- β_2 is the coefficient on population density. It gives a measure of the effect of a marginal increase in population density on invasions. We expect it to be positive.

- β_3 is the coefficient on the dummy variable, ISL. Since ISL only takes on values of 1 or 0, β_3 cannot measure a marginal change in the

variable. Instead, it gives the change in the dependent variable when the dummy variable has a value of 1 (as opposed to 0), i.e. the increased invasiveness of island versus continental nations. We expect it to be positive.

- β_4 is the marginal increase in invasiveness with an increase in GDP. We expect it to be positive.
- β_5 is the marginal increase in A with an increase in land used for agriculture. This is expected to be positive.

Running this multiple regression in S-Plus gives the following output:

```
*** Linear Model ***

Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept) -0.0184  0.0859    -0.2141  0.8326
     Island  0.0623  0.0821     0.7588  0.4564
   Pop.dens  0.0010  0.0002     6.2330  0.0000
        GDP  0.0000  0.0000     3.3249  0.0032
        Agr -0.0014  0.0015    -0.9039  0.3763

Residual standard error: 0.1804 on 21 degrees of freedom
Multiple R-Squared: 0.7986
F-statistic: 20.81 on 4 and 21 degrees of freedom, the p-value
    is 4.632e-007
```

Based on the results from this regression, the answer to our previously posed questions are:

1. Population density appears to have a small, though significantly positive effect on invasions. The fact that the p-value from the t-test is low (.0000) tells us that although the effect is small, it is significantly different from zero. A one unit increase in population density results in about a .001 unit increase in A.
2. Island nations are more heavily invaded than continental nations. Since the dummy variable $Island$ does not interact with any other variables, the coefficient (.0623) can be interpreted as follows: Island nations

have a higher proportion of alien to native species (by about .06), all other variables being held constant. However, the higher p value (.456) suggests that this value is highly variable, and the estimated parameter is not statistically different from zero.

3. Predicting with a model such as this is more of an art than a science. In particular, we must decide which variables to include, and which to omit. Just because certain variables are not statistically significant (the intercept, island, and agricultural activity in this case), doesn't mean that they provide no assistance in predicting. Leaving in all variables, and noting that the average values of the right hand side variables are:

```
***  Summary Statistics for data in:  ExoticSpecies ***

                A     Island  Pop.dens       GDP      Agr
    Min:  0.01754190  0.0000000    0.1600   610.000  0.00000
 1st Qu.:  0.04105672  0.0000000   14.2200  1584.500  7.25000
   Mean:  0.22893174  0.3461538  117.2550  7062.615 32.80769
 Median:  0.12180475  0.0000000   27.0000  4647.000 32.50000
 3rd Qu.:  0.21222179  1.0000000  110.1650 11244.750 55.75000
    Max:  1.83636364  1.0000000 1220.0000 18105.000 74.00000
 Total N: 26.00000000 26.0000000   26.0000    26.000 26.00000
   NA's :  0.00000000  0.0000000    0.0000     0.000  0.00000
Std Dev.:  0.36841784  0.4851645  246.7725  5984.539 25.16191
```

So, the desired prediction is $\hat{A} = -.01839 + .062277(1) + .001009(117.26) + .0000215(7062.62) - .001357(32.81) \approx 0.2696$. That is, we would predict about 27 alien species for every 100 native species on an average island nation (with some rounding error).

Chapter 10

Maximum Likelihood Estimation

10.1 Motivation

We have studied the Ordinary Least Squares (OLS) estimator for regression analysis. But the OLS estimator is only applicable under fairly restrictive assumptions. Perhaps most importantly: (1) the OLS estimator is only applicable to models that are linear (in parameters, recall) and (2) the OLS estimator requires some rather stringent assumptions about the distribution of the error term. In many applications, these assumptions do not hold. An alternative estimator, called the Maximum Likelihood (ML) Estimator (Maximum Likelihood Estimation is denoted MLE), is often the most applicable[1]. In fact, you can never go wrong by replacing the OLS estimator with the ML estimator.

10.2 Example & Question

The federal government and the State of California impose different ozone standards on each county in California. If the measured ozone level on a given day is below the federal (or state) standard, the county is said to be "out of attainment" for that day. From the regulator's perspective, the important

[1]In fact, it can be shown that if the assumptions to use the OLS estimator hold, then the OLS estimator and the Maximum Likelihood Estimator are equivalent.

statistic is the number of days per year that the county is out of attainment (that is, the number of days the ozone standard was violated).

You will be provided with data for a particular California county on the number of days out of attainment for each of 20 years. The task is to model the probability distribution of the number of exceedances per year.

