Statistička analiza podataka:
Popis literature za učenje

UNIZG FER, ak. god. 2016./2017.

Predavanje 1: Uvod u statističku analizu podataka

- Benšić, Šuvak. *Primijenjena statistika*, Poglavlja 1 i 2
- Walpole et al. *Probability & Statistics for Engineers & Scientists*, Poglavlja 1.1 i 1.2
- Diez et al. *OpenIntro Statistics*, Poglavlja 1.2, 1.3 i 1.4.1

Predavanje 2: Deskriptivna statistika

- Benšić, Šuvak. *Primijenjena statistika*, Poglavlje 3
- Walpole et al. *Probability & Statistics for Engineers & Scientists*, Poglavlja 1.3–1.6
- Diez et al. *OpenIntro Statistics*, Poglavlje 1.6

Predavanje 3: Uvod u statističko zaključivanje

- Walpole et al. *Probability & Statistics for Engineers & Scientists*, Poglavlja 8.1–8.6, 8.9, 9.1–9.7, 9.10 i 9.12
- Diez et al. *OpenIntro Statistics*, Poglavlja 1.4 i 1.5
- Harward Statistics 110, Predavanje 29: Law of Large Numbers and Central Limit Theorem (cijelo predavanje). `https://www.youtube.com/watch?v=OprNqnHsVIA`

- Harward Statistics 110, Predavanje 30: Chi-Square, Student-t, Multivariate Normal (prvih 25 minuta). `https://www.youtube.com/watch?v=MF-XSJOsGqw`

Predavanje 4: Testiranje statističkih hipoteza

- Benšić, Šuvak. *Primijenjena statistika*, Poglavlja 5.4 i 5.4.1
- Walpole et al. *Probability & Statistics for Engineers & Scientists*, Poglavlja 10.1–10.3
- Diez et al. *OpenIntro Statistics*, Poglavlje 4.3

Savjet: pročitati cijelo 4. poglavlje koje pokriva 3. i 4. predavanje.

Predavanje 5: Statističko zaključivanje za metričke podatke

- Walpole et al. *Probability & Statistics for Engineers & Scientists*, Poglavlja 9.13, 10.4–10.7 i 10.10
- Barttlet's test. `http://www.itl.nist.gov/div898/handbook/eda/section3/eda357.htm`

Predavanje 6: Statističko zaključivanje za kategorijske podatke

- Walpole et al. *Probability & Statistics for Engineers & Scientists*, Poglavlja 10.8, 10.9 i 10.11–10.13
- Hodges et al. *Basic Concepts of Probability and Statistics*, Poglavlje 12.2

Predavanje 7: Postupci ponovnog uzorkovanja

- B. Efron & G. Gong (1983). *A leisurely look at the bootstrap, the jackknife, and cross-validation*. The American Statistician, 37(1), 36-48. `http://www.sas.rochester.edu/psc/clarke/405/EfronGong.pdf`
- Avery I. McIntosh. *The Jackknife Estimation Method*. `http://people.bu.edu/aimcinto/jackknife.pdf`
- S. Sawyer. *Resampling Data: Using a Statistical Jackknife*. `http://www.math.wustl.edu/~sawyer/handouts/Jackknife.pdf`
- K. Singh & M. Xie. *Bootstrap: A Statistical Method*. `http://stat.rutgers.edu/home/mxie/RCPapers/bootstrap.pdf`
- J. Canny. *Lecture 11*. `https://inst.eecs.berkeley.edu/~cs174/sp08/lecs/lec11/lec11.pdf`
- D. C. Howell. *Overview of Randomization Tests*. `https://tinyurl.com/ycodpo2r`

Predavanja 8 i 9: Linearna regresija

- Walpole et al. *Probability & Statistics for Engineers & Scientists*, Poglavlja 11 i 12.1–12.6
- N. Elezović. *Statistika i procesi*, str. 72 i 73
- Wikipedia. *Standardized residuals*. `https://en.wikipedia.org/wiki/Studentized_residual`

Predavanje 10: Regresija – binarna zavisna varijabla

- G. Rodriguez *Logit Models for Binary Data*. `http://data.princeton.edu/wws509/notes/c3.pdf`
- Walpole et al. *Probability & Statistics for Engineers & Scientists*, Poglavlja 12.8 i 12.9

Predavanje 11: Analiza varijance – ANOVA

- Walpole et al. *Probability & Statistics for Engineers & Scientists*, Poglavlja 13.1–13.5, 14.1–14.3

Predavanje 12: Neparametarski postupci

- Walpole et al. *Probability & Statistics for Engineers & Scientists*, Poglavlja 16.1–16.4, 16.7

Predavanje 13: Alternativni pristupi analizi podataka

- Walpole et al. *Probability & Statistics for Engineers & Scientists*, Poglavlja 18.1 i 18.2

- P. Ipeirotis. *Are you a Bayesian or a Frequentist? (Or Bayesian Statistics 101).* `https://tinyurl.com/cuyvfka`

A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation

Bradley Efron; Gail Gong

The American Statistician, Vol. 37, No. 1 (Feb., 1983), 36-48.

Stable URL:
http://links.jstor.org/sici?sici=0003-1305%28198302%2937%3A1%3C36%3AALLATB%3E2.0.CO%3B2-Q

The American Statistician is currently published by American Statistical Association.

A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation

BRADLEY EFRON and GAIL GONG*

This is an invited expository article for *The American Statistician*. It reviews the nonparametric estimation of statistical error, mainly the bias and standard error of an estimator, or the error rate of a prediction rule. The presentation is written at a relaxed mathematical level, omitting most proofs, regularity conditions, and technical details.

KEY WORDS: Bias estimation; Variance estimation; Nonparametric standard errors; Nonparametric confidence intervals; Error rate prediction.

1. INTRODUCTION

This article is intended to cover lots of ground, but at a relaxed mathematical level that omits most proofs, regularity conditions, and technical details. The ground in question is the nonparametric estimation of statistical error. "Error" here refers mainly to the bias and standard error of an estimator, or to the error rate of a data-based prediction rule.

All of the methods we discuss share some attractive properties for the statistical practitioner: they require very little in the way of modeling, assumptions, or analysis, and can be applied in an automatic way to any situation, no matter how complicated. (We will give an example of a very complicated prediction rule indeed). An important theme of what follows is the substitution of raw computing power for theoretical analysis.

The references upon which this article is based (Efron 1979a,b, 1981a,b,c, 1982; Efron and Gong 1982) explore the connections between the various nonparametric methods, and also the relationship to familiar parametric techniques. Needless to say, there is no danger of parametric statistics going out of business. A good parametric analysis, when appropriate, can be far more efficient than its nonparametric counterpart. Often, though, parametric assumptions are difficult to justify, in which case it is reassuring to have available the comparatively crude but trustworthy nonparametric answers.

What are the bootstrap, the jackknife, and cross-validation? For a quick answer, before we begin the main exposition, we consider a problem where none of the three methods are necessary, estimating the standard error of a sample average.

The data set consists of a random sample of size n from an unknown probability distribution F on the real line,

$$X_1, X_2, \ldots, X_n \sim F. \tag{1}$$

Having observed $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, we compute the sample average $\bar{x} = \sum_{i=1}^n x_i/n$ for use as an estimate of the expectation of F.

An interesting fact, and a crucial one for statistical applications, is that the data set provides more than the estimate $\bar{x}$. It also gives an estimate for the accuracy of $\bar{x}$, namely

$$\hat{\sigma} = \left[\frac{1}{n(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2\right]^{1/2}; \tag{2}$$

$\hat{\sigma}$ is the estimated standard error of $\bar{X} = \bar{x}$, the root mean squared error of estimation.

The trouble with formula (2) is that it does not, in any obvious way, extend to estimators other than $\bar{X}$, for example the sample median. The jackknife and the bootstrap are two ways of making this extension. Let

$$\bar{x}_{(i)} = \frac{n\bar{x} - x_i}{n-1} = \frac{1}{n-1} \sum_{j \neq i} x_i, \tag{3}$$

the sample average of the data set deleting the nth point. Also, let $\bar{x}_{(\cdot)} = \sum_{i=1}^n x_{(i)}/n$, the average of the deleted averages. (Actually $\bar{x}_{(\cdot)} = \bar{x}$, but we need the dot notation below.) The jackknife estimate of standard error is

$$\hat{\sigma}_J = \left[\frac{n-1}{n} \sum_{i=1}^n (\bar{x}_{(i)} - \bar{x}_{(\cdot)})^2\right]^{1/2}. \tag{4}$$

The reader can verify that this is the same as (2). The advantage of (4) is an easy generalizability to any estimator $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$. The only change is to substitute $\hat{\theta}_{(i)} = \hat{\theta}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ for $\bar{x}_{(i)}$ and $\hat{\theta}_{(\cdot)} = \sum_{i=1}^n \hat{\theta}_{(i)}/n$ for $\bar{x}_{(\cdot)}$.

The bootstrap generalizes (2) in an apparently different way. Let $\hat{F}$ be the empirical probability distribution of the data, putting probability mass $1/n$ on each x_i, and let $X_1^*, X_2^*, \ldots, X_n^*$ be a random sample from $\hat{F}$,

$$X_1^*, X_2^*, \ldots, X_n^* \sim \hat{F}. \tag{5}$$

In other words each X_i^* is drawn independently with replacement and with equal probability from the set $\{x_1, x_2, \ldots, x_n\}$. Then $\bar{X}^* = \sum_{i=1}^n X_i^*/n$ has variance

*Bradley Efron is Professor of Statistics and Biostatistics at Stanford University. Gail Gong is Assistant Professor of Statistics at Carnegie-Mellon University. The authors are grateful to Rob Tibshirani who suggested the final example in Section 7; to Samprit Chatterjee and Werner Stuetzle who suggested looking at estimators like "BootAve" in Section 9; and to Dr. Peter Gregory of the Stanford Medical School who provided the original analysis as well as the data in Section 10. This work was partially supported by the National Science Foundation and the National Institutes of Health.

$$\text{var.}\,\bar{X}^* = \frac{1}{n^2}\sum_{i=1}^{n}(x_i - \bar{x})^2, \tag{6}$$

var. indicating variance under sampling scheme (5). The bootstrap estimate of standard error for an estimator $\hat{\theta}(X_1, X_2, \ldots, X_n)$ is

$$\hat{\sigma}_B = [\text{var.}\,\hat{\theta}(X_1^*, X_2^*, \ldots, X_n^*)]^{1/2}. \tag{7}$$

Comparing (7) with (2) we see that $[n/(n-1)]^{1/2}\,\hat{\sigma}_B = \hat{\sigma}$ for $\hat{\theta} = \bar{X}$. We could make $\hat{\sigma}_B$ exactly equal $\hat{\sigma}$, for $\hat{\theta} = \bar{X}$, by adjusting definition (7) with the factor $[n/(n-1)]^{1/2}$, but there is no general advantage in doing so. A simple algorithm described in Section 2 allows the statistician to compute $\hat{\sigma}_B$ no matter how complicated $\hat{\theta}$ may be. Section 3 shows the close connection between $\hat{\sigma}_B$ and $\hat{\sigma}_J$.

Cross-validation relates to another, more difficult, problem in estimating statistical error. Going back to (1), suppose we try to predict a new observation from F, call it X_0, using the estimator $\bar{X}$ as a predictor. The expected squared error of prediction $E[X_0 - \bar{X}]^2$ equals $((n+1)/n)\mu_2$ where μ_2 is the variance of the distribution F. An unbiased estimate of $((n+1)/n)\mu_2$ is

$$(n+1)\hat{\sigma}^2. \tag{8}$$

Cross-validation is a way of obtaining nearly unbiased estimators of prediction error in much more complicated situations. The method consists of (a) deleting the points x_i from the data set one at a time; (b) recalculating the prediction rule on the basis of the remaining $n-1$ points; (c) seeing how well the recalculated rule predicts the deleted point; and (d) averaging these predictions over all n deletions of an x_i. In the simple case above, the cross-validated estimate of prediction error is

$$\frac{1}{n}\sum_{i=1}^{n}[x_i - \bar{x}_{(i)}]^2. \tag{9}$$

A little algebra shows that (9) equals (8) times $n^2/(n^2-1)$, this last factor being nearly equal to one.

The advantage of the cross-validation algorithm is that it can be applied to arbitrarily complicated prediction rules. The connection with the bootstrap and jackknife is shown in Section 9.

2. THE BOOTSTRAP

This section describes the simple idea of the bootstrap (Efron 1979a). We begin with an example. The 15 points in Figure 1 represent various entering classes at American law schools in 1973. The two coordinates for law school i are $x_i = (y_i, z_i)$,

y_i = average LSAT score of entering students at school i,

z_i = average undergraduate GPA score of entering students at school i.

(The LSAT is a national test similar to the Graduate Record Exam, while GPA refers to undergraduate grade point average.)

The observed Pearson correlation coefficient for these $n = 15$ pairs is $\hat{\rho}(x_1, x_2, \ldots, x_{15}) = .776$. We want to attach a nonparametric estimate of standard error to $\hat{\rho}$. The bootstrap idea is the following:

1. Suppose that the data points $x_1, x_2, \ldots, x_{15}$ are independent observations from some bivariate distribution F on the plane. Then the true standard error of $\hat{\rho}$ is a function of F, indicated $\sigma(F)$,

$$\sigma(F) = [\text{var}_F\,\hat{\rho}(X_1, X_2, \ldots, X_n)]^{1/2}.$$

(It is also a function of sample size n, and the functional form of the statistic $\hat{\rho}$, but both of these are known to the statistician.)

2. We don't know F, but we can estimate it by the empirical probability distribution $\hat{F}$,

$$\hat{F}\text{: mass } \frac{1}{n} \text{ on each observed data point } x_i,$$

$$i = 1, 2, \ldots, n.$$

3. The bootstrap estimate of $\sigma(F)$ is

$$\hat{\sigma}_B = \sigma(\hat{F}). \tag{10}$$

For the correlation coefficient and for most statistics, even very simple ones, the function $\sigma(F)$ is impossible to express in closed form. That is why the bootstrap is not in common use. However in these days of fast and cheap computation $\hat{\sigma}_B$ can easily be approximated by Monte Carlo methods:

(i) Construct $\hat{F}$, the empirical distribution function, as just described.

(ii) Draw a *bootstrap sample* $X_1^*, X_2^*, \ldots, X_n^*$ by independent random sampling from $\hat{F}$. In other words, make n random draws *with replacement* from $\{x_1, x_2, \ldots, x_n\}$. In the law school example a typical bootstrap sample might consist of 2 copies of point 1, 0 copies of point 2, 1 copy of point 3, and so on, the total number of copies adding up to $n = 15$. Compute the *bootstrap replication*, $\hat{\rho}^* = \hat{\rho}(X_1^*, X_2^*, \ldots, X_n^*)$, that is, the value of the statistic, in this case the correlation coefficient, evaluated for the bootstrap sample.

(iii) Do step (ii) some large number "B" of times,

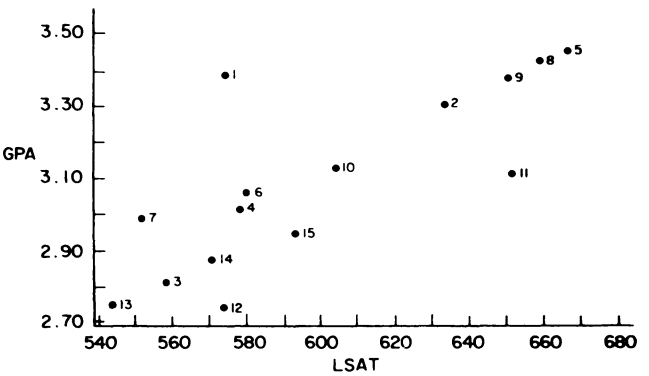


Figure 1. The law school data (Efron 1979B). The data points, beginning with School #1, are (576, 3.39), (635, 3.30), (558, 2.81), (578, 3.03), (666, 3.44), (580, 3.07), (555, 3.00), (661, 3.43), (651, 3.36), (605, 3.13), (653, 3.12), (575, 2.74), (545, 2.76), (572, 2.88), (594, 2.96).