10.3 Evidence or Data

Suppose the number of exceedences are:

Year	Number of Exceedances
1980	1
1981	1
1982	3
1983	1
1984	3
1985	2
1986	2
1987	0
1988	3
1989	4
1990	2
1991	1
1992	5
1993	4
1994	2
1995	2
1996	2
1997	5
1998	2
1999	4

10.4 Technique

Maximum likelihood is a conceptually different estimator than the ordinary least squares estimator. With MLE, we ask the question: What value of the parameter(s) makes it most likely that we will observe the data?

As a very simple example, suppose we flip a coin (not necessarily a fair coin) 5 times and obtain 4 heads and 1 tail. Suppose we denote by p, the probability of heads. We may wish to ask the question, what is the probability of obtaining heads on the next flip? If we have no reason to believe the coin is fair, then it is intuitively obvious that our best estimate of p is $4/5$ or 0.8[2]. Let's see if the maximum likelihood estimator of p (the unknown parameter in this case) is, in fact, 0.8. In this case, assuming that outcome of the flips are independent, each flip can be considered an independent "trial" with probability of success, p. Therefore, this is precisely the definition of the binomial process, where the probability (or likelihood) of x successes in n trials, given a probability of success of p is:

$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, ..., n \tag{10.1}$$

Where I have written $f(x|p)$ to denote the probability of x successes *given* the parameter p. The problem is, we do not know what p is. The maximum likelihood estimator is simply the value of p that gives the highest "likelihood" of observing our particular data. For example, suppose $p = .6$. Then the likelihood of observing 4 successes in 5 trials would be:

$$f(4|.6) = \binom{5}{4} .6^4 (.4)^1 = 0.2592 \tag{10.2}$$

And the likelihood of observing 4 successes in 5 trials given that $p = .7$ is:

$$f(4|.7) = \binom{5}{4} .7^4 (.3)^1 = 0.3602 \tag{10.3}$$

Figure 10.1 shows the likelihood of observing 4 heads and 1 tail for different values of the unknown parameter, p, which is maximized at 0.8, as expected.

For more complicated problems, we will not be able to describe the likelihood function as a known pdf. More typically, we will have many observations, all drawn from the same distribution. Under the assumption of independence between samples, we can calculate the likelihood function as

[2]On the other hand, if we believe the coin is probably fair, then we might want to ask a more subtle question, such as, given that the coin is fair, what is the probability that we would have observed 4 heads and 1 tail?

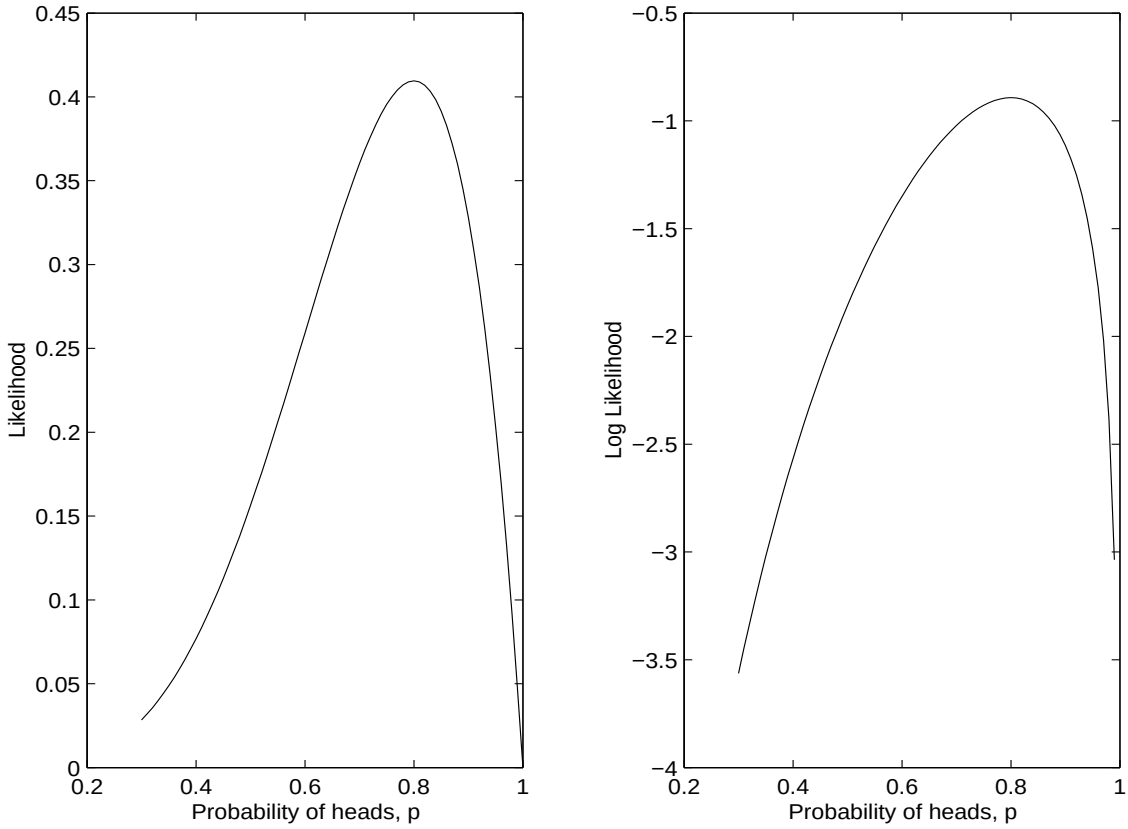


Figure 10.1: Likelihood and log-likelihood functions for p when there are 4 heads and one tail.