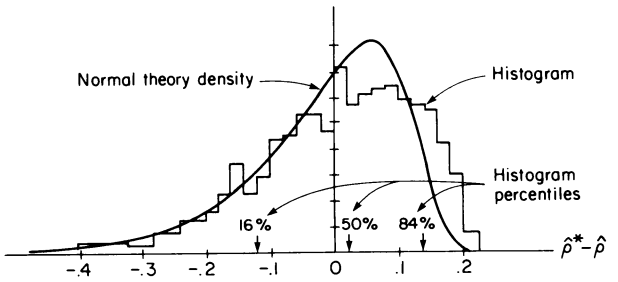


Figure 2. Histogram of B = 1000 bootstrap replications $\hat{\rho}^*$ for the law school data. The normal theory density curve has a similar shape, but falls off more quickly at the upper tail.

obtaining independent bootstrap replications $\hat{\rho}^{*1}$, $\hat{\rho}^{*2}, \ldots, \hat{\rho}^{*B}$, and approximate $\hat{\sigma}_B$ by

$$\hat{\sigma}_B = \left[\left(\sum_{b=1}^{B} (\hat{\rho}^{*b} - \hat{\rho}^{*\cdot})^2\right)\Big/(B-1)\right]^{1/2}, \quad \hat{\rho}^{*\cdot} \equiv \frac{\sum \hat{\rho}^{*b}}{B} \quad (11)$$

As $B \to \infty$, (11) approaches the original definition (10). The choice of B is further discussed below, but meanwhile we won't distinguish between (10) and (11), calling both estimates $\hat{\sigma}_B$.

Figure 2 shows $B = 1000$ bootstrap replications $\hat{\rho}^{*1}, \ldots, \hat{\rho}^{*1000}$ for the law school data. The abscissa is plotted in terms of $\hat{\rho}^* - \hat{\rho} = \hat{\rho}^* - .776$. Formula (11) gives $\hat{\sigma}_B = .127$. This can be compared with the normal theory estimate of standard error for $\hat{\rho}$, (Johnson and Kotz 1970, p. 229),

$$\hat{\sigma}_{\text{NORM}} \equiv \frac{1-\hat{\rho}^2}{\sqrt{n-3}} = .115.$$

One thing is obvious about the bootstrap procedure: it can be applied just as well to any statistic, simple or complicated, as to the correlation coefficient. In Table 1 the statistic is the 25 percent trimmed mean for a sample of size $n = 15$. The true distribution F (now defined on the line rather than on the plane) is the standard normal $\mathcal{N}(0, 1)$ for the left side of the table, or one-sided negative exponential for the right side. The true standard errors $\sigma(F)$ are .286 and .232, respectively. In both cases, $\hat{\sigma}_B$, calculated with $B = 200$ bootstrap replications, is nearly unbiased for $\sigma(F)$.

The jackknife estimate of standard error $\hat{\sigma}_J$, described in Section 3, is also nearly unbiased in both cases, but has higher variability than $\hat{\sigma}_B$, as shown by its higher coefficient of variation. The minimum possible coefficient of variation (C.V.), for a scale-invariant estimate of $\sigma(F)$, assuming full knowledge of the parametric model, is shown in brackets. In the normal case, for example, .19 is the C.V. of $[\Sigma(x_i - \bar{x})^2/14]^{1/2}$. The bootstrap estimate performs well by this standard considering its totally nonparametric character and the small sample size.

Table 1. A Sampling Experiment Comparing the Bootstrap and Jackknife Estimates of Standard Error for the 25% Trimmed Mean, Sample Size n = 15

	F Standard Normal			F Negative Exponential		
	Ave	Sd	Coeff Var	Ave	Sd	Coeff Var
Bootstrap $\hat{\sigma}_B$: ($B = 200$)	.287	.071	.25	.242	.078	.32
Jackknife $\hat{\sigma}_J$:	.280	.084	.30	.224	.085	.38
True : [Minimum C.V.]	.286		[.19]	.232		[.27]

Table 2 returns to the case of $\hat{\rho}$, the correlation coefficient. Instead of real data we have a sampling experiment in which F is bivariate normal, true correlation $\rho = .5$, and the sample size is $n = 14$. The left side of Table 2 refers to $\hat{\rho}$, while the right side refers to the statistic $\hat{\phi} = \tanh^{-1} \hat{\rho} = .5 \log(1+\hat{\rho})/(1-\hat{\rho})$. For each estimator $\hat{\sigma}$, the root mean squared error of estimation $[E(\hat{\sigma} - \sigma)^2]^{1/2}$ is given in the column headed $\sqrt{\text{MSE}}$.

The bootstrap was run with $B = 128$ and $B = 512$, the latter value yielding only slightly better estimates $\hat{\sigma}_B$. Further increasing B would be pointless. It can be shown that $B = \infty$ would give $\sqrt{\text{MSE}} = .063$ in the $\hat{\rho}$ case, only .001 less than using $B = 512$. As a point of comparison, the normal theory estimate for the standard error of $\hat{\rho}$, $\hat{\sigma}_{\text{NORM}} = (1-\hat{\rho}^2)/(n-3)^{1/2}$, has $\sqrt{\text{MSE}} = .056$.

Why not generate the bootstrap observations from an estimate of $\hat{F}$ which is smoother than F? This is done in lines 3, 4, and 5 of Table 2. Let $\hat{\boldsymbol{\Sigma}} = \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})'/n$ be the sample covariance matrix of the observed data. The *normal smoothed bootstrap* draws the bootstrap sample $X_1^*, X_2^*, \ldots, X_n^*$ from $\hat{F} \oplus \mathcal{N}_2(0, .25\hat{\boldsymbol{\Sigma}})$, $\oplus$ indicating convolution. This amounts to estimating F by an equal mixture of the n distributions $\mathcal{N}_2(x_i, .25\hat{\boldsymbol{\Sigma}})$, that is by a normal window estimate. Smoothing makes little difference on the left side of the table, but is spectacularly effective in the $\hat{\phi}$ case. The latter result is suspect since the true sampling distribution is bivariate normal, and the function $\hat{\phi} = \tanh^{-1} \hat{\rho}$ is specifically chosen to have nearly constant standard error in the bivariate-normal family. The *uniform smoothed bootstrap* samples $X_1^*, \ldots, X_n^*$ from $\hat{F} \oplus \mathcal{U}(0, .25\hat{\boldsymbol{\Sigma}})$, where $\mathcal{U}(0, .25\hat{\boldsymbol{\Sigma}})$ is the uniform distribution on a rhombus selected so $\mathcal{U}$ has mean vector 0 and covariance matrix $.25\hat{\boldsymbol{\Sigma}}$. It yields moderate reductions in $\sqrt{\text{MSE}}$ for both sides of the table.

The standard normal-theory estimates of line 8, Table 2, are themselves bootstrap estimates, carried out in a parametric framework. The bootstrap sample $X_1^*, \ldots, X_n^*$ is drawn from the parametric maximum likelihood distribution

$$\hat{F}_{\text{NORM}} \sim \mathcal{N}_2(\bar{x}, \hat{\boldsymbol{\Sigma}}),$$

rather than the nonparametric maximum likelihood distribution $\hat{F}$, and with only this change the bootstrap algorithm proceeds as previously described. In practice the bootstrap process is not actually carried out. If it were, and if $B \to \infty$, then a high-order Taylor series analysis shows that $\hat{\sigma}_B$ would equal approximately $(1-\hat{\rho}^2)/(n-3)^{1/2}$, the formula actually used to compute line 8 for the $\hat{\rho}$ side of Table 2. Notice that the normal

Table 2. Estimates of Standard Error for the Correlation Coefficient $\hat{\rho}$ and for $\hat{\phi} = \tanh^{-1} \hat{\rho}$; Sample Size $n = 14$, Distribution F Bivariate Normal With True Correlation $\rho = .5$. From a Larger Table in Efron (1981b)

	Summary Statistics for 200 Trials							
	Standard Error Estimates for $\hat{\rho}$				Standard Error Estimates for $\hat{\phi}$			
	Ave	Std Dev	CV	$\sqrt{MSE}$	Ave	Std Dev	CV	$\sqrt{MSE}$
1. Bootstrap B = 128	.206	.066	.32	.067	.301	.065	.22	.065
2. Bootstrap B = 512	.206	.063	.31	.064	.301	.062	.21	.062
3. Normal Smoothed Bootstrap B = 128	.200	.060	.30	.063	.296	.041	.14	.041
4. Uniform Smoothed Bootstrap B = 128	.205	.061	.30	.062	.298	.058	.19	.058
5. Uniform Smoothed Bootstrap B = 512	.205	.059	.29	.060	.296	.052	.18	.052
6. Jackknife	.223	.085	.38	.085	.314	.090	.29	.091
7. Delta Method (Infinitesimal Jackknife)	.175	.058	.33	.072	.244	.052	.21	.076
8. Normal Theory	.217	.056	.26	.056	.302	0	0	.003
True Standard Error	.218				.299			

smoothed bootstrap can be thought of as a compromise between using $\hat{F}$ and $\hat{F}_{\text{NORM}}$ to begin the bootstrap process.

3. THE JACKKNIFE

The jackknife estimate of standard error was introduced by Tukey in 1958 (see Miller 1974). Let $\hat{\rho}_{(i)} = \hat{\rho}(x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ be the value of the statistic when x_i is deleted from the data set, and let $\hat{\rho}_{(\cdot)} = (1/n) \sum_{i=1}^{n} \hat{\rho}_{(i)}$. The jackknife formula is

$$\hat{\sigma}_J = \left[((n-1)/n) \sum_{i=1}^{n} (\hat{\rho}_{(i)} - \hat{\rho}_{(\cdot)})^2 \right]^{1/2}.$$

Like the bootstrap, the jackknife can be applied to any statistic that is a function of n independent and identically distributed variables. It performs less well than the bootstrap in Tables 1 and 2, and in most cases investigated by the author (see Efron 1982), but requires less computation. In fact the two methods are closely related, which we shall now show.

Suppose the statistic of interest, which we will now call $\hat{\theta}(x_1, x_2, \ldots, x_n)$, is of *functional form*: $\hat{\theta} = \theta(\hat{F})$, where $\theta(F)$ is a functional assigning a real number to any distribution F on the sample space. Both examples in Section 2 are of this form. Let $\mathbf{P} = (P_1, P_2, \ldots, P_n)$ be a probability vector having nonnegative weights summing to one, and define the reweighted empirical distribution $\hat{F}(\mathbf{P})$: mass P_i on x_i, $i = 1, 2, \ldots, n$. Corresponding to $\mathbf{P}$ is a *resampled value* of the statistic of interest, say $\hat{\theta}(\mathbf{P}) = \theta(\hat{F}(\mathbf{P}))$. The shorthand notation $\hat{\theta}(\mathbf{P})$ assumes that the data points $x_1, x_2, \ldots, x_n$ are fixed at their observed values.

Another way to describe the bootstrap estimate $\hat{\sigma}_B$ is as follows. Let $\mathbf{P}^*$ indicate a vector drawn from the rescaled multinomial distribution

$$\mathbf{P}^* \sim \text{Mult}_n(n, \mathbf{P}^o)/n, \ (\mathbf{P}^o \equiv (1/n)\ (1, 1, \ldots, 1)'), \quad (12)$$

meaning the observed proportions from n random draws on n categories, with equal probability $1/n$ for each category. Then

$$\hat{\sigma}_B = [\text{var}_* \hat{\theta}(\mathbf{P}^*)]^{1/2}, \quad (13)$$

where var_* indicates variance under distribution (12). (This is true because we can take $P_i^* = \#\{X_j^* = x_i\}/n$ in step 2 of the bootstrap algorithm.)

Figure 3 illustrates the situation for the case $n = 3$. There are 10 possible bootstrap points. For example, the point $\mathbf{P}^* = (\frac{2}{3}, \frac{1}{3}, 0)'$ is the second dot from the left on the lower side of the triangle, and occurs with bootstrap probability $\frac{1}{9}$, under (12). It indicates a bootstrap sample X_1^*, X_2^*, X_3^* consisting of two x_1's and one x_2. The center point $\mathbf{P}^o = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})'$ has bootstrap probability $\frac{2}{9}$.

The jackknife resamples the statistic at the n points

$$\mathbf{P}_{(i)} = (1/(n-1))\ (1, 1, \ldots, 1, 0, 1, \ldots, 1)' \quad (0 \text{ in } i\text{th place}),$$

$i = 1, 2, \ldots, n$. These are indicated by the open circles in Figure 3. In general there are n jackknife points, compared with $\binom{2n-1}{n}$ bootstrap points.

The trouble with bootstrap formula (13) is that $\hat{\theta}(\mathbf{P})$ is usually a complicated function of $\mathbf{P}$ (think of the examples in Sec. 2), and so $\text{var}_* \hat{\theta}(\mathbf{P}^*)$ cannot be evalu-

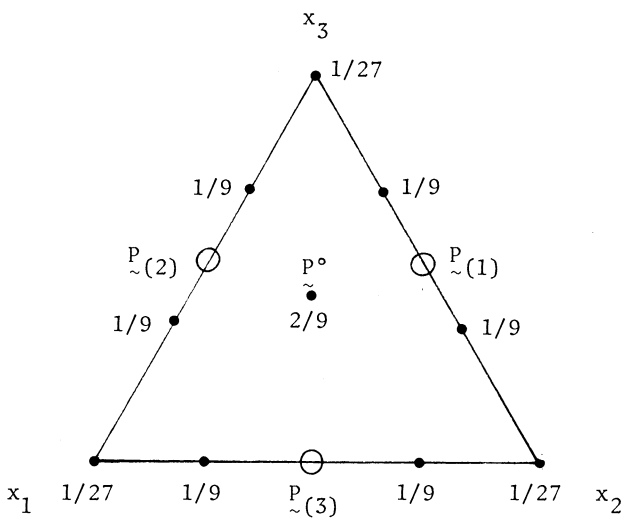


Figure 3. The bootstrap and jackknife sampling points in the case $n = 3$. The bootstrap points ($\cdot$) are shown with their probabilities.

ated except by Monte Carlo methods. The jackknife trick approximates $\hat{\theta}(\mathbf{P})$ by a linear function of $\mathbf{P}$, say $\hat{\theta}_L(\mathbf{P})$, and then uses the known covariance structure of (12) to evaluate var. $\hat{\theta}_L(\mathbf{P}^*)$. The approximator $\hat{\theta}_L(\mathbf{P})$ is chosen to match $\hat{\theta}(\mathbf{P})$ at the n points $\mathbf{P} = \mathbf{P}_{(i)}$. It is not hard to see that

$$\hat{\theta}_L(\mathbf{P}) = \hat{\theta}_{(\cdot)} + (\mathbf{P} - \mathbf{P}^o)'\mathbf{U} \tag{14}$$

where $\hat{\theta}_{(\cdot)} = (1/n)\ \Sigma\ \hat{\theta}_{(i)} = (1/n)\ \Sigma\ \hat{\theta}(\mathbf{P}_{(i)})$, and $\mathbf{U}$ is a column vector with coordinates $U_i = (n-1)(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)})$.

Theorem. The jackknife estimate of standard error equals

$$\hat{\sigma}_J = \left[\frac{n}{n-1}\ \text{var.}\ \hat{\theta}_L(\mathbf{P}^*)\right]^{1/2},$$

which is $[n/(n-1)]^{1/2}$ times the bootstrap estimate of standard error for $\hat{\theta}_L$ (Efron 1982).

In other words the jackknife is, almost,[1] a bootstrap itself. The advantage of working with $\hat{\theta}_L$ rather than $\hat{\theta}$ is that there is no need for Monte Carlo: var. $\hat{\theta}_L(\mathbf{P}^*) = \text{var.}\ (\mathbf{P}^* - \mathbf{P}^o)'\mathbf{U} = \Sigma U_i^2/n^2$, using the covariance matrix for (12) and the fact that $\Sigma U_i = 0$. The disadvantage is (usually) increased error of estimation, as seen in Tables 1 and 2.

The fact that $\hat{\sigma}_J$ is almost $\hat{\sigma}_B$ for a linear approximation of $\hat{\theta}$ does not mean that $\hat{\sigma}_J$ is a reasonable approximation for the actual $\hat{\sigma}_B$. That depends on how well $\hat{\theta}_L$ approximates $\hat{\theta}$. In the case where $\hat{\theta}$ is the sample median, for instance, the approximation is very poor.

4. THE DELTA METHOD, INFLUENCE FUNCTIONS, AND THE INFINITESIMAL JACKKNIFE

There is a more obvious linear approximation to $\hat{\theta}(\mathbf{P})$ than $\hat{\theta}_L(\mathbf{P})$, (14). Why not use the first-order Taylor series expansion for $\hat{\theta}(\mathbf{P})$ about the point $\mathbf{P} = \mathbf{P}^o$? This is the idea of Jaeckel's *infinitesimal jackknife* (1972). The Taylor series approximation turns out to be

$$\hat{\theta}_T(\mathbf{P}) = \hat{\theta}(\mathbf{P}^o) + (\mathbf{P} - \mathbf{P}^o)'\mathbf{U}^o,$$

where

$$U_i^o = \lim_{\varepsilon \to 0} \frac{\hat{\theta}((1-\varepsilon)\mathbf{P}^o + \varepsilon\boldsymbol{\delta}_i) - \hat{\theta}(\mathbf{P}^o)}{\varepsilon},$$

$\boldsymbol{\delta}_i$ being the ith coordinate vector. This suggests the infinitesimal jackknife estimate of standard error

$$\hat{\sigma}_{IJ} = [\text{var.}\ \hat{\theta}_T(\mathbf{P}^*)]^{1/2} = [\Sigma U_i^{o\,2}/n^2]^{1/2}, \tag{15}$$

with var. still indicating variance under (12). The ordinary jackknife can be thought of as taking $\varepsilon = -1/(n-1)$ in the definition of U_i^o, while the infinitesimal jackknife lets $\varepsilon \to 0$, thereby earning the name.

The U_i^o are values of what Mallows (1974) calls the *empirical influence function*. Their definition is a nonparametric estimate of the true influence function

$$IF(x) = \lim_{\varepsilon \to 0} \frac{\theta((1-\varepsilon)F + \varepsilon\delta_x) - \theta(F)}{\varepsilon},$$

δ_x being the degenerate distribution putting mass 1 on x. The right side of (15) is then the obvious estimate of the influence function approximation to the standard error of $\hat{\theta}$, (Hampel 1974), $\sigma(F) \doteq [\int IF^2(x)dF(x)/n]^{1/2}$. The empirical influence function method and the infinitesimal jackknife give identical estimates of standard error.

How have statisticians gotten along for so many years without methods like the jackknife or the bootstrap? The answer is the *delta method*, which is still the most commonly used device for approximating standard errors. The method applies to statistics of the form $t(\overline{Q}_1, \overline{Q}_2, \ldots, \overline{Q}_A)$, where $t(\cdot, \cdot, \ldots, \cdot)$ is a known function and each $\overline{Q}_a$ is an observed average, $\overline{Q}_a = \sum_{i=1}^n Q_a(X_i)/n$. For example, the correlation $\hat{\rho}$ is a function of $A = 5$ such averages: the average of the first coordinate values, the second coordinates, the first coordinates squared, the second coordinates squared, and the cross-products.

In its nonparametric formulation, the delta method works by (a) expanding t in a linear Taylor series about the expectations of the $\overline{Q}_a$; (b) evaluating the standard error of the Taylor series using the usual expressions for variances and covariances of averages; and (c) substituting $\gamma(\hat{F})$ for any unknown quantity $\gamma(F)$ occurring in (b). For example, the nonparametric delta method estimates the standard error of $\hat{\rho}$ by

$$\left\{\frac{\hat{\rho}^2}{4n}\left[\frac{\hat{\mu}_{40}}{\hat{\mu}_{20}^2} + \frac{\hat{\mu}_{04}}{\hat{\mu}_{02}^2} + \frac{2\hat{\mu}_{22}}{\hat{\mu}_{20}\hat{\mu}_{02}} + \frac{4\hat{\mu}_{22}}{\hat{\mu}_{11}^2} - \frac{4\hat{\mu}_{31}}{\hat{\mu}_{11}\hat{\mu}_{02}} - \frac{4\hat{\mu}_{13}}{\hat{\mu}_{11}\hat{\mu}_{02}}\right]\right\}^{1/2}$$

where, in terms of $x_i = (y_i,\ z_i)$, $\hat{\mu}_{gh} \equiv \Sigma(y_i - \bar{y})^g (z_i - \bar{z})^h/n$ (Cramér 1946, p. 359).

Theorem. For statistics of the form $\hat{\theta} = t(\overline{Q}_1, \ldots, \overline{Q}_A)$, the nonparametric delta method and the infinitesimal jackknife give the same estimate of standard error (Efron 1981b).

The infinitesimal jackknife, the delta method, and the empirical influence function approach are three names for the same method. *Notice that the results reported in line 7 of Table 2 show a severe downward bias.* Efron and Stein (1981) show that the ordinary jackknife is always biased upwards, in a sense made precise in that paper. In the authors' opinion the ordinary jackknife is the method of choice if one does not want to do the bootstrap computations.

5. NONPARAMETRIC CONFIDENCE INTERVALS

In applied work, the usual purpose of estimating a standard error is to set confidence intervals for the un-

[1]The factor $[n/(n-1)]^{1/2}$ makes $\hat{\sigma}_J^2$ unbiased for σ^2 if $\hat{\theta}$ is a linear statistic, e.g., $\hat{\theta} = \bar{X}$. We could multiply $\hat{\sigma}_B$ by this same factor, and achieve the same unbiasedness, but there doesn't seem to be any general advantage to doing so.

known paramater. These are typically of the crude form $\hat{\theta} \pm z_\alpha \hat{\sigma}$, with z_α being the $100(1-\alpha)$ percentile point of a standard normal distribution. We can, and do, use the bootstrap and jackknife estimates $\hat{\sigma}_B$, $\hat{\sigma}_J$ in this way. However in small-sample parametric situations, where we can do exact calculations, confidence intervals are often highly asymmetric about the best point estimate $\hat{\theta}$. This asymmetry, which is $0(1/\sqrt{n})$ in magnitude, is substantially more important than the Student's t correction (replacing $\hat{\theta} \pm z_\alpha \hat{\sigma}$ by $\hat{\theta} \pm t_\alpha \hat{\sigma}$, with t_α the $100(1-\alpha)$ percentile point of the appropriate t distribution), which is only $0(1/n)$. This section discusses some nonparametric methods of assigning confidence intervals, which attempt to capture the correct asymmetry. It is abbreviated from a longer discussion in Efron (1981c), and also Chapter 10 of Efron (1982). All of this work is highly speculative, though encouraging.

We return to the law school example of Section 2. Suppose for the moment that we believe the data come from a bivariate normal distribution. The standard 68 percent central confidence interval (i.e., $\alpha = .16$, $1 - 2\alpha = .68$) for ρ in this case is $[.62, .87] = [\hat{\rho} - .16, \hat{\rho} + .09]$, obtained by inverting the approximation $\hat{\phi} \sim \mathcal{N}(\phi + \rho/(2(n-1)), 1/(n-3))$. Compared to the crude interval $\hat{\rho} \pm z_{.16}\, \hat{\sigma}_{\text{NORM}} = \hat{\rho} \pm \hat{\sigma}_{\text{NORM}} = [\hat{\rho} - .12, \hat{\rho} + .12]$, this demonstrates the magnitude of the asymmetry effect described previously.

The asymmetry of the confidence interval $[\hat{\rho} - .16, \hat{\rho} + .09]$ relates to the asymmetry of the normal-theory density curve for $\hat{\rho}$, as shown in Figure 2. The bootstrap histogram shows this same asymmetry. The striking similarity between the histogram and the density curve suggests that we can use the bootstrap results more ambitiously than simply to compute $\hat{\sigma}_B$.

Two ways of forming nonparametric confidence intervals from the bootstrap histogram are discussed in Efron (1981c). The first, called the *percentile method,* uses the 100α and $100(1-\alpha)$ percentiles of the bootstrap histogram, say

$$\theta \in [\hat{\theta}(\alpha), \hat{\theta}(1-\alpha)], \tag{16}$$

as a putative $1 - 2\alpha$ central confidence interval for the unknown parameter θ. Letting

$$\hat{C}(t) \equiv \frac{\#\{\hat{\theta}^{*b} < t\}}{B},$$

then $\hat{\theta}(\alpha) = \hat{C}^{-1}(\alpha)$, $\hat{\theta}(1-\alpha) = \hat{C}^{-1}(1-\alpha)$. In the law school example, with $B = 1000$ and $\alpha = .16$, the 68 percent interval is $\rho \in [.65, .91] = [\hat{\rho} - .12, \hat{\rho} + .13]$, almost exactly the same as the crude normal-theory interval $\hat{\rho} \pm \hat{\sigma}_{\text{NORM}}$.

Notice that the median of the bootstrap histogram is substantially higher than $\hat{\rho}$ in Figure 2. In fact, $\hat{C}(\hat{\rho}) = .433$, only 433 out of 1000 bootstrap replications having $\hat{\rho}^* < \hat{\rho}$. The *bias-corrected percentile method* makes an adjustment for this type of bias. Let $\Phi(z)$ indicate the CDF of the standard normal distribution, so $\Phi(z_\alpha) = 1 - \alpha$, and define

$$z_0 \equiv \Phi^{-1}\{\hat{C}(\hat{\theta})\}.$$

The bias-corrected putative $1 - 2\alpha$ central confidence interval is defined to be

$$\theta \in [\hat{C}^{-1}\{\Phi(2z_0 - z_\alpha)\},\ \hat{C}^{-1}\{\Phi(2z_0 + z_\alpha)\}]. \tag{17}$$

If $\hat{C}(\hat{\theta}) = .50$, the median unbiased case, then $z_0 = 0$ and (8) reduce to the uncorrected percentile interval (16). Otherwise the results can be quite different. In the law school example $z_0 = \Phi(.433) = -.17$, and for $\alpha = .16$, (8) gives $\rho \in [\hat{C}^{-1}\{\Phi(-1.34)\},\ \hat{C}^{-1}\{\Phi(.66)\}] = [\hat{\rho} - .17, \hat{\rho} + .10]$. This agrees nicely with the normal-theory interval $[\hat{\rho} - .16, \hat{\rho} + .09]$.

Table 3 shows the results of a small sampling experiment, only 10 trials, in which the true distribution F was bivariate normal, $\rho = .5$. The bias-corrected percentile method shows impressive agreement with the normal-theory intervals. Even better are the smoothed intervals, last column. Here the bootstrap replications were obtained by sampling from $\hat{F} \oplus \mathcal{N}(0, .25\hat{\Sigma})$, as in line 3 of Table 2, and then applying (17) to the resulting histogram.

There are some theoretical arguments supporting (16) and (17). If there exists a normalizing transformation, in the same sense as $\hat{\phi} = \tanh^{-1} \hat{\rho}$ is normalizing for the correlation coefficient under bivariate-normal sampling, then the bias-corrected percentile method automatically produces the appropriate confidence intervals. This is interesting since we do not have to know the form of the normalizing transformation to apply (17). Bayesian and frequentist justifications are given also in Efron (1981c). None of these arguments is overwhelming, and in fact (17) and (16) sometimes perform poorly. Some other methods are suggested in Efron (1981c), but the appropriate theory is still far from clear.

6. BIAS ESTIMATION

Quenouille (1949) originally introduced the jackknife as a nonparametric device for estimating bias. Let us denote the bias of a functional statistic $\hat{\theta} = \theta(\hat{F})$ by

Table 3. Central 68% Confidence Intervals for ρ, 10 Trials of $X_1, X_2, \ldots, X_{15}$ Bivariate Normal With True $\rho = .5$. Each Interval Has $\hat{\rho}$ Subtracted From Both Endpoints

Trial	$\hat{\rho}$	Normal Theory	Percentile Method	Bias-Corrected Percentile Method	Smoothed and Bias-Corrected Percentile Method
1	.16	(−.29, .26)	(−.29, .24)	(−.28, .25)	(−.28, .24)
2	.75	(−.17, .09)	(−.05, .08)	(−.13, .04)	(−.12, .08)
3	.55	(−.25, .16)	(−.24, .16)	(−.34, .12)	(−.27, .15)
4	.53	(−.26, .17)	(−.16, .16)	(−.19, .13)	(−.21, .16)
5	.73	(−.18, .10)	(−.12, .14)	(−.16, .10)	(−.20, .10)
6	.50	(−.26, .18)	(−.18, .18)	(−.22, .15)	(−.26, .14)
7	.70	(−.20, .11)	(−.17, .12)	(−.21, .10)	(−.18, .11)
8	.30	(−.29, .23)	(−.29, .25)	(−.33, .24)	(−.29, .25)
9	.33	(−.29, .22)	(−.36, .24)	(−.30, .27)	(−.30, .26)
10	.22	(−.29, .24)	(−.50, .34)	(−.48, .36)	(−.38, .34)
AVE	.48	(−.25, .18)	(−.21, .19)	(−.26, .18)	(−.25, .18)

β, $\beta = E\{\theta(\hat{F}) - \theta(F)\}$. In the notation of Section 3, Quenouille's estimate is

$$\hat{\beta}_J = (n-1)(\hat{\theta}_{(\cdot)} - \hat{\theta}). \tag{18}$$

Subtracting $\hat{\beta}_J$ from $\hat{\theta}$, to correct the bias leads to the *jackknife estimate* of θ, $\hat{\theta}_J = n\hat{\theta} - (n-1)\hat{\theta}_{(\cdot)}$, see Miller (1974), and also Schucany, Gray, and Owen (1971).

There are many ways to justify (18). Here we follow the same line of argument as in the justification of $\hat{\sigma}_J$. The bootstrap estimate of β, which has an obvious motivation, is introduced, and then (18) is related to the bootstrap estimate by a Taylor series argument.

The bias can be thought of as a function of the unknown probability distribution F, $\beta = \beta(F)$. The bootstrap estimate of bias is simply

$$\hat{\beta}_B = \beta(\hat{F}) = E_*\{\theta(\hat{F}^*) - \theta(\hat{F})\}. \tag{19}$$

Here E_* indicates expectation with respect to bootstrap sampling, and $\hat{F}^*$ is the empirical distribution of the bootstrap sample.

In practice $\hat{\beta}_B$ must be approximated by Monte Carlo methods. The only change in the algorithm described in Section 2 is at step (iii), when instead of (or in addition to) $\hat{\sigma}_B$ we calculate

$$\hat{\beta}_B = \hat{\theta}^{*\cdot} - \hat{\theta} = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}^{*b} - \hat{\theta}).$$

In the sampling experiment of Table 2 the true bias, of $\hat{\rho}$ for estimating ρ, is $\beta = -.014$. The bootstrap estimate $\hat{\beta}_B$, taking $B = 128$, has expectation $-.014$ and standard deviation .031 in this case, while $\hat{\beta}_J$ has expectation $-.017$, standard deviation .040. Bias is a negligible source of statistical error in this situation compared with variability. In applications this is usually made clear by comparison of $\hat{\beta}_B$ with $\hat{\sigma}_B$.

The estimates (18) and (19) are closely related to each other. The argument is the same as in Section 3, except that we approximate $\hat{\theta}(\mathbf{P})$ with a quadratic rather than a linear function of P, say $\hat{\theta}_Q(\mathbf{P}) = a + (\mathbf{P} - \mathbf{P}^o)'\mathbf{b} + \frac{1}{2}(\mathbf{P} - \mathbf{P}^o)'\mathbf{c}(\mathbf{P} - \mathbf{P}^o)$. Let $\hat{\theta}_Q(\mathbf{P})$ be any such quadratic satisfying

$$\hat{\theta}_Q(\mathbf{P}^o) = \hat{\theta}(\mathbf{P}^o) = \hat{\theta} \text{ and } \hat{\theta}_Q(\mathbf{P}_{(i)}) = \hat{\theta}(\mathbf{P}_{(i)}),\ i = 1, 2, \ldots, n.$$

Theorem. The jackknife estimate of bias equals

$$\hat{\beta}_J = \frac{n}{n-1}[E_*\{\hat{\theta}_Q(\mathbf{P}^*) - \hat{\theta}\}],$$

which is $n/(n-1)$ times the bootstrap estimate of bias for $\hat{\theta}_Q$ (Efron 1982).