the product of the individual probabilities. In the literature, you will often see people maximizing the *log likelihood* function. You should convince yourself that since the logarithm is a monotonic transformation, the parameter that maximizes the log likelihood is the same as the parameter that maximizes the likelihood function. The right hand panel of figure 10.1 shows the value of the log likelihood function.

10.5 Application of Technique

We are now in a position to use the maximum likelihood estimator to model the number of days our hypothetical California county is out of attainment of the ozone standard. Recall from a previous lecture that count data are typically modeled as a Poisson random variable. This requires some assumptions that may not hold here, such as the rate of violations is constant[3] But, given that the rate is constant, we can model this process as a Poisson Random variable, with unknown "rate" parameter, θ. The density for each observation is:

$$f(x_i|\theta) = \frac{e^{-\theta}\theta^{x_i}}{x_i!} \tag{10.4}$$

And since the observations are independent (by assumption), the likelihood of observing $x_1 = 1$ the first year, $x_2 = 1$ the second year, and so on, is just the product of the individual densities, as follows:

$$\begin{aligned} f(x_1, x_2, ..., x_{20}|\theta) &= \prod_{i=1}^{20} f(x_i|\theta) && (10.5)\\ &= \frac{e^{-20\theta}\theta^{\sum x_i}}{\prod x_i!} && (10.6)\\ &= \frac{e^{-20\theta}\theta^{49}}{5.5038\text{x}10^{12}} && (10.7) \end{aligned}$$

The likelihood function and the log likelihood function, as functions of the unknown parameter θ, are plotted in figure 10.2. which both appear to be maximized at approximately 2.4. Based on our assumptions, it appears as though the number of exceedances per year follows a Poisson distribution

[3]This assumption may be violated, for example, if a multi-day weather event is the cause of the violation.