Once again, the jackknife is, almost, a bootstrap estimate itself, except applied to a convenient approximation of $\hat{\theta}(\mathbf{P})$.

More general problems. There is nothing special about bias and standard error as far as the bootstrap is concerned. The bootstrap procedure can be applied to almost any estimation problem.

Suppose that $R(X_1, X_2, \ldots, X_n; F)$ is a random variable, and we are interested in estimating some aspect of R's distribution. (So far we have taken $R = \theta(\hat{F}) - \theta(F)$ and have been interested in the expectation β and the standard deviation σ of R.) The bootstrap algorithm proceeds as described in Section 2, with these two changes: at step (ii), we calculate the bootstrap replication $R^* = R(X_1^*, X_2^*, \ldots, X_n^*; \hat{F})$, and at step (iii) we calculate the distributional property of interest from the empirical distribution of the bootstrap replications $R^{*1}, R^{*2}, \ldots, R^{*B}$.

For example, we might be interested in the probability that the usual t statistic $\sqrt{n}(\bar{X} - \mu)/S$ exceeds 2, where $\mu = E\{X\}$ and $S^2 = \Sigma(X_i - \bar{X})^2/(n-1)$. Then $R^* = \sqrt{n}(\bar{X}^* - \bar{x})/S^*$, and the bootstrap estimate is $\#\{R^{*b} > 2\}/B$. This calculation is used in Section 9 of Efron (1981c) to get confidence intervals for the mean μ in a situation where normality is suspect.

The cross-validation problem of Sections 8 and 9 involves a different type of error random variable R. It will be useful there to use a jackknife-type approximation to the bootstrap expectation of R,

$$E_*\{R^*\} \doteq R^o + (n-1)(R_{(\cdot)} - R^o). \tag{20}$$

Here $R^o = R(x_1, x_2, \ldots, x_n; \hat{F})$ and $R_{(\cdot)} = (1/n)\Sigma R_{(i)}$, $R_{(i)} = R(x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n; \hat{F})$. The justification of (20) is the same as for the theorem of this section, being based on a quadratic approximation formula.

7. MORE COMPLICATED DATA SETS

So far we have considered the simplest kind of data sets, where all the observations come from the same distribution F. The bootstrap idea, and jackknife-type approximations (which are not discussed here), can be applied to much more complicated situations. We begin with a two-sample problem.

The data in our first example consist of two independent random samples,

$$X_1, X_2, \ldots, X_m \sim F \text{ and } Y_1, Y_2, \ldots, Y_n \sim G,$$

F and G being two possibly different distributions on the real line. The statistic of interest is the Hodges-Lehmann shift estimate

$$\hat{\theta} = \text{median}\{y_j - x_i;\ i = 1, \ldots, m,\ j = 1, \ldots, n\}.$$

We desire an estimate of the standard error $\sigma(F, G)$.

The bootstrap estimate is simply

$$\hat{\sigma}_B = \sigma(\hat{F}, \hat{G}),$$

$\hat{G}$ being the empirical distribution of the y_i. This is evaluated by Monte Carlo, as in Section 3, with obvious modifications: a bootstrap sample now consists of a random sample $X_1^*, X_2^*, \ldots, X_m^*$ drawn from $\hat{F}$ *and* an independent random sample $Y_1^*, \ldots, Y_n^*$ drawn from $\hat{G}$. (In other words, m draws with replacement from $\{x_1, x_2, \ldots, x_m\}$, and n draws with replacement from $\{y_1, y_2, \ldots, y_n\}$.) The bootstrap replication $\hat{\theta}^*$ is the median of the mn differences $Y_j^* - X_i^*$. Then $\hat{\sigma}_B$ is approximated from B independent such replications as on the right side of (11).

Table 4 shows the results of a sampling experiment in

Table 4. Bootstrap Estimates of Standard Error for the Hodges-Lehmann Two-Sample Shift Estimate; $m = 6$, $n = 9$; True Distributions Both F and G Uniform [0, 1]

		Expectation	St. Dev.	C.V.	$\sqrt{MSE}$
Separate	B = 100	.165	.030	.18	.030
	B = 200	.166	.031	.19	.031
Combined	B = 100	.145	.028	.19	.036
	B = 200	.149	.025	.17	.031
True Standard Error		.167			

which $m = 6$, $n = 9$, and both F and G were uniform distributions on the interval [0, 1]. The table is based on 100 trials of the situation. The true standard error is $\sigma(F, G) = .167$. "Separate" refers to $\hat{\sigma}_B$ calculated exactly as described in the previous paragraph. The improvement in going from $B = 100$ to $B = 200$ is too small to show up in the table.

"Combined" refers to the following idea: suppose we believe that G is really a translate of F. Then it wastes information to estimate F and G separately. Instead we can form the combined empirical distribution

$$\hat{H}: \text{mass } \frac{1}{m+n} \text{ on}$$

$$x_1, x_2, \ldots, x_m, y_1 - \hat{\theta}, y_2 - \hat{\theta}, \ldots, y_n - \hat{\theta}.$$

All $m + n$ bootstrap variates $X_1^*, \ldots, X_m^*, Y_1^*, \ldots, Y_n^*$ are then sampled independently from $\hat{H}$. (We could add $\hat{\theta}$ back to the Y_j^* values, but this has no effect on the bootstrap standard error estimate, since it just adds the constant $\hat{\theta}$ to each bootstrap replication $\hat{\theta}^*$.)

The combined method gives no improvement here, but it might be valuable in a many-sample problem where there are small numbers of observations in each sample, a situation that arises in stratified sampling. (See Efron 1982, Ch. 8.) The main point here is that "bootstrap" is not a well-defined verb, and that there may be more than one way to proceed in complicated situations. Next we consider regression problems, where again there is a choice of bootstrapping methods.

In a typical regression problem we observe n independent real-valued quantitives $Y_i = y_i$,

$$Y_i = g_i(\beta) + \varepsilon_i, \; i = 1, 2, \ldots, n. \tag{21}$$

The functions $g_i(\cdot)$ are of known form, usually $g_i(\beta) = g(\beta; t_i)$, where t_i is an observed **p**-dimensional vector of covariates; β is a vector of unknown parameters we wish to estimate. The ε_i are an independent and identically distributed random sample from some distribution F on the real line,

$$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n \sim F,$$

where F is assumed to be centered at zero in some sense, perhaps $E\{\varepsilon\} = 0$ or $\text{Prob}\{\varepsilon < 0\} = 0.5$.

Having observed the data vector $\mathbf{Y} = \mathbf{y} = (y_1, \ldots, y_n)$, we estimate β by minimizing some measure of distance between $\mathbf{y}$ and the vector of predicted values $\boldsymbol{\eta}(\beta) = (g_1(\beta), \ldots, g_n(\beta))$,

$$\hat{\beta}: \min_{\beta} D(\mathbf{y}, \boldsymbol{\eta}(\beta)).$$

The most common choice of D is $D(\mathbf{y}, \boldsymbol{\eta}) = \sum_{i=1}^n (y_i - \eta_i)^2$.

Having calculated $\hat{\beta}$, we can modify the one-sample bootstrap algorithm of Section 2, and obtain an estimate of $\hat{\beta}$'s variability:

(i) Construct $\hat{F}$ putting mass $1/n$ at each observed residual,

$$\hat{F}: \text{mass } 1/n \text{ on } \hat{\epsilon}_i = y_i - g_i(\hat{\beta}).$$

(ii) Construct a bootstrap data set

$$Y_i^* = g_i(\hat{\beta}) + \epsilon_i^*, \; i = 1, 2, \ldots, n,$$

where the ϵ_i^* are drawn independently from $\hat{F}$, and calculate

$$\hat{\beta}^*: \min_{\beta} D(\mathbf{Y}^*, \boldsymbol{\eta}(\beta)).$$

(iii) Do step (ii) some large number B of times, obtaining independent bootstrap replications $\hat{\beta}^{*1}, \hat{\beta}^{*2}, \ldots, \hat{\beta}^{*B}$, and estimate the covariance matrix of $\hat{\beta}$ by

$$\hat{\boldsymbol{\Sigma}}_B = \left[\left(\sum_{b=1}^{B} (\hat{\beta}^{*b} - \hat{\beta}^{*\cdot})(\hat{\beta}^{*b} - \hat{\beta}^{*\cdot})'\right) \Big/ (B-1)\right],$$

$$(\hat{\beta}^{*\cdot} = \frac{1}{B} \Sigma \hat{\beta}^{*b}).$$

In ordinary linear regression we have $g_i(\beta) = t_i' \beta$ and $D(\mathbf{y}, \boldsymbol{\eta}) = \Sigma (y_i - \eta_i)^2$. Section 7 of Efron (1979a) shows that in this case the algorithm above can be carried out theoretically, $B = \infty$, and yields

$$\hat{\boldsymbol{\Sigma}}_B = \hat{\sigma}^2 \left(\sum_{i=1}^{n} t_i t_i'\right)^{-1}, \; \hat{\sigma}^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 / n. \tag{22}$$

This is the usual answer, except for dividing by n instead of $n - p$ in $\hat{\sigma}^2$. Of course the advantage of the bootstrap approach is that $\hat{\boldsymbol{\Sigma}}_B$ can just as well be calculated if, say, $g_i(\beta) = \exp(t_i \beta)$ and $D(\mathbf{y}, \boldsymbol{\eta}) = \sum_{i=1}^n |y_i - \eta_i|$.

There is another simpler way to bootstrap the regression problem. We can consider each covariate-response pair $x_i = (t_i, y_i)$ to be a single data point obtained by random sampling from a distribution F on $p + 1$ dimension space. Then we apply the one-sample bootstrap of Section 2 to the data set $x_1, x_2, \ldots, x_n$.

The two bootstrap methods for the regression problem are asymptotically equivalent, but can perform quite differently in small-sample situations. The simple method, described last, takes less advantage of the special structure of the regression problem. It does *not* give answer (22) in the case of ordinary least squares. On the other hand the simple method gives a trustworthy estimate of $\hat{\beta}$'s variability *even if the regression model (21) is not correct.* For this reason we use the simple method of bootstrapping on the error rate prediction problem of Sections 9 and 10.

As a final example of bootstrapping complicated data

we consider a two-sample problem with censored data. The data are the leukemia remission times listed in Table 1 of Cox (1972). The sample sizes are $m = n = 21$. Treatment-group remission times (weeks) are 6+, 6, 6, 6, 7, 9+, 10+, 10, 11+, 13, 16, 17+, 19+, 20+, 22, 23, 25+, 32+, 32+, 34+, 35+; control-group remission times (weeks) are 1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23. Here 6+ indicates a *censored* remission time, known only to exceed 6 weeks, while 6 is an uncensored remission time of exactly 6 weeks. None of the control-group times were censored.

We assume Cox's proportional hazards model, the hazard rate in the control group equaling e^{β} times that in the Treatment group. The partial likelihood estimate of β is $\hat{\beta} = 1.51$, and we want to estimate the standard error of $\hat{\beta}$. (Cox gets 1.65, not 1.51. Here we are using Breslow's convention for ties (1972), which accounts for the discrepancy.)

Figure 4 shows the histogram for 1000 bootstrap replications of $\hat{\beta}^*$. Each replication was obtained by the two-sample method described for the Hodges-Lehmann estimate:

(i) Construct $\hat{F}$ putting mass $\frac{1}{21}$ at each point 6+, 6, 6, . . . , 35+, and $\hat{G}$ putting mass $\frac{1}{21}$ at each point 1, 1, . . . , 23. (Notice that the "points" in $\hat{F}$ include the censoring information.)

(ii) Draw $X_1^*, X_2^*, \ldots, X_{21}^*$ by random sampling from $\hat{F}$, and likewise $Y_1^*, Y_2^*, \ldots, Y_{21}^*$ by random sampling from $\hat{G}$. Calculate $\hat{\beta}^*$ by applying the partial-likelihood method to the bootstrap data.

The bootstrap estimate of standard error for $\hat{\beta}$, as given by (11), is $\hat{\sigma}_B = .42$. This agrees nicely with Cox's asymptotic estimate $\hat{\sigma} = .41$. However, the percentile method gives quite different confidence intervals from those obtained by the usual method. For $\alpha = .05$, $1 - 2\alpha = .90$, the latter interval is $1.51 \pm 1.65 \cdot .41 = [.83, 2.19]$. The percentile method gives the 90 percent central interval [.98, 2.35]. Notice that $(2.35 - 1.51)/(1.51 - .98) = 1.58$, so that the percentile interval is considerably larger to the right of $\hat{\beta}$ than to the left. (The bias-corrected percentile method gives almost the same answers as the uncorrected method in this case since $\hat{C}(\hat{\beta}) = .49$.)

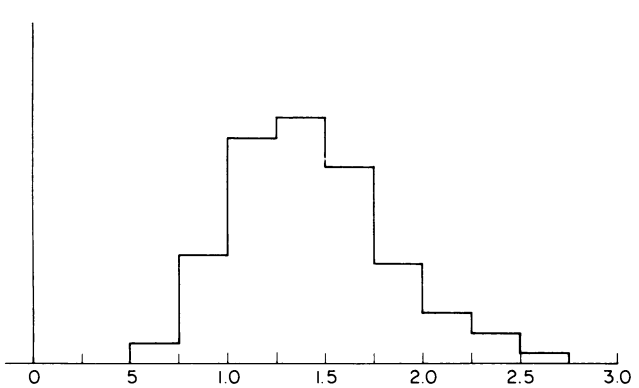


Figure 4. Histogram of 1000 bootstrap replications of $\hat{\beta}^*$ for the leukemia data, proportional hazards model. Courtesy of Rob Tibshirani, Stanford.

There are other reasonable ways to bootstrap censored data. One of these is described in Efron (1981a), which also contains a theoretical justification for the method used to construct Figure 4.

8. CROSS-VALIDATION

Cross-validation is an old but useful idea, whose time seems to have come again with the advent of modern computers. We discuss it in the context of estimating the error rate of a prediction rule. (There are other important uses; see Stone 1974; Geisser 1975.)

The prediction problem is as follows: each data point $x_i = (t_i, y_i)$ consists of a p-dimensional vector of explanatory variables t_i, and a response variable y_i. Here we assume y_i can take on only two possible values, say 0 or 1, indicating two possible responses, live or dead, male or female, success or failure, and so on. We observe $x_1, x_2, \ldots, x_n$, called collectively *the training set*, and indicated $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. We have in mind a formula $\eta(t; \mathbf{x})$ for constructing a *prediction rule* from the training set, also taking on values either 0 or 1. Given a new explanatory vector t_0, the value $\eta(t_0; \mathbf{x})$ is supposed to predict the corresponding response y_0.

We assume that each x_i is an independent realization of $X = (T, Y)$, a random vector having some distribution F on $p+1$-dimensional space, and likewise for the "new case" $X_0 = (T_0, Y_0)$. The *true error rate* err of the prediction rule $\eta(\cdot; \mathbf{x})$ is the expected probability of error over $X_0 \sim F$ with $\mathbf{x}$ fixed,

$$\text{err} \equiv E\{Q[Y_0, \eta(T_0, \mathbf{x})]\},$$

where $Q[y, \eta]$ is the error indicator

$$Q[y, \eta] = \begin{cases} 0 & \text{if } y = \eta \\ 1 & \text{if } y \neq \eta. \end{cases}$$

An obvious estimate of err is the *apparent error rate*

$$\overline{\text{err}} \equiv \hat{E}\{Q[Y_0, \eta(T_0; \mathbf{x})]\} = \frac{1}{n}\sum_{i=1}^{n} Q[y_i, \eta(t_i; \mathbf{x})].$$

The symbol $\hat{E}$ indicates expectation with respect to the empirical distribution $\hat{F}$, putting mass $1/n$ on each x_i. The apparent error rate is likely to underestimate the true error rate, since we are evaluating $\eta(\cdot, \mathbf{x})$'s performance on the same set of data used in its construction. A random variable of interest is the *overoptimism*, true minus apparent error rate,

$$R(\mathbf{x}, F) = \text{err} - \overline{\text{err}}$$
$$= E\{Q[Y_0, \eta(T_0; \mathbf{x})]\} - \hat{E}\{Q[Y_0, \eta(T_0; \mathbf{x})]\}. \quad (23)$$

The expectation of $R(\mathbf{X}, F)$ over the random choice of $X_1, X_2, \ldots, X_n$ from F,

$$\omega(F) \equiv ER(\mathbf{X}, F) \quad (24)$$

is the *expected overoptimism*.

The cross-validated estimate of err is

$$\text{err}^{\cdot} = \frac{1}{n}\sum_{i=1}^{n} Q[y_i, \eta(t_i; \mathbf{x}_{(i)})],$$

$\eta(t_i; \mathbf{x}_i)$ being the prediction rule based on $\mathbf{x}_{(i)} =$

$(x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$. In other words $\text{err}^{\dagger}$ is the error rate over the observed data set, *not allowing $x_i = (t_i, y_i)$ to enter into the construction of the rule for its own prediction.* It is intuitively obvious that $\text{err}^{\dagger}$ is a less biased estimator of err than is $\widehat{\text{err}}$. In what follows we consider how well $\text{err}^{\dagger}$ estimates err, or equivalently how well

$$\omega^{\dagger} \equiv \text{err}^{\dagger} - \overline{\text{err}}$$

estimates $R(\mathbf{x}, F) = \text{err} - \overline{\text{err}}$. (These are equivalent problems since $\text{err}^{\dagger} - \text{err} = \omega^{\dagger} - R(\mathbf{x}, F)$.) We have used the notation $\omega^{\dagger}$, rather than $R^{\dagger}$, because it turns out later that it is actually ω being estimated.

We consider a sampling experiment involving Fisher's linear discriminant function. The dimension is $p = 2$ and the sample size of the training set is $n = 14$. The distribution F is as follows: $Y = 0$ or 1 with probability $\frac{1}{2}$, and given $Y = y$ the predictor vector T is bivariate normal with identity covariance matrix and mean vector $(y - \frac{1}{2}, 0)$. If F were known to the statistician, the ideal prediction rule would be to guess $y_0 = 0$ if the first component of t_0 was ≤ 0, and to guess $y_0 = 1$ otherwise. Since F is assumed unknown, we must estimate a prediction rule from the training set.

We use the prediction rule based on Fisher's estimated linear discriminant function (Efron 1975),

$$\eta(t; \mathbf{x}) = \begin{cases} 0 \\ 1 \end{cases} \text{ if } \hat{\alpha} + t'\hat{\beta} \text{ is } \begin{cases} \leq 0 \\ > 0 \end{cases}.$$

The quantities $\hat{\alpha}$ and $\hat{\beta}$ are defined in terms of n_0 and n_1, the number of y_i equal to zero and one, respectively; $\bar{t}_0$ and $\bar{t}_1$, the averages of the t_i corresponding to those y_i equaling zero and one, respectively; and $S = [\sum_{i=1}^{n} t_i t_i' - n_0 \bar{t}_0 \bar{t}_0' - n_1 \bar{t}_1 \bar{t}_1']/n$:

$$\hat{\alpha} = [\bar{t}_1' S^{-1} \bar{t}_1 - \bar{t}_2' S^{-1} \bar{t}_2]/2,$$

$$\hat{\beta} = (\bar{t}_2 - \bar{t}_1) S^{-1}.$$

Table 5 shows the results of 10 simulations ("trials") of this situation. The expected overoptimism, obtained from 100 trials, is $\omega = .098$, so that $R = \text{err} - \overline{\text{err}}$ is typically quite large. However, R is also quite variable from trial to trial, often being negative. The cross-validation estimate $\omega^{\dagger}$ is positive in all 10 cases, and does not correlate with R. This relates to the comment that $\omega^{\dagger}$ is trying to estimate ω rather than R. We will see later that $\omega^{\dagger}$ has expectation .091, and so is nearly unbiased for ω. However, $\omega^{\dagger}$ is too variable itself to be very useful for estimating R, which is to say that $\text{err}^{\dagger}$ is not a particularly good estimate of err. These points are discussed further in Section 9, where the two other estimates of ω appearing in Table 5, $\hat{\omega}_J$ and $\hat{\omega}_B$, are introduced.

Table 5. The First 10 Trials of a Sampling Experiment Involving Fisher's Linear Discriminant Function. The Training Set Has Size n = 14. The Expected Overoptimism is ω = .096, see Table 6

		Error Rates			Estimates of Overoptimism		
Trial	n_0, n_1	True $\overline{\text{err}}$	Apparent err	Overoptimism R	Cross-validation $\omega^{\dagger}$	Jackknife $\hat{\omega}_J$	Bootstrap ($B = 200$) $\hat{\omega}_B$
1	9,5	.458	.286	.172	.214	.214	.083
2	6,8	.312	.357	−.045	.000	.066	.098
3	7,7	.313	.357	−.044	.071	.066	.110
4	8,6	.351	.429	−.078	.071	.066	.107
5	8,6	.330	.357	−.027	.143	.148	.102
6	8,6	.318	.143	.175	.214	.194	.073
7	8,6	.310	.071	.239	.071	.066	.087
8	6,8	.382	.286	.094	.071	.056	.097
9	7,7	.360	.429	−.069	.071	.087	.127
10	8,6	.335	.143	−.192	.000	.010	.048

9. BOOTSTRAP AND JACKKNIFE ESTIMATES FOR THE PREDICTION PROBLEMS

At the end of Section 6 we described a method for applying the boostrap to any random variable $R(\mathbf{X}, F)$. Now we use that method on the overoptimism random variable (23), and obtain a bootstrap estimate of the expected overoptimism $\omega(F)$.

The bootstrap estimate of $\omega = \omega(F)$, (24), is simply

$$\hat{\omega}_B = \omega(\hat{F}).$$

As usual $\hat{\omega}_B$ must be approximated by Monte Carlo. We generate independent bootstrap replications $R^{*1}, R^{*2}, \ldots, R^{*B}$, and take

$$\hat{\omega}_B \doteq \frac{1}{B} \sum_{b=1}^{B} R^{*b}.$$

As B goes to infinity this last expression approaches $E_*\{R^*\}$, the expectation of R^* under bootstrap resampling, which is by definition the same quantity as $\omega(\hat{F}) = \hat{\omega}_B$. The bootstrap estimates $\hat{\omega}_B$ seen in the last column of Table 5 are considerably less variable than the cross-validation estimates $\omega^{\dagger}$.

What does a typical bootstrap replication consist of in this situation? As in Section 3 let $\mathbf{P}^* = (P_1^*, P_2^*, \ldots, P_n^*)$ indicate the bootstrap resampling proportions $P_i^* = \#\{X_j^* = x_i\}/n$. (Notice that we are considering each vector $x_i = (t_i, y_i)$ as a single sample point for the purpose of carrying out the bootstrap algorithm.) Following through definition (13), it is not hard to see that

$$R^* = R(\mathbf{X}^*, \hat{F}) = \sum_{i=1}^{n} (P_i^o - \mathbf{P}_i^*)\, Q[y_i, \eta(t_i; \mathbf{X}^*)], \quad (25)$$

where $\mathbf{P}^o = (1, 1, \ldots, 1)'/n$ as before, and $\eta(\cdot, \mathbf{X}^*)$ is the prediction rule based on the bootstrap sample.

Table 6 shows the results of two simulation experiments (100 trials each) involving Fisher's linear discriminant fraction. The left side relates to the bivariate normal situation described in Section 8: sample size $n = 14$, dimension $d = 2$, mean vectors for the two randomly selected normal distributions $= (\pm\frac{1}{2}, 0)$. The right side still has $n = 14$, but the dimension has been raised to 5, with mean vectors $(\pm 1, 0, 0, 0, 0)$. Fuller descriptions appear in Chapter 7 of Efron (1982).

Seven estimates of overoptimism were considered. In the $d = 2$ situation, the cross-validation estimate $\omega^{\dagger}$, for example, had expectation .091, standard deviation .073, and correlation −.07 with R. This gave root mean

Table 6. Two Sampling Experiments Involving Fisher's Linear Discriminant Function. The Left Side of the Table Relates to the Situation of Table 5: $n = 14$, $d = 2$, True Mean Vectors $= (\pm\frac{1}{2}, 0)$. The Right Side Relates to $n = 14$, $d = 5$, True Mean Vectors $= (\pm 1, 0, 0, 0, 0)$

Overoptimism $R(\mathbf{X}, F)$	Dimension 2 Exp. $\omega = .096$	Sd. .113	Corr.	$\sqrt{MSE}$	Dimension 5 Exp. $\omega = .184$	Sd. .099	Corr.	$\sqrt{MSE}$
1. Ideal Constant	.096	0	0	.113	.184	0	0	.099
2. Cross-Validation	.091	.073	−.07	.139	.170	.094	−.15	.147
3. Jackknife	.093	.068	−.23	.145	.167	.089	−.26	.150
4. Bootstrap ($B = 200$)	.080	.028	−.64	.135	.103	.031	−.58	.145
5. BootRand ($B = 200$)	.087	.026	−.55	.130	.147	.020	−.31	.114
6. BootAve ($B = 200$)	.100	.036	−.18	.125	.172	.041	−.25	.118
7. Zero	0	0	0	.149	0	0	0	.209

squared error, of ω^+ for estimating R or equivalently of err^+ for estimating err,

$$[E[\omega^+ - R]^2]^{\frac{1}{2}} = [E(\text{err}^+ - \text{err})^2]^{\frac{1}{2}} = .139.$$

The bootstrap, line 4, did only slightly better, $\sqrt{\text{MSE}} = .135$.

The zero estimate $\hat{\omega} \equiv 0$, line 7, had $\sqrt{\text{MSE}} = .149$, which is also $[E(\text{err} - \overline{\text{err}})^2]^{\frac{1}{2}}$, the $\sqrt{\text{MSE}}$ of estimating err by the apparent error $\overline{\text{err}}$, with zero correction for overoptimism. The "ideal constant" is ω itself. If we knew ω, which we don't in genuine applications, we would use the bias-corrected estimate $\overline{\text{err}} + \omega$. Line 1, left side, says that this ideal correction gives $\sqrt{\text{MSE}} = .113$.

We see that neither cross-validation nor the bootstrap are much of an improvement over making no correction at all, though the situation is more favorable on the right side of Table 6. Estimators 5 and 6, which will be described later, perform noticeably better.

The "jackknife," line 3, refers to the following idea: since $\hat{\omega}_B = E_*\{R^*\}$ is a bootstrap expectation, we can approximate that expectation by (19). In this case (25) gives $R^o = 0$, so the jackknife approximation is simply $\hat{\omega}_J = (n-1)\, R_{(\cdot)}$. Evaluating this last expression, as in Chapter 7 of Efron (1982), gives

$$\hat{\omega}_J = \frac{1}{n}\sum_{i=1}^{n}\left\{ Q[y_i, \eta(t_i, \mathbf{x}_{(i)})] - \left(\sum_{j=1}^{n} Q[y_i, \eta(t_i, \mathbf{x}_{(j)})]\right)\Big/ n \right\}.$$

This looks very much like the cross-validation estimate, which can be written

$$\omega^+ = \frac{1}{n}\sum_{i=1}^{n}\{Q[y_i, \eta(t_i, \mathbf{x}_{(i)})] - Q[y_i, \eta(t_i, \mathbf{x})]\}.$$

As a matter of fact, $\hat{\omega}_J$ and ω^+ have asymptotic correlation one (Gong 1982). Their nearly perfect correlation can be seen in Table 5. In the sampling experiments of Table 6, $\text{corr}(\hat{\omega}_J, \omega^+) = .93$ on the left side, and .98 on the right side. The point here is that *the cross-validation estimate ω^+ is, essentially, a Taylor series approximation to the bootstrap estimate $\hat{\omega}_B$.*

Even though $\hat{\omega}_B$ and ω^+ are closely related in theory and are asymptotically equivalent, they behave very differently in Table 6: ω^+ is nearly unbiased and uncorrelated with R, but has enormous variability; $\hat{\omega}_B$ has small variability, but is biased downwards, particularly in the right-hand case, and highly negatively correlated with R. The poor performances of the two estimators are due to different causes, and there are some grounds of hope for a favorable hybrid.

"BootRand," line 5, modified the bootstrap estimate in just one way: instead of drawing the bootstrap sample $X_1^*, X_2^*, \ldots, X_n^*$ from $\hat{F}$, it was drawn from

$$\hat{F}_{\text{RAND}}\text{: mass } \begin{cases} \hat{\pi}_i/n & \text{on } (t_i, 1) \\ (1-\hat{\pi}_i)/n & \text{on } (t_i, 0) \end{cases} \qquad i = 1, 2, \ldots, n.$$

This is a distribution supported on $2n$ points, the observed points $x_i = (t_i, y_i)$ and also the complementary points $(t_i, 1 - y_i)$. The probabilities $\hat{\pi}_i$ were those naturally associated with the linear discriminant function,

$$\hat{\pi}_i = 1/[1 + \exp - (\hat{\alpha} + t_i'\hat{\beta})]$$

(see Efron 1975), except that $\hat{\pi}_i$ was always forced to lie in the interval $[.1, .9]$.

Drawing the bootstrap sample $X_1^*, \ldots, X_n^*$ from $\hat{F}_{\text{RAND}}$ instead of $\hat{F}$ is a form of smoothing, not unlike the smoothed bootstraps of Section 2. In both cases we support the estimate of F on points beyond those actually observed in the sample. Here the smoothing is entirely in the response variable y. In complicated problems, such as the one described in Section 10, t_i can have complex structure (censoring, missing values, cardinal and ordinal scales, discrete and continuous variates, etc.) making it difficult to smooth in the t space. Notice that in Table 6 BootRand is an improvement over the ordinary bootstrap in every way: it has smaller bias, smaller standard deviation, and smaller negative correlation with R. The decrease in $\sqrt{\text{MSE}}$ is especially impressive on the right side of the table.

"BootAve," line 6, involves a quantity we shall call $\tilde{\omega}_0$. Generating B bootstrap replications involves making nB predictions $\eta(t_i, \mathbf{X}^{*b})$, $i = 1, 2, \ldots, n$, $b = 1, 2, \ldots, B$. Let

$$I_{ib}^* = \begin{cases} 1 & \text{if } P_i^{*b} = 0 \\ 0 & \text{if } P_i^{*b} > 0 \end{cases}$$

Then

$$\tilde{\omega}_0 \equiv \Sigma_{i,b}\, I_{i,b}^*\, Q[y_i, \eta(t_i, \mathbf{X}^{*b})]/\Sigma_{ib}\, I_{ib}^* - \overline{\text{err}}.$$

In other words, $\tilde{\omega}_0 + \overline{\text{err}}$ is the observed bootstrap error rate for prediction of those y_i where x_i is not involved in the construction of $\eta(\cdot, \mathbf{X}^*)$. Theoretical arguments can be mustered to show that $\tilde{\omega}_0$ will usually have expectation greater than ω, while $\tilde{\omega}_B$ usually has expectation less than ω. "BootAve" is the compromise estimator $\hat{\omega}_{\text{AVE}} = (\hat{\omega}_B + \tilde{\omega}_0)/2$. It also performs well in Table 6, though there is not yet enough theoretical or numerical evidence to warrant unqualified enthusiasm.