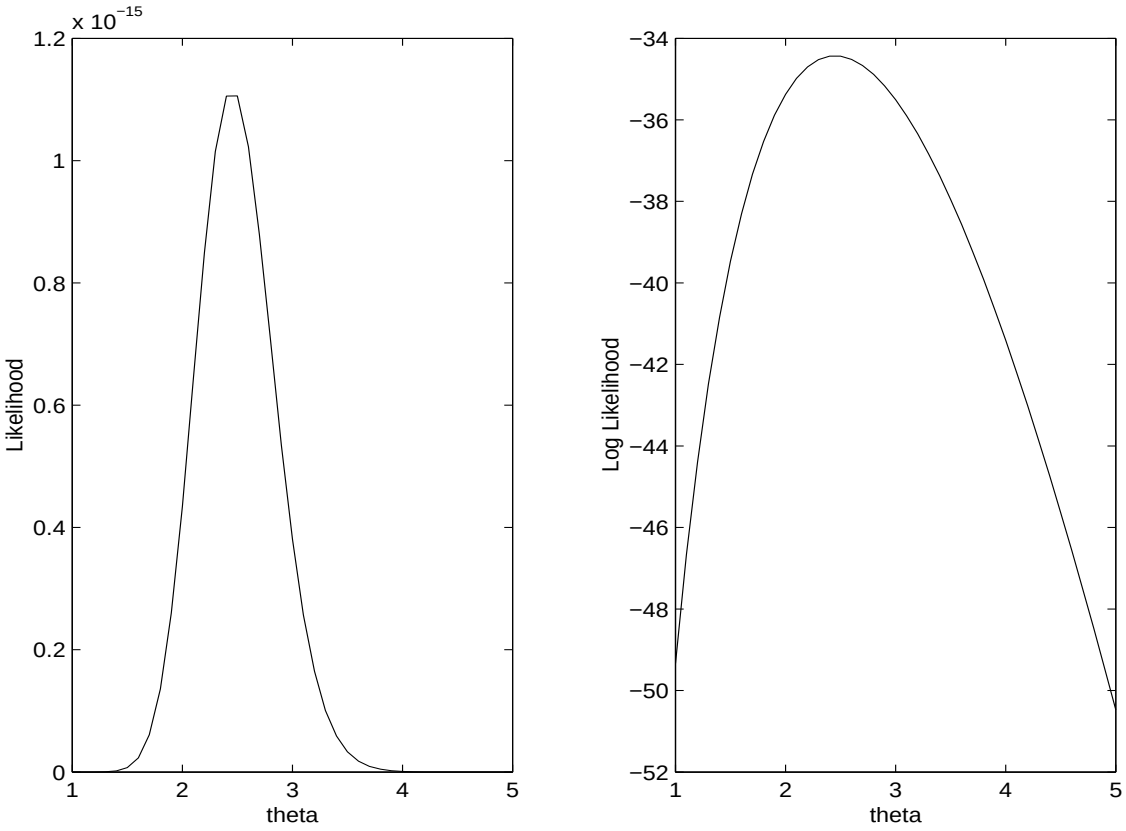


Figure 10.2: Likelihood and log-likelihood functions for the Poisson model of out-of-attainment days.

with parameter (and mean), 2.4. Using S-Plus, this number is more precisely calculated as 2.45, indicating a rate of approximately 2.45 exceedances per year.

Note that this result does not prove that exceedences are Poisson-distributed; it only says that *if* they were Poisson-distributed, then the most likely value of θ is 2.45. It is always a good idea to check whether the model with the ML parameters gives a good overall description of the data.

10.6 When are MLE and OLS Equivalent?

We mentioned above that when the CLRM assumptions hold, the MLE and OLS estimators are equivalent. To illustrate this point, we will revisit the model of lake water quality, as measured by chlorophyll. Suppose you have the following model (introduced earlier in the course) $\log(C) = \alpha + \beta \log(P) + \gamma \log(N) + \epsilon$ where C=chlorophyll, P=phosphorus, and N=nitrogen. Under the CLRM assumptions (basically that $\epsilon \sim N(0, \sigma^2)$ and the errors are uncorrelated with one another), we can use the OLS estimator to calculate the following: $\hat{\alpha} = -4.28$, $\hat{\beta} = 1.24$, and $\hat{\gamma} = 0.91$ (notice your textbook incorrectly estimated $\hat{\alpha}$).

The equivalence of the MLE estimator and the OLS estimator when the assumptions of the CLRM hold can be demonstrated by estimating the parameters with the MLE estimator and checking to see if they are equivalent. To run the model using MLE in S-Plus, got to Statistics $->$ Regression $->$ Generalized Linear. Then, under the CLRM assumptions, choose the error distribution "Gaussian" (i.e. Normal). Then run the regression. Even though a totally different estimation method (MLE instead of OLS) is being used, the results are identical!

This simple example illustrates the very important point that the maximum likelihood estimator, though sometimes more difficult to interpret, applies to more cases than the OLS estimator; using OLS cannot improve upon MLE-generated estimates. However, if the assumptions of the CLRM do hold, then using the OLS estimator does have the advantage that results are very straightforward to interpret (e.g. for t and F tests of restrictions), confidence intervals are straightforward to calculate[4], and the actual computation

[4]These sorts of tests have been worked out for MLE, and are fairly straightforward to apply, but we have not covered them in any detail in this course.

of results does not require numerical optimization, so you never run into a problem that cannot converge to a solution.

Chapter 11

Discrete Dependent Variables

11.1 Motivation

There exist many settings in which the dependent variable takes on discrete (as opposed to continuous) values. A firm's decision about whether to adopt pollution abatement technology is one example. Imagine a data set for analyzing such a problem. We might have data on, say, 1000 firms, where the dependent variable takes on a value of, say, 1 if the firm adopts the technology, and 0 otherwise.

The dependent variable need not only take on two possible values. Suppose we are studying land use change. We may wish to distinguish between the following land-use categories: (1) urban, (2) production (e.g. agriculture), and (3) preserved open space. We might have, say, time-series data on the characteristics of a place and the type of land-use practiced there. Analyzing these data (through a variation of the regression techniques discussed earlier) can facilitate making a forecast of land-use as a function of land characteristics.

In these notes we focus on what are called "binary choice estimators", which, as the name implies, refer to those estimators that are used to model situations in which the dependent variable can take on 2 possible values (typically labeled 1 and 0). Multiple choice models, such as the land-use example above, can be dealt with in a number of ways, but are beyond the scope of this course. For the most part, multinomial models are extensions of binomial models.

11.2 Example & Question

As a precursor to a more lengthy discussion of risk assessment, here we estimate the relationship between the degree of exposure and the chance of adverse consequences. This example is very straightforward, but will illustrate the concept of a discrete (binary) dependent variable. The key concept is the *dose-response curve* which relates exposure to observed adverse outcomes.

Treated foods often contain a byproduct of a fungicide called ethylete thiourea (ETU), which may be harmful to health. An experiment was carried out in the 1970's in which rats were exposed to different doses of ETU (from 0 to 500 ppm/day), and the number of rats that developed tumors was recorded. The objective here is to estimate the dose-response function for rats exposed to ETU. Suppose that based on previous experience, we know that an impact is only detected in humans if it is detected in greater than 10% of rats. The question is, what dose of ETU corresponds to this threshold?