The bootstrap is a general all-purpose device that can be applied to almost any problem. This is very handy,

Table 7. The Last 11 Liver Patients. Negative Numbers Indicate Missing Values

y	Constant 1	Age 2	Sex 3	Steroid 4	Antiviral 5	Fatigue 6	Malaise 7	Anorexia 8	Liver Big 9	Liver Firm 10	Spleen Palp 11	Spiders 12	Ascites 13	Varices 14	Bilirubin 15	Alk Phos 16	SGOT 17	Albumin 18	Protein 19	Histology 20	#
1	1	45	1	2	2	1	1	1	2	2	2	1	1	2	1.90	−1	114	2.4	−1	−3	145
0	1	31	1	1	2	1	2	2	2	2	2	2	2	2	1.20	75	193	4.2	54	2	146
1	1	41	1	2	2	1	2	2	2	1	1	1	2	1	4.20	65	120	3.4	−1	−3	147
1	1	70	1	1	2	1	1	1	−3	−3	−3	−3	−3	−3	1.70	109	528	2.8	35	2	148
0	1	20	1	1	2	2	2	2	2	−3	2	2	2	2	.90	89	152	4.0	−1	2	149
0	1	36	1	2	2	2	2	2	2	2	2	2	2	2	.60	120	30	4.0	−1	2	150
1	1	46	1	2	2	1	1	1	2	2	2	1	1	1	7.60	−1	242	3.3	50	−3	151
0	1	44	1	2	2	1	2	2	2	1	2	2	2	2	.90	126	142	4.3	−1	2	152
0	1	61	1	1	2	1	1	2	1	1	2	1	2	2	.80	95	20	4.1	−1	2	153
0	1	53	2	1	2	1	2	2	2	2	1	1	2	1	1.50	84	19	4.1	48	−3	154
1	1	43	1	2	2	1	2	2	2	2	1	1	1	2	1.20	100	19	3.1	42	2	155

but it implies that in situations with special structure the bootstrap may be outperformed by more specialized methods. Here we have done so in two different ways. BootRand uses an estimate of F that is better than the totally nonparametric estimate F. BootAve makes use of the particular form of R for the overoptimism problem.

10. A COMPLICATED PREDICTION PROBLEM

We end this article with the bootstrap analysis of a genuine prediction problem, involving many of the complexities and difficulties typical of genuine problems. The bootstrap is not necessarily the best method here, as discussed in Section 9, but it is impressive to see how much information this simple idea, combined with massive computation, can extract from a situation that is hopelessly beyond traditional theoretical solutions. A fuller discussion appears in Efron and Gong (1981).

Among $n = 155$ acute chronic hepatitis patients, 33 were observed to die from the disease, while 122 survived. Each patient had associated a vector of 20 covariates. On the basis of this training set it was desired to produce a rule for predicting, from the covariates, whether a given patient would live or die. If an effective prediction rule were available, it would be useful in choosing among alternative treatments. For example, patients with a very low predicted probability of death could be given less rigorous treatment.

Let $x_i = (t_i, y_i)$ represent the data for patient i, $i = 1, 2, \ldots, 155$. Here t_i is the 20-dimensional vector of covariates, and y_i equals 1 or 0 as the patient died or lived. Table 7 shows the data for the last 11 patients. Negative numbers represent missing values. Variable 1 is the constant 1, included for convenience. The meaning of the 19 other predictors, and their coding in Table 7, will not be explained here.

A prediction rule was constructed in 3 steps:

1. An $\alpha = .05$ test of the importance of predictor j, $H_0: \beta_j = 0$ versus $H_1: \beta_j \neq 0$, was run separately for $j = 2, 3, \ldots, 20$, based on the logistic model

$$\log \frac{\pi(t_i)}{1 - \pi(t_i)} = \beta_i + \beta_j t_{ij},$$

$$\pi(t_i) \equiv \text{Prob}\{\text{patient } i \text{ dies}\}.$$

Among these 19 tests, 13 predictors indicated predictive power by rejecting $H_0: j = 18, 13, 15, 12, 14, 7, 6, 19, 20, 11, 2, 5, 3$. These are listed in order of achieved significance level, $j = 18$ attaining the smallest alpha.

2. These 13 predictors were tested in a forward multiple-logistic-regression program, which added predictors one at a time (beginning with the constant) until no further single addition achieved significance level $\alpha = .10$. Five predictors besides the constant survived this step, $j = 13, 20, 15, 7, 2$.

3. A final forward, stepwise multiple-logistic-regression program on these five predictors, stopping this time at level $\alpha = .05$, retained four predictors besides the constant, $j = 13, 15, 7, 20$.

At each of the three steps, only those patients having no relevant data missing were included in the hypothesis tests. At step 2 for example, a patient was included only if all 13 variables were available.

The final prediction rule was based on the estimated logistic regression

$$\log \frac{\pi(t_i)}{1 - \pi(t_i)} = \sum_{J=1,\ 13,\ 15,\ 7,\ 20} \hat{\beta}_j\, t_{ij},$$

where $\hat{\beta}_j$ was the maximum likelihood estimate in this model. The prediction rule was

$$\eta(t; \mathbf{x}) = \begin{cases} 1 \\ 0 \end{cases} \text{ if } \Sigma_j\, \hat{\beta}_j\, t_{ij} \begin{cases} \geq c \\ < c \end{cases}, \tag{26}$$

$c = \log 33/122$.

Among the 155 patients, 133 had none of the predictors 13, 15, 7, 20 missing. When the rule $\eta(t; \mathbf{x})$ was applied to these 133 patients, it misclassified 21 of them, for an apparent error rate $\overline{\text{err}} = 21/133 = .158$. We would like to estimate how overoptimistic $\overline{\text{err}}$ is.

To answer this question, the simple bootstrap was applied as described in Section 9. A typical bootstrap sample consisted of $X_1^*, X_2^*, \ldots, X_{155}^*$, randomly drawn with replacement from the training set $x_1, x_2, \ldots, x_{155}$. The bootstrap sample was used to construct the bootstrap prediction rule $\eta(\cdot, \mathbf{X}^*)$, following the same three steps used in the construction of $\eta(\cdot, \mathbf{x})$, (26). This gives a bootstrap replication R^* for the overoptimism random variable $R = \text{err} - \overline{\text{err}}$, essentially as in (25), but with a modification to allow for difficulties caused by missing predictor values.

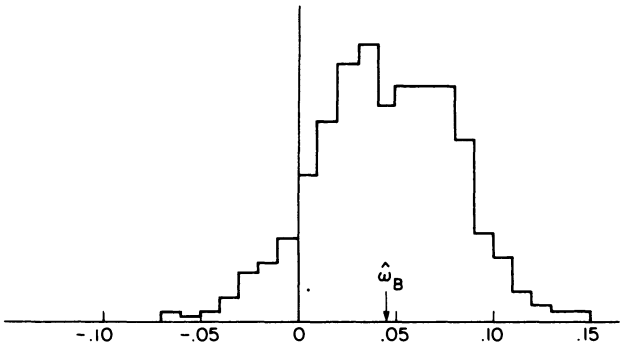


Figure 5. Histogram of 500 bootstrap replications of overoptimism for the hepatitis problem.

Figure 5 shows the histogram of $B = 500$ such replications. 95 percent of these fall in the range $0 \le R^* \le .12$. This indicates that the unobservable true overoptimism err $-$ $\overline{\text{err}}$ is likely to be positive. The average value is

$$\hat{\omega}_B = \tfrac{1}{B} \sum_{b=1}^{B} R^{*b} = .045,$$

suggesting that the expected overoptimism is about $\frac{1}{3}$ as large as the apparent error rate .158. Taken literally, this gives the bias-corrected estimated error rate .158 + .045 = .203. There is obviously plenty of room for error in this last estimate, given the spread of values in Figure 5, but at least we now have some idea of the possible bias in err.

The bootstrap analysis provided more than just an estimate of $\omega(F)$. For example, the standard deviation of the histogram in Figure 5 is .036. This is a dependable estimate of the true standard deviation of R

13	7	20	15			
13	19	6				
20	16	19				
20	19					
14	18	7	16	2		
18	20	7	11			
20	19	15				
20						
13	12	15	8	18	7	19
15	13	19				
13	4					
12	15	3				
15	16	3				
15	20	4				
16	13	2	19			
18	20	3				
13	15	20				
15	13					
15	20	7				
13						
15						
13	14					
12	20	18				
2	20	15	7	19	12	
13	20	15	19			

Figure 6. Predictors selected in the last 25 bootstrap replications for the hepatitis program. The predictors selected by the actual data were 13, 15, 7, 20.

(see Efron 1982, Ch. VII), which by definition equals $[E(\text{err} - \overline{\text{err}} - \omega)^2]^{1/2}$, the $\sqrt{\text{MSE}}$ of $\overline{\text{err}} + \omega$ as an estimate of err. Comparing line 1 with line 4 in Table 6, we expect $\overline{\text{err}} + \hat{\omega}_B = .203$ to have $\sqrt{\text{MSE}}$ at least this big for estimating err.

Figure 6 illustrates another use of the bootstrap replications. The predictions chosen by the three-step selection procedure, applied to the bootstrap training set $\mathbf{X}^*$, are shown for the last 25 of the 500 replications. Among all 500 replications, predictor 13 was selected 37 percent of the time, predictor 15 selected 48 percent, predictor 7 selected 35 percent, and predictor 20 selected 59 percent. No other predictor was selected more than 50 percent of the time. No theory exists for interpreting Figure 6, but the results certainly discourage confidence in the casual nature of the predictors 13, 15, 7, 20.

[*Received January 1982. Revised May 1982.*]

REFERENCES

BRESLOW, N. (1972), Discussion of Cox (1974), *Journal of the Royal Statistical Society*, Ser. B, 34, 216–217.

COX, D.R. (1972), "Regression Models With Life Tables," *Journal of the Royal Statistical Society*, Ser. B, 34, 187–000.

CRAMÉR, H. (1946), *Mathematical Methods of Statistics*, Princeton: Princeton University Press.

EFRON, B. (1975), "The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis," *Journal of the American Statistical Association*, 70, 897–898.

——— (1979a), "Bootstrap Methods: Another Look at the Jackknife," *Annals of Statistics*, 7, 1–26.

——— (1979b), "Computers and the Theory of Statistics: Thinking the Unthinkable," *SIAM Review*, 21, 460–480.

——— (1981a), "Censored Data and the Bootstrap," *Journal of the American Statistical Association*, 76, 312–319.

——— (1981b), "Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap, and Other Resampling Methods," *Biometrika*, 00, 000–000.

——— (1981c), "Nonparametric Standard Errors and Confidence Intervals," *Canadian Journal of Statistics*, 9, 139–172.

——— (1982), "The Jackknife, the Bootstrap, and Other Resampling Plans," *SIAM*, monograph #38, CBMS–NSF.

EFRON, B., and GONG, G. (1981), "Statistical Theory and the Computer," unpublished manuscript.

GEISSER, S. (1975), "The Predictive Sample Reuse Method With Applications," *Journal of the American Statistical Association*, 70, 320–328.

GONG, G. (1982), "Cross-validation, the Jackknife, and the Bootstrap: Excess Error Estimation in Forward Logistic Regression", Ph.D. dissertation, Dept. of Statistics, Stanford University.

HAMPEL, F. (1974), "The Influence Curve and its Role in Robust Estimation," *Journal of the American Statistical Association*, 69, 383–393.

JAECKEL, L. (1972), "The Infinitesimal Jackknife," Bell Laboratories Memorandum #MM 72-1215-11.

JOHNSON, N., and KOTZ, S. (1970), *Continuous Univariate Distributions* (vol. 2), Boston: Houghton Mifflin.

MALLOWS, C.L. (1974), "On Some Topics in Robustness", Memorandum, Bell Laboratories, Murray Hill, New Jersey.

QUENOUILLE, M. (1949), "Approximate Tests of Correlation in Time Series," *Journal of The Royal Statistical Society*, Ser. B, 11, 18–84.

SHUCANY, W.; BRAY, H.; and OWEN, O. (1971), "On Bias Reduction in Estimation," *Journal of the American Statistical Association*, 66, 524–533.

STONE, M. (1974), "Cross-Validatory Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society*, Ser. B, 36, 111–147.

1 Introduction

Statistical resampling methods have become feasible for parametric estimation, hypothesis testing, and model validation now that the computer is a ubiquitous tool for statisticians. This essay focuses on the resampling technique for parametric estimation known as the Jackknife procedure. To outline the usefulness of the method and its place in the general class of statistical resampling techniques, I will quickly delineate two similar resampling methods: the bootstrap and the permutation test.

1.1 Other Sampling Methods: The Bootstrap

The bootstrap is a broad class of usually non-parametric resampling methods for estimating the sampling distribution of an estimator. The method was described in 1979 by Bradley Efron, and was inspired by the previous success of the Jackknife procedure.[1]

Imagine that a sample of n independent, identically distributed observations from an unknown distribution have been gathered, and a mean of the sample, $\bar{Y}$, has been calculated. To make inferences about the population mean we need to know the variability of the sample mean, which we know from basic statistical theory is $\mathrm{V}[\bar{Y}] = \mathrm{V}[Y]/n$. Here, since the distribution is unknown, we do not know the value of $\mathrm{V}[Y] = \sigma^2$. The central limit theorem (CLT) states that the standardized sample mean converges in distribution to a standard normal Z as the sample size grows large—and we can invoke Slutsky's theorem to demonstrate that the sample standard deviation is an adequate estimator for standard deviation σ when the distribution is unknown. However, for other statistics of interest that do not admit the CLT, and for small sample sizes, the bootstrap is a viable alternative.

Briefly, the bootstrap method specifies that B samples be generated from the data by sampling with replacement from the original sample, with each sample set being of identical size as the original sample (here, n). The larger B is, the closer the set of samples will be to the ideal exact bootstrap sample, which is of the order of an n-dimensional simplex: $|C_n| = (2n-1)\mathbf{C}(n)$. The computation of this number, never mind the actual sample, is generally unfeasible for all but the smallest sample sizes (for example a sample size of 12 has about 1.3 million with-replacement subsamples). Furthermore, the bootstrap follows a multinomial distribution, and the most likely sample is in fact the original sample, hence it is almost certain that there will be random bootstrap samples that are replicates of the original sample. This means that the computation of the exact bootstrap is all but impossible in practice. However, Efron and Tibshirani have argued that in some instances, as few as 25 bootstrap samples can be large enough to form a reliable estimate.[2]

The next step in the process is to perform the action that derived the initial statistic—here the mean: so we sum each bootstrap sample and divide the total by n, and use those quantities to generate an estimate of the variance of $\bar{Y}$ as follows:

$$SE(\bar{Y})_B = \left\{ \frac{1}{B} \sum_{b=1}^{B} (\bar{Y}_b - \bar{Y})^2 \right\}^{1/2}$$

The empirical distribution function (EDF) used to generate the bootstrap samples can be shown to be a consistent, unbiased estimator for the actual cumulative distribution function (CDF) from which the samples were drawn, F. In fact, the bootstrap performs well because it has a faster rate of convergence than the CLT: $O(1/n)$ vs. $O(1/\sqrt{n})$, as the bootstrap relies on the strong law of large numbers (SLLN), a more robust condition than the CLT.

1.2 Other Sampling Methods: Permutation

Permutation testing is done in many arenas, and a classical example is that of permuted y's in a pair of random vectors $(\mathbf{X}, \mathbf{Y})$ to get a correlation coefficient p-value. For an observed sample $\mathbf{z} = \{(X_1, \ldots, X_n), (Y_1, \ldots, Y_n)\}$, the elements of (only) the $\mathbf{Y}$ vector are permuted B times. Then for permutation function $\pi(\cdot)$, we have that an individual permutation sample $\mathbf{z}_b$ is:

$$\mathbf{z}_b = \{(X_1, \ldots, X_n), (Y_{\pi(1)}, \ldots, Y_{\pi(n)})\}$$

The next step is to compute the number of times that the original correlation statistic is in absolute value greater than the chosen percentile threshold (say, 0.025 and 0.975 for an empirical α level of 0.05), divided by B. This value is the empirical p-value. If $\mathrm{B} = n!$ then the test is called *exact*; if all of the permutations are not performed, then there is an inflated Type I error rate, as we are less likely to sample those values in the tails of the null distribution, and hence we are less likely to say that there are values greater in absolute value than our original statistic. This method is entirely non-parametric, and is usually approximated by Monte Carlo methods for large sample sizes where the exact permutation generation is computationally impractical.

2 The Jackknife: Introduction and Basic Properties

The Jackknife was proposed by M.H. Quenouille in 1949 and later refined and given its current name by John Tukey in 1956. Quenouille originally developed the method as a procedure for correcting bias. Later, Tukey described its use in constructing confidence limits for a large class of estimators. It is similar to the bootstrap in that it involves resampling, but instead of sampling with replacement, the method samples *without* replacement.