11.3 Evidence or Data

The data are given on the course web in a file called RatsA.xls. In summary, there are 6 groups of rats with approximately 70 rats per group. Figure 11.1 gives the proportion of rats in each group that develop tumors as a function of the dose. The relationship appears to be positive, but it is usually inappropriate to simply fit a linear model to binary dependent variable data (ask yourself why).

11.4 Technique

Suppose we applied the linear regression model we analyzed earlier. If we let Y be the presence or absence of a tumor and X be the dose, the linear model would look something like this:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{11.1}$$

What's wrong with this? Many things are wrong with this picture, but most importantly, predicted values of Y are in no way bounded by 0 and 1. They

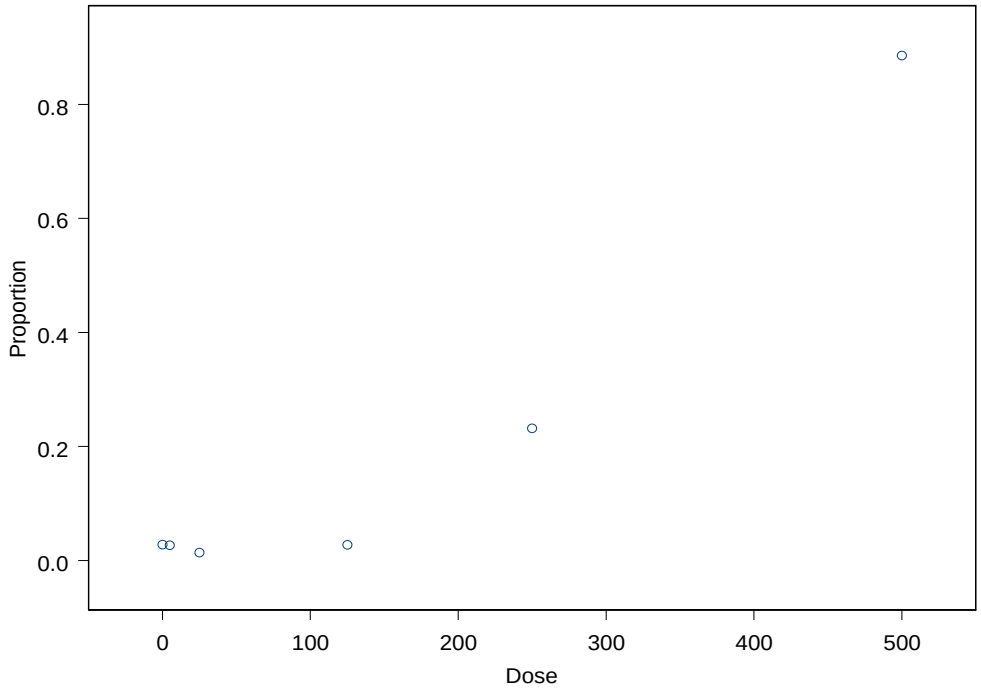


Figure 11.1: Fraction of rats developing tumors after exposure to various doses of ETU.

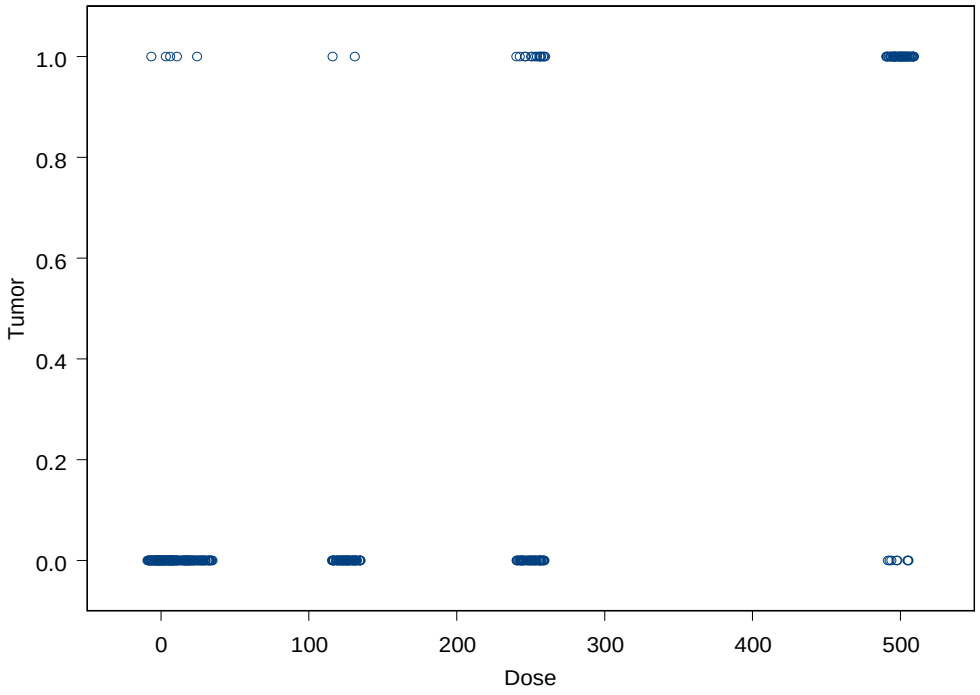


Figure 11.2: Presence or absence of tumors in rats after exposure to various doses of ETU.