Many situations arise where it is impractical or even impossible to calculate good estimators or find those estimators' standard errors. The situation may be one where there is no theoretical basis to fall back on, or it may be that in estimating the variance of a difficult function of a statistic, say $g(\bar{X})$ for some function with no closed-form integral, making use of the usual route of estimation—the delta method theorem—is impossible. In these situations the Jackknife method can be used to derive an estimate of bias and standard error. Keith Knight has noted, in his book *Mathematical Statistics*, that the Jackknife estimate of the standard error is roughly equivalent

to the delta method for large samples.[3]

Definition: The delete-1 **Jackknife Samples** are selected by taking the original data vector and deleting one observation from the set. Thus, there are n unique Jackknife samples, and the ith Jackknife sample vector is defined as:

$$\mathbf{X}_{[i]} = \{X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_{n-1}, X_n\}$$

This procedure is generalizable to k deletions, which is discussed further below.

The ith **Jackknife Replicate** is defined as the value of the estimator $s(\cdot)$ evaluated at the ith Jackknife sample.

$$\hat{\theta}_{(i)} := s(\mathbf{X}_{[i]})$$

The Jackknife Standard Error is defined

$$SE(\hat{\theta})_{jack} = \Big\{ \frac{n-1}{n} \sum_{i=1}^{n} (\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)})^2 \Big\}^{1/2},$$

where $\hat{\theta}_{(\cdot)}$ is the empirical average of the Jackknife replicates:

$$\hat{\theta}_{(\cdot)} = \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}_{(i)}$$

The $(n-1)/n$ factor in the formula above looks similar to the formula for the standard error of the sample mean, except that there is a quantity $(n-1)$ included in the numerator. As motivation for this estimator, I consider the case that does not actually need any resampling methods: that of the sample mean. Here, the Jackknife estimator above is an unbiased estimator of the variance of the sample mean.

To demonstrate this claim, I need to show that

$$\frac{n-1}{n} \sum_{i=1}^{n} (\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)})^2 = \frac{1}{n(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

I note that here the Jackknife replicates in the inner squared term on the left simplify as follows:

$$(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}) = \frac{n\bar{x} - x_i}{n-1} - \frac{1}{n} \sum_{i=1}^{n} \bar{x}_{(i)} = \frac{1}{n-1} \Big(n\bar{x} - x_i - \frac{1}{n} \sum_{i=1}^{n} n\bar{x} - x_i \Big) = \frac{1}{n-1} (\bar{x} - x_i)$$

Once the term is squared, the equation is complete, and is identically equal to the right hand term above. Thus, in the case of the sample mean, the Jackknife estimate of the standard error reduces

to the regular, unbiased estimator commonly used. The standard error estimate is somewhat *ad hoc*, but it is also intuitive. A more formal derivation was provided by Tukey and involves *pseudovalues*, which are discussed briefly below. It has been shown, however, that the Jackknife estimate of variance is slightly biased upward,[4] and does not work in all situations, for example, as an estimator of the median (see Knight, ibid.). In instances such as quantile estimation, it has been shown that the **delete** $-$ **d** Jackknife, where $\sqrt{n} < d < (n-1)$, is a consistent estimator.[5] The delete-d variance estimator has similar form as the delete-1 estimator, with a different normalizing constant:

$$SE(\hat{\theta})_{d-jack} = \left\{ \frac{n-d}{d\binom{n}{d}} \sum_{z} (\hat{\theta}_{(z)} - \hat{\theta}_{(\cdot)})^2 \right\}^{1/2}$$

The **Jackknife Bias** is defined as

$$\widehat{\text{bias}}_{jack} = (n-1)(\hat{\theta}_{(\cdot)} - \hat{\theta}),$$

where $\hat{\theta}$ is the estimator taking the entire sample as argument. Jackknife Bias is just the average of the deviations of the replicates, which are sometimes called $Jackknife\ Influence\ Values$, multiplied by a factor $(n-1)$. The bias of the sample mean is 0, so I cannot take the function $s(\cdot) = \bar{x}$ to get an idea of what the multiplier should be, as I did previously for the Jackknife SE. Instead, I consider as an estimator the uncorrected variance of the sample:

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

If I take as my estimator $\hat{\theta}$ the (biased) sample variance, I have that its bias, that is, $\mathrm{E}[\hat{\theta} - \theta]$, is equal to $-\sigma^2/n$. If I use the Jackknife bias as an estimate for the bias of my estimator, and I have that my estimator $\hat{\theta}$ is equal to the uncorrected sample variance, then the Jackknife bias formula reduces to $-S^2/n$, where S^2 is now the regular, corrected, unbiased estimator of sample variance. Thus, the bias here is constructed from a heuristic notion to emulate the bias of the uncorrected sample variance.

The above synopsis gave a rationale based on familiar sample-based estimators. Here is another justification: Assume that for any fixed n the expected value of an estimator is the parameter estimand plus some bias term, call it $b_1(\theta)/n$. Then, as the average of the Jackknife replicates has $(n-1)$ terms, the expected value of the average is

$$E[\hat{\theta}_{(\cdot)}] = \frac{1}{n} \sum_{i=1}^{n} E[\hat{\theta}_{(i)}] = \theta + \frac{b_1(\theta)}{n-1}$$

From this observation it follows that the bias of the Jackknife replicates estimator is

$$E[\hat{\theta} - \hat{\theta}_{(\cdot)}] = \theta + \frac{b_1(\theta)}{n} - \theta - \frac{b_1(\theta)}{n-1} = \frac{b_1(\theta)}{n(n-1)}$$

Hence if we multiply this difference above by $(n-1)$, we get an unbiased estimator of the bias of our original estimator.

With the estimate of bias in hand, an obvious extension is to define a Jackknife estimate of the parameter of interest as

$$\hat{\theta}_{jack} = \hat{\theta} - \widehat{\text{bias}}_{jack} = \hat{\theta} - (n-1)(\hat{\theta}_{(\cdot)} - \hat{\theta}) = n\hat{\theta} - (n-1)\hat{\theta}_{(\cdot)}$$

The estimator is made clear if we remember than the Jackknife bias of the original estimator is $(n-1)(\hat{\theta}_{(\cdot)} - \hat{\theta})$, and hence the bias of the new estimator $\hat{\theta}_{jack}$ is 0. In practice, it is not always exactly 0, as the above treatment is really just a first-order Taylor series approximation, but the bias of biased estimators is often reduced by this method. If we imagine that the situation described in the explanation of the Jackknife bias where the bias is a linear combination of the estimand and a bias term were expanded so that the bias term is now an infinite series of terms, $b_1(\theta)/n + b_2(\theta)/n^2 + b_3(\theta)/n^3 + \ldots$, and the expected value of the original estimator was that summation plus the estimand, then we have that

$$E[\hat{\theta}_{(\cdot)}] = \frac{1}{n}\sum_{i=1}^{n} E[\hat{\theta}_{(i)}] = \theta + \frac{b_1(\theta)}{n-1} + \frac{b_2(\theta)}{(n-1)^2} + \frac{b_3(\theta)}{(n-1)^3} + \ldots$$

and

$$E[\widehat{\text{bias}}_{jack}] = (n-1)E[\hat{\theta} - \hat{\theta}_{(\cdot)}] = \frac{b_1(\theta)}{n} + \frac{(2n-1)b_2(\theta)}{n^2(n-1)} + \frac{(3n^2-3n+1)b_3(\theta)}{n^3(n-1)} + \ldots$$

and finally that the expected value of the Jackknife estimator is

$$E[\hat{\theta}_{jack}] = E[n\hat{\theta} - (n-1)\hat{\theta}_{(\cdot)}] = \theta - \frac{b_2(\theta)}{n(n-1)} - \frac{(2n-1)b_3(\theta)}{n^2(n-1)^2} + \ldots \approx \theta - \frac{b_2(\theta)}{n^2} - \frac{2b_3(\theta)}{n^3} - \ldots$$

The above formulation shows that, as there is no first-order term n in the denominators of the infinite sum terms (i.e. the first term in the Taylor expansion has cancelled out), the Jackknife bias is asymptotically smaller than the bias of any given biased estimator.

Another construction sometimes used in Jackknife estimation is the "pseudovalue," and can be seen as a bias-corrected version of the estimator. The scheme is to treat the jackknife pseudovalues as if they were independent random variables.

Definition: The ith Pseudovalue of estimator $\varphi_n(\mathbf{X})$ for sample vector $\mathbf{X}$ is defined as

$$ps_i = n\varphi_n(\mathbf{X}) - (n-1)\varphi_{n-1}(\mathbf{X}_{[i]})$$

The pseudovalues can also be written as

$$ps_i = \varphi_n(\mathbf{X}) + (n-1)(\varphi_n(\mathbf{X}) - \varphi_{n-1}(\mathbf{X}_{[i]})$$

These constructs can be used in place of the replicate terms in the Jackknife SE to give confidence intervals from the t distribution. However, the method is criticized by Efron and Tibshirani, who write "This interval does not work very well: in particular, it is not significantly better than cruder intervals based on normal theory (ibid., p. 145)."

Nevertheless, it can easily be shown that pseudovalues can be used to construct a normal test of hypotheses. Since each pseudovalue is independent and identically distributed (iid), it follows that their average conforms to a normal distribution as the sample size grows large. The average of the pseudovalues is just $\hat{\theta}_{jack}$, and the expected value of that average, owing by construction to the unbiasedness of the estimator, is the parameter under investigation, θ. Thus, we have that

$$\frac{\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}(n\varphi_n(\mathbf{X}) - (n-1)\varphi_{n-1}(\mathbf{X}_{[i]})) - \theta\right)}{\hat{S}} \quad \rightarrow \quad N(0,1),$$

where $\hat{S}$ is the square root of the sum of the squared differences of the pseudovalue compared against $\hat{\theta}_{jack}$, divided by the sample size minus 1 (the unbiased estimate of variance).

The Jackknife can be used in many situations. However, the method is inappropriate for correlated data or time series data. The method assumes independence between the random variables (and identically distributed data points), and if that assumption is violated, the results will be of no use. Another condition of note is that the Jackknife estimate is composed of a linear function (subtraction) and hence will only work properly for linear functions of the data and/or parameters, or on functions that are smooth enough to be modeled as continuous without much of a problem.

3 Examples

Imagine that we know that we are sampling from a uniform distribution on interval $[0, \theta]$, $\theta > 0$, and we are interested in estimating the upper bound. A simple, intuitive estimator is the sample maximum, but is this biased? The expectation of the maximum, $Y = \max(\{X_n\})$, is

$$\int_{-\infty}^{\infty} y\ f(y)dy = \int_{0}^{\theta} yn\left(\frac{y}{\theta}\right)^n dy = \frac{n}{n+1}\theta$$

This estimate is clearly biased. Since the maximum of a given fixed sample drawn from a continuous CDF is the same element for $(n-1)$ out of the n Jackknife samples, and is the second-largest term in the single Jackknife sample subset that excludes the largest element of the original sample, it is clear that the average of the Jackknife replicates is

$$\hat{\theta}_{(\cdot)} = \frac{n-1}{n}X_{(n)} + \frac{1}{n}X_{(n-1)}$$

and so the Jackknife estimate of the maximum is

$$\hat{\theta}_{jack} = X_{(n)} + \frac{n-1}{n}(X_{(n)} - X_{(n-1)})$$

By the results above, we have that the bias of this estimator will be smaller than that of the sample maximum, but if we had wanted to we could have just corrected the sample maximum by the constant $(n+1)/n$ to yield a totally unbiased estimator. However, the Jackknife estimator is generalizable to *any* distribution that has an upper bound that we would like to estimate, regardless of whether the random variable is distributed uniformly or not. This property is extremely useful when the distribution is unknown, or when it is unclear how to correct for the bias of the sample maximum.

To show empirically that the Jackknife bias is smaller than the maximum for $X \sim U(0, \theta)$, I wrote an `R` script that sampled from a uniform distribution on the interval $[0, 5]$, and recorded the average number of times that the absolute bias between the Jackknife estimate and 5 was less than the absolute bias between the sample maximum and 5. I ran 100,000 simulations for sample sizes 10, 30, and 100. The `R` code for sample size 100 and results follow:

```
numvec<-rep(NA,100000)
maxbiasvec<-rep(NA,100000)
jackbiasvec<-rep(NA,100000)
for (i in 1:100000){
samp<-runif(100, min = 0, max = 5)
jack<-max(samp)+(100-1)/100*(max(samp)-max(samp[!samp==max(samp)]))
numvec[i]<-ifelse(abs(5-jack)<abs(max(samp)-5), 1,0)
maxbiasvec[i]<-abs(5-max(samp))
jackbiasvec[i]<-abs(5-jack)      }
mean(numvec)
mean(jackbiasvec)
mean(maxbiasvec)
```

100K Samples	N=10	N=30	N=100
% JN bias < Sample Max bias	68.9%	67.73%	66.94%
Average Bias Jack: abs(JK-5)	0.4324	0.1582	0.0492
Average Bias Max: abs(max-5)	0.4549	0.1608	0.0496

For each sample size N the Jackknife bias was smaller than the bias of the sample maximum about 68% of the time. Of course, simply multiplying the maximum by a constant would be ideal in this case, but again, the method is generalizable to more complicated situations.

4 Applications

One interesting application I came upon in a course on Microarray analysis is the genomic software application EASE (Expression Analysis Systematic Explorer).[6] In experiments and analyses that use gene expression data, researchers often finalize their study with an annotated list of genes found to be differentially expressed between biological conditions. For example gene X may be expressed more in cancerous cells than in normal cells. Researchers often annotate a list of candidate genes one-by-one by looking up relevant information on the genes from an online database, or automating the process for a large gene set. The results of such a search are often difficult to interpret, especially for those not versed in the finer elements of biochemistry and cell biology. The gene set annotation will not inform nonspecialists as to whether the group of genes discovered in the analysis stage of the study is related to a plausible biological function category that has already been investigated previously: for example, genes known to regulate hemoglobin characteristics in a study to discover genes associated with sickle cell anemia.

The EASE software queries publicly available databases with the option of adding user-defined ontology categories. The program initially gives a Fisher exact (hypergeometric) test for each known class. As an example, researchers discover gene set **A** is found to be associated with some phenotype of interest. They compare gene set **A** against annotation ontology set **X**, which lists all known genes related to some biological function, say, apoptosis (programmed cell death) to see if gene set **A** is overrepresented in this biological theme more than would be expected by chance. If so, this knowledge can be used to confirm a hypothesis, or suggest a new avenue of research. Each gene in set **A** either is or is not in gene set **X**, hence, we can give an exact p-value to the probability of observing at least as many genes in discovery set **A** as in the known ontology gene set **X**.

The software also computes an "EASE" score, which is a Jackknifed score for the exact test. The goal of this test is to construct a conservative score that is similar to the score given to most ontology sets, but which penalizes those known gene sets that have very few members. The program takes the discovery set **A**, and in calculating the hypergeometric test for each ontology gene set it removes a single gene from the discovery set and computes Fisher p-values anew for every subset of genes in the discovery set having one gene excluded.

As an example, if there is only a *single* gene in some ontology category **X**, and that gene happens to appear in discovery set **A** which contains 206 genes, then the p-value for that gene set **X** is 0.0152. However a different ontology gene set **Y** would be only slightly less significant if it had 787 genes, with 20 of our discovery set being members of that class. This is obviously a problem in commensurability. The authors of the paper write:

> "From the perspective of global biological themes, a theme based on the presence of a single gene is neither global nor stable and is rarely interesting. If the single [discovery] gene happens to be a false positive, then the significance of the dependent [ontology] theme is entirely false. However, the EASE score for these two situations is $p = 1$ for category X and $p < 0.0274$ for category Y, and thus the EASE score eliminates the significance of the 'unstable' category X while only slightly penalizing the significance of the more global theme Y."

5 Differences Between Jackknife and other Resampling Methods

That the Jackknife is asymptotically equivalent to the bootstrap is inventively shown by Efron and Tibshirani in their monograph (ibid.). They discuss the process of sampling from blocks of data as equivalent to defining a multinomial distribution on the sample n-tuple with equal probabilities for each sample point. In the case of $n = 3$, the list of possible points is described as a triangle where each point can be selected from the set of all possible resamples of the three elements, and is graphically represented as points along the three edges of the triangle. The graphic below is reproduced from the Efron and Tibshirani paper presented in 1986.[7]

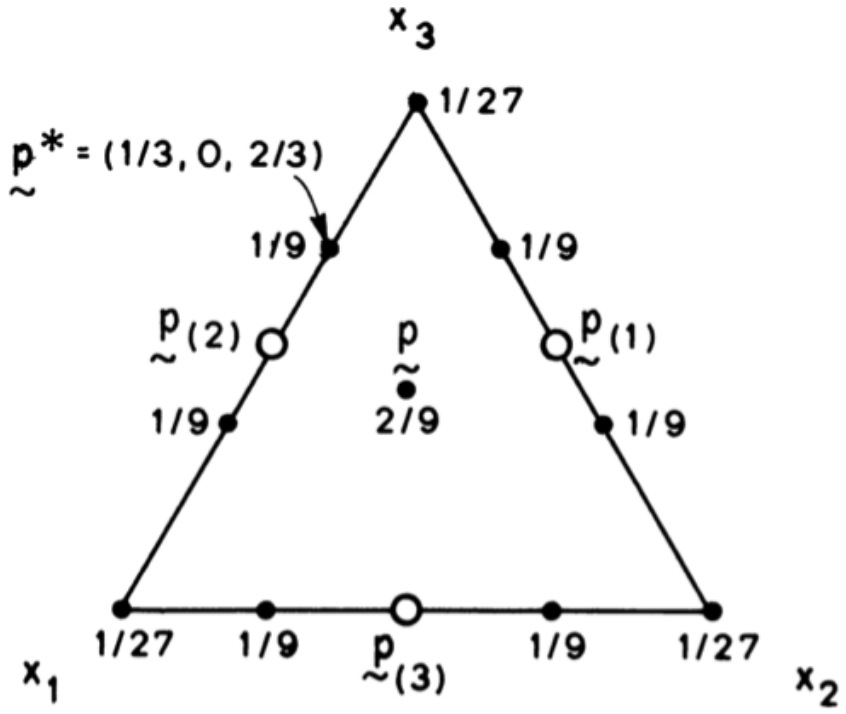


FIG. 15. *The bootstrap and jackknife sampling points in the case $n = 3$. The bootstrap points (·) are shown with their probabilities.*

It shows the domain of the sample functional, where the ideal bootstrap sample represents the surface attained by the domain points on the simplex. Then the jackknife sample is an approximating hyperplane to the bootstrap surface.

I now present the theorem given by Efron and Tibshirani (cf. 2, p. 287-288).

Define $\mathbf{P}^0 = (1/n, \ldots, 1/n)^T$, and $\mathbf{U} = (U_1, U_2, \ldots, U_n)^T$ such that the elements of $\mathbf{U}$ sum to 0. Then $\mathrm{T}(\mathbf{P}^0)$ is the original sample statistic, and $\mathrm{T}(\mathbf{P}_{(i)})$ is the jackknife replicate for sample point i: $\mathbf{P}_{(i)} = (1/(n-1), \ldots, 0, \ldots 1/(n-1))^T$.

A linear statistic T($\mathbf{P}^*$), where T is the functional of the vector of all probabilities such that each element is in [0, 1] and the sum of the elements is 1, has the following form:

$$T(\mathbf{P}^*) = c_0 + (\mathbf{P}^* - \mathbf{P}^0)^T \mathbf{U}$$

The linear statistic defines a hyperplane over simplex S_n. The following result states that for any statistic the jackknife estimate of the variance of T($\mathbf{P}^*$) is almost the same as the bootstrap estimate for a certain linear approximation to T($\mathbf{P}^*$).

Theorem: Let T^{LIN} be the unique hyperplane passing through the Jackknife points $(\mathbf{P}_{(i)}, T(\mathbf{P}_{(i)}))$, $i = 1, 2, \ldots, n$. Then $var * T^{LIN} = (n-1)/n var_{jack}$, where $var*$ is the variance under the multinomial distribution of all probability vectors. In the case $n = 3$, it would just be the addition of the possible samples weighted by their probabilities; var_{jack} is the usual formula given above, and reprinted here for clarity:

$$\frac{n-1}{n} \sum_{i=1}^{n} (\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)})^2$$

Thus, the Jackknife estimate of the variance for our estimator of interest $\hat{\theta}$ is $n/(n-1)$ times the bootstrap estimate of variance for the linear approximation to the surface described by the bootstrap simplex.

Proof: By solving n linear equations of the form $\hat{\theta}_{(i)} = T^{LIN}(\mathbf{P}_{(i)})$ for c_0 and the components of the $\mathbf{U}$ vector, we find that $c_0 = \hat{\theta}_{(i)}$ and $U_i = (n-1)(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)})$. Using the fact that $\mathbf{P}^*$ is distributed as a multiple of the multinomial distribution with mean and covariance matrix $(\mathbf{P}^0, [I/n^2 - \mathbf{P}^0\mathbf{P}^0 T/n])$, and that the U_i terms sum to 0 we have that $var * T^{LIN}(\mathbf{P}^*) = \mathbf{U}^T (var * \mathbf{P}^*)\mathbf{U} = 1/n^2 \mathbf{U}^T\mathbf{U} = (n-1)/n\{(n-1)/n \sum_{i=1}^{n} (\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)})^2\}$. QED.

Thus, the accuracy of the Jackknife as a linear approximation to the bootstrap depends on how well the hyperplane T^{LIN} approximates $T(\mathbf{P}^*)$.

References

[1] B. Efron, "Bootstrap methods: another look at the jackknife," *The annals of Statistics*, pp. 1–26, 1979.

[2] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap*. CRC press, 1994.

[3] K. Knight, *Mathematical Statistics*. New York Chapman and Hall/CRC, 2000, Pg. 218.

[4] B. Efron and C. Stein, "The jackknife estimate of variance," *The Annals of Statistics*, pp. 586–596, 1981.

[5] J. Shao and C. J. Wu, "A general theory for jackknife variance estimation," *The Annals of Statistics*, pp. 1176–1197, 1989.

[6] D. A. Hosack, G. Dennis Jr, B. T. Sherman, H. C. Lane, R. A. Lempicki, *et al.*, "Identifying biological themes within lists of genes with ease," *Genome Biol*, vol. 4, no. 10, p. R70, 2003.

[7] B. Efron and R. Tibshirani, "Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy," *Statistical science*, pp. 54–75, 1986.

Resampling Data:
Using a Statistical Jackknife

S. Sawyer — Washington University — March 11, 2005

1. Why Resample? Suppose that we want to estimate a parameter θ that depends on a random quantity sample $X = (X_1, X_2, \ldots, X_n)$ in a complicated way. For example, θ might be the sample variance of X or the log sample variance. If the X_i are vector valued, θ could be the Pearson correlation coefficient.

Assume that we have an estimator $\phi_n(X_1, X_2, \ldots, X_n)$ of θ but do not know the probability distribution of $\phi_n(X)$ given θ. This means that we cannot estimate the error involved in estimating θ by $\phi_n(X_1, \ldots, X_n)$, and that we cannot tell if we can conclude $\theta \neq 0$ from an observed $\phi_n(X) \neq 0$, no matter how large.

More generally, can we get a confidence interval for θ depending only on the observed $X_1, X_2, \ldots, X_n$, or test $H_0 : \theta = \theta_0$ just using the data $X_1, X_2, \ldots, X_n$?

Methods that try to estimate the bias and variability of an estimator $\phi_n(X_1, X_2, \ldots, X_n)$ by using the values of $\phi_n(X)$ on subsamples from $X_1, X_2, \ldots, X_n$ are called *resampling* methods. Two common resampling methods are the *jackknife*, which is discussed below, and the *bootstrap*.

The jackknife was invented by Quenouille in 1949 for the more limited purpose of correcting possible bias in $\phi_n(X_1, X_2, \ldots, X_n)$ for small n. Tukey in 1958 noticed that the procedure could be used to construct reasonably reliable confidence intervals for a wide variety of estimators $\phi_n(X)$, and so might be viewed as being as useful to a statistician as a regular jackknife would be to an outdoorsman. Bootstrap methods were invented by Bradley Efron around 1979. These are computationally more intensive (although easier to program) and give more accurate results in some cases.

2. The Jackknife Recipe. Let $\phi_n(X) = \phi_n(X_1, \ldots, X_n)$ be an estimator defined for samples $X = (X_1, X_2, \ldots, X_n)$. The i^{th} *pseudovalue* of $\phi_n(X)$ is

$$ps_i(X) = n\phi_n(X_1, X_2, \ldots, X_n) - (n-1)\phi_{n-1}(\,(X_1, \ldots, \ldots, X_n)_{[i]}) \qquad (1)$$

In (1), $X_{[i]}$ means the sample $X = (X_1, X_2, \ldots, X_n)$ with the i^{th} value X_i *deleted* from the sample, so that $X_{[i]}$ is a sample of size $n-1$. Note

$$ps_i(X) = \phi_n(X) + (n-1)\big(\phi_n(X) - \phi_{n-1}(X_{[i]})\big)$$

so that $ps_i(X)$ can be viewed as a bias-corrected version of $\phi_n(X)$ determined by the trend in the estimators $\phi_n(X)$ from $\phi_{n-1}(X_{[i]})$ to $\phi_n(X)$.

The basic jackknife recipe is to treat the pseudovalues $ps_i(X)$ as if they were independent random variables with mean θ. One can then obtain confidence intervals and carry out statistical tests using the Central Limit Theorem. Specifically, let

$$ps(X) = \frac{1}{n}\sum_{i=1}^{n} ps_i(X) \quad \text{and} \quad V_{ps}(X) = \frac{1}{n-1}\sum_{i=1}^{n}\big(ps_i(X) - ps(X)\big)^2 \quad (2)$$

be the mean and sample variance of the pseudovalues. The sample mean $ps(X)$ was Quenouille's (1949) bias-corrected version of $\phi_n(X)$. The jackknife 95% confidence interval for θ is

$$\left(ps(X) - 1.960\sqrt{\frac{1}{n}V_{ps}(X)}\,,\quad ps(X) + 1.960\sqrt{\frac{1}{n}V_{ps}(X)}\right) \quad (3)$$

Similarly, one can define a jackknife P-value for the hypothesis $H_0 : \theta = \theta_0$ by comparing

$$Z \;=\; \frac{\sqrt{n}\,\big(ps(X) - \theta_0\big)}{\sqrt{V_{ps}(X)}} \;=\; \frac{ps(X) - \theta_0}{\sqrt{(1/n)V_{ps}(X)}} \quad (4)$$

with a standard normal variable.

Remark: Technically speaking, the pseudovalues in (1) are for what is called the *delete-one* jackknife. There is also a more general *delete-k* or *block* jackknife that we discuss below.

3. Examples (1) If $\phi_n(X) = \frac{1}{n}\sum_{j=1}^{n} X_j = \overline{X}$ is the sample mean for $\theta = E(X_i)$, then the pseudovalues

$$ps_i(X) \;=\; n\overline{X} \;-\; (n-1)\overline{X_{[i]}} \;=\; X_i$$

are the same as the original values. Thus

$$ps(X) = \frac{1}{n}\sum_{i=1}^{n} ps_i(X) = \overline{X} \qquad \text{and} \qquad V_{ps}(X) = \frac{1}{n-1}\sum_{i=1}^{n}\big(X_i - \overline{X}\big)^2 \quad (5)$$

are the usual sample mean and variance.

(2) If $\phi_n(X) = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})^2$ is the sample variance, then, after some algebra, the pseudovalues of $\phi_n(X)$ are

$$ps_i(X) \;=\; \frac{n}{n-2}(X_i - \overline{X})^2 \;-\; \frac{1}{(n-1)(n-2)}\sum_{j=1}^{n}(X_j - \overline{X})^2 \quad (6)$$

The mean of the pseudovalues

$$ps(X) \;=\; \frac{1}{n}\sum_{i=1}^{n} ps_i(X) \;=\; \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})^2$$

is the same as $\phi_n(X)$ in this case also.

(3) If $\phi_n(X) = \frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X})^2$ is the sample variance with $1/(n-1)$ replaced by $1/n$, then the pseudovalues of $\phi_n(X)$ are

$$ps_i(X) \;=\; \frac{n}{n-1}(X_i - \overline{X})^2 \tag{7}$$

This implies that

$$ps(X) = \frac{1}{n}\sum_{i=1}^{n} ps_i(X) = \frac{1}{n-1}\sum_{i=1}^{n}\big(X_i - \overline{X}\big)^2$$

is the usual sample variance. Note that $E\big(\phi_n(X)\big) = \frac{n-1}{n}\sigma^2$ for $\sigma^2 = \mathrm{Var}(X)$ while $E\big(ps(X)\big) = \sigma^2$, so that $ps(X)$ is a bias-corrected version of $\phi_n(X)$.

4. A Simple Example. Suppose that we have four observations $\{1, 2, 3, 4\}$ with $\phi_4(X) = \overline{X}$. Thus $\phi_4(X) = (1/4)\sum_{i=1}^{4} X_i = (1/4)(1+2+3+4) = 2.5$.

The four delete-one values are $\phi_3(X_{[1]}) = (1/3)(2+3+4) = 3.0$, $\phi_3(X_{[2]}) = (1/3)(1+3+4) = 2.67$, $\phi_3(X_{[3]}) = (1/3)(1+2+4) = 2.33$, and $\phi_3(X_{[4]}) = (1/3)(1+2+3) = 2.00$.

The four pseudovalues are $ps_1(X) = 4\phi_4(X) - 3\phi_3(X_{[1]}) = 4(2.50) - 3(3.0) = 10 - 9 = 1.0$, $ps_2(X) = 4\phi_4(X) - 3\phi_3(X_{[2]}) = 4(2.50) - 3(2.67) = 10-8 = 2.0$, $ps_2(X) = 4\phi_4(X) - 3\phi_3(X_{[3]}) = 4(2.50) - 3(2.33) = 10-7 = 3.0$, and $ps_2(X) = 4\phi_4(X) - 3\phi_3(X_{[4]}) = 4(2.50) - 3(2.00) = 10 - 6 = 4.0$. Thus the four pseudovalues are the same as the original observations, as they should be for $\phi_n(X) = \overline{X}$.

5. Another Example. Suppose that we have 16 observations

17.23	13.93	15.78	14.91	18.21	14.28	18.83	13.45
18.71	18.81	11.29	13.39	11.57	10.94	15.52	15.25

and that we are interested in estimating the variance σ^2 of the data and in finding a 95% confidence interval for σ^2. In order to minimize any possible effect of outliers, we apply the jackknife to the log sample variance

$$\phi_n(X_1, \ldots, X_n) \;=\; \log(s^2) \;=\; \log\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\right)$$

instead of to s^2 directly. For these 16 observations, $s^2 = 7.171$ and $\phi_n(X) = \log(s^2) = 1.9701$.

The delete-one values $\phi_{n-1}(X_{[i]})$ on the 16 subsamples with $n - 1 = 15$ are

1.994	2.025	2.035	2.039	1.940	2.032	1.893	2.011
1.903	1.895	1.881	2.009	1.905	1.848	2.038	2.039

The corresponding pseudovalues $ps_i(X) = n\phi_n(X) - (n-1)\phi_{n-1}(X_{[i]})$ are

1.605	1.151	0.998	0.942	2.416	1.043	3.122	1.362
2.972	3.097	3.308	1.393	2.951	3.806	0.958	0.937

The mean of the pseudovalues is 2.00389, which is a little larger than the initial estimate $\phi_n(X) = 1.9701$. The sample variance of the 16 pseudovalues is 1.091. The jackknife 95% confidence interval for the log variance $\log(\sigma^2)$ is $(1.492, 2.516)$.

The 16 values in this case were drawn from a probability distribution whose true variance is 5.0, with $\log(5.0) = 1.609$, which is well within the 95% jackknife confidence interval.

6. The Delete-*k* or Block Jacknife. If n is large, the pseudovalues $ps_i(X)$ in (1) may be too close together, and the variance $V_{ps}(X)$ may be mostly sampling error. In that case, we can define a *block jackknife* instead of the *delete-one* jackknife defined above by proceeding as follow. Assume $n = n_b k$, where k will be the block size and n_b is the number of blocks. Define

$$ps_i(X) = n_b\phi_n(X_1, X_2, \ldots, X_n) - (n_b - 1)\phi