should be bounded by 0 and 1 because Y can take on only values of 0 or 1. A natural way to think about Y is the "probability that a rat exposed to an ETU dose of X will contract a thyroid tumor". Since we wish to estimate a model that predicts a probability, we want it to be bounded between 0 and 1.

Let's first answer the question the wrong way; by fitting an OLS model to the data. The figure shows another way of plotting the data. Each data point represents one rat, where the variable "Tumor" is 1 if the rat has a tumor and 0 otherwise. Figure 11.2 shows the data points with a small "jitter" so you can see data points that are stacked on one another. If we fit a linear model to these data, the estimated parameters are: $\hat{\beta}_0 = -.04889$ and $\hat{\beta}_1 = .00167$.

You can confirm that at least two things are wrong with this model: (1) If the rat is not exposed at all, he has a negative chance of contracting a tumor (which is, of course, nonsense) and (2) If the rat is exposed to a dose of greater than 628, he has a greater than 100% chance of contracting cancer: also nonsense.

The above illustration demonstrates one major problem with fitting an ordinary linear model to binary data. Two alternatives are commonly used. They are both based on commonly-used probability distributions: the normal cdf (called the "Probit" model) and the logistic cdf (called the "Logit" model). For the purposes of this discussion, we will focus attention on the Logit model, the Probit model is conceptually identical. The logistic cdf is as follows:

$$F(x) = \frac{1}{1+e^{-x}}, \quad -\infty < x < \infty \tag{11.2}$$

To transform the logistic cdf into the Logit probability model, we replace the variable x with a linear function of (possibly many) explanatory variables. For example, suppose we are trying to explain firms' decisions about whether to adopt a pollution reducing technology. This might depend, say, on (1) firm size (S), (2) pollution output (P), and (3) time since the last upgrade of technology (T). Then the logistic model would look as follows:

$$P(Adopt) = \frac{1}{1+e^{-(\beta_0+\beta_1 S+\beta_2 P+\beta_3 T)}} \tag{11.3}$$

The main alternative to the Logit model is the Probit model, which is based on the normal cdf. This is much more difficult to interpret, because there is no closed-form expression for the cdf (one needs to evaluate an integral). However, the model is straightforward to analyze using a software package such as S-Plus.

11.5 Application of Technique

Returning to our question, we specify the following Logit model to estimate the probability of a rat contracting a thyroid tumor as a function of the dose of ETU:

$$P(Tumor) = \frac{1}{1+e^{-(\beta_0+\beta_1 D)}} \tag{11.4}$$

Estimating this expression in S-Plus gives the following coefficients: $\hat{\beta}_0 = -4.344$ and $\hat{\beta}_1 = .01261$. So, the dose-response function we were after is:

$$P(Tumor) = \frac{1}{1 + e^{-(-4.344 + .01261D)}} \tag{11.5}$$

Our original question asked us to find the dose associated with the threshold probability of 0.1. With a little algebra, we can find this value (call it D^*) as follows:

$$0.1 = \frac{1}{1 + e^{-(-4.344 + .01261D^*)}} \tag{11.6}$$

$$1 + e^{-(-4.344 + .01261D^*)} = 10 \tag{11.7}$$

$$e^{-(-4.344 + .01261D^*)} = 9 \tag{11.8}$$

$$-(-4.344 + .01261D^*) = \ln 9 \tag{11.9}$$

$$-4.344 + .01261D^* = -2.197 \tag{11.10}$$

$$.01261D^* = 2.1468 \tag{11.11}$$

$$D^* = 170.24 \tag{11.12}$$

By this estimate, a dose of about 170 gives approximately a 10% chance of a rat contracting a thyroid tumor.

Chapter 12

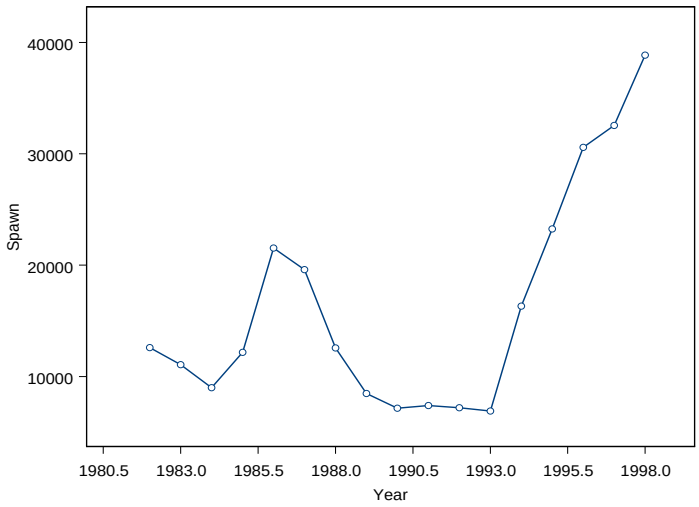
Autocorrelation

12.1 Motivation

Autocorrelation arises when correlation exists between the error terms for observations in a model. Just as heteroskedasticity is typically observed in cross-sectional data, autocorrelation is often observed in models of time series data. Autocorrelation is a violation of the Classical Linear Regression Model, so we would like to know how to correct for it. It turns out that correcting for the most common type of autocorrelation, a first order autoregressive process, is straightforward. Specifically, if we ignore autocorrelation, our OLS estimates are still unbiased, but we cannot trust our t–statistics or P–values, and therefore will not know whether we can appropriately reject null hypotheses.

Autocorrelation can arise for several reasons (and they can arise in models that do not have a time component). Some examples are (adapted from Kennedy "A Guide to Econometrics"):

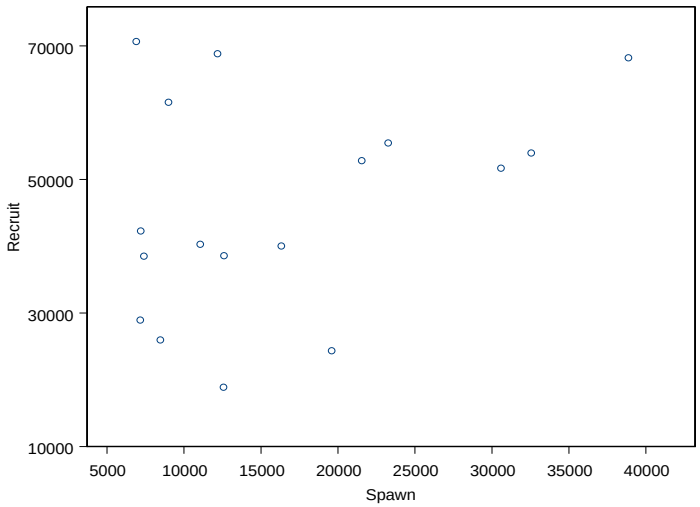
- Spatial Autocorrelation.
- Prolonged influence of shocks. This is probably the most common cause of autocorrelation in environmental data. A large shock, such as El Niño, can cause autocorrelation if its influence doesn't die out immediately.
- Data manipulation. Often published data are smoothed over time, which averages true disturbances over successive time periods.



- Misspecification. Both omitted variables (such as omitting Sea Surface Temperature in a model of a marine population) and incorrect functional form can cause autocorrelation.

12.2 Question

We are interested in modeling they dynamics of weakfish, an east coast species of fish that is becoming commercially popular. In a previous problem set, we introduced the Ricker spawner recruit model which is used to model many populations. The spawning population of weakfish over the period 1982-1998 is shown in figure 12.2, and a plot of recruits vs. spawners is shown in figure 12.2. Once we have fully estimated this model, an example of the type of question we can ask is, "What would be the effect on the population of a regulation that increased the number of spawners by X fish per year?



12.3 Estimating the Ricker Model

Recall the Ricker model that we estimated in problem set 2:

$$R_t = \alpha S_{t-1}^{\theta} e^{\beta S_{t-1}} \epsilon_t \tag{12.1}$$

The reason R has a t subscript and S has a $t-1$ subscript is that spawners last year make recruits this year (for some species this lag is more than one period).

We all know how to estimate this using S-Plus. The regression results are:

```
    *** Linear Model ***

Call: lm(formula = log(Recruit) ~ log(Spawn) + Spawn, data = weakfish,
    na.action = na.exclude)
Residuals:
       Min          1Q      Median          3Q       Max
 -0.687061 -0.09186404 -0.00862516 0.2490089 0.6037112

Coefficients:
```

```
                 Value  Std. Error     t value    Pr(>|t|)
(Intercept)  17.1509148   6.1529603   2.7874249   0.0145356
 log(Spawn)  -0.7768903   0.7133766  -1.0890325   0.2945313
      Spawn   0.0000569   0.0000412   1.3809284   0.1889480

Residual standard error: 0.3778412 on 14 degrees of freedom
Multiple R-Squared: 0.1828108
F-statistic: 1.565947 on 2 and 14 degrees of freedom, the p-value is 0.2433652
```

This model does not appear to fit very well, neither explanatory (independent) variable appears to be significant. However, considering the motivation for this lecture, you should be wondering whether autocorrelation of the residuals might be present.

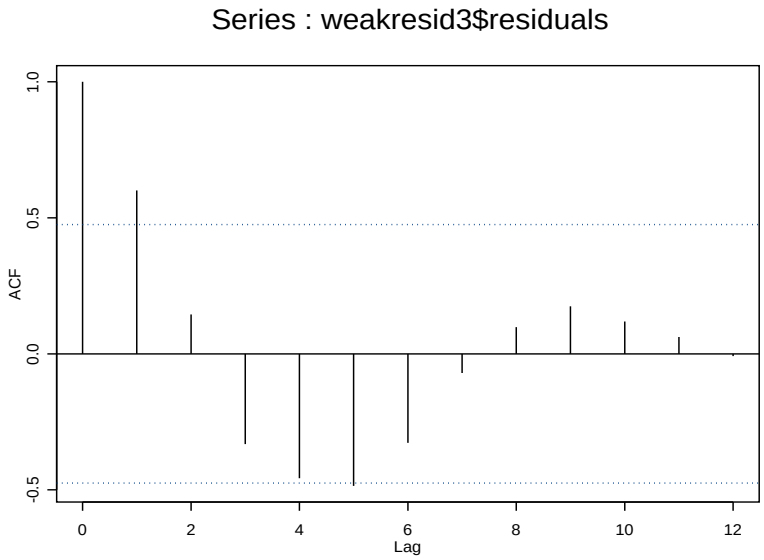
If autocorrelation is present, then we cannot trust our p-values, and we would ideally like to re-estimate the parameters of this model, taking the autocorrelation into account.

12.4 Detecting Autocorrelation

By far the most common type of autocorrelation is called the AR(1) process. This is a process where the errors are autoregressive of order 1, meaning that the error for one period is determined as a function of the error last period plus a random component. The best way to test for AR(1) errors is with a fairly straightforward statistical test called the "Durbin Watson" statistic. The DW statistics essentially looks at the residuals of your regression and tests whether successive residuals are correlated.

Since S-Plus does not perform the DW test for you, we'll just take a graphical look at the residuals to see if evidence of an AR(1) process exists. To do so, we first run our Ricker model above and save the residuals. Then we graph what is called the "Autocorrelation Function", which plots the correlation between the residual an its first lag, second lag, and so forth. S-Plus has this functionality within the "Time Series" header.

The plot of the autocorrelation function for the residuals of our weakfish Ricker model are shown in figure 12.4. S-Plus also plots the 95% confidence interval for the null hypothesis that there is no correlation between the variable and its lag. The plot suggests that we might have an AR(1) process.



If we have an AR(1) process, our model becomes:

$$R_t = \alpha S_{t-1}^{\theta} e^{\beta S_{t-1}} \epsilon_t \tag{12.2}$$

$$\epsilon_t = \rho \epsilon_{t-1} + \mu_t \tag{12.3}$$

where μ_t are independently and identically distributed. Notice that ϵ_t and ϵ_{t-1} are related via the (unknown) constant, ρ.

12.5 Correcting Autocorrelation

Once we have detected autocorrelation and we have determined its form (AR(1) in this case), it is straightforward to re-estimate our model, taking the autocorrelation into account. To do so, we use a modified version of OLS, called "Generalized Least Squares" or GLS. S-Plus performs GLS via a dropdown menu under the Statistics header. We simply need to input our model again, and tell S-Plus that the errors follow an AR(1) process. This procedure will estimate new parameters of the Ricker model, as well as providing an estimate of ρ, the AR(1) parameter.

Note first that $|\rho| \leq 1$ (otherwise, shocks could grow infinitely large). Note also that $\rho > 0$ means the errors are positively related (positive shock

followed by another positive shock), and $\rho < 0$ means they are negatively related (negative shock often followed by positive shock).

The GLS estimates (obtained by estimating the AR(1) model with Maximum Likelihood in S-Plus) are:

```
    *** Generalized Least Squares ***

Generalized least squares fit by REML
  Model: log(Recruit) ~ log(Spawn) + Spawn
  Data: weakfish
          AIC         BIC       logLik
  34.65078081 37.84606746 -12.32539041

Correlation Structure: AR(1)
 Parameter estimate(s):
          Phi
 0.7798985693

Coefficients:
                    Value   Std.Error      t-value p-value
(Intercept)  23.30617483 5.547520040  4.201188037  0.0009
 log(Spawn)  -1.47202721 0.647252081 -2.274271890  0.0392
      Spawn   0.00008687 0.000040381  2.151226751  0.0494
```

Note that the p-values for both explanatory variables are now < 0.05 and that the estimate of ρ (called Phi by S-Plus) is 0.78, suggesting fairly strong, positive first order autocorrelation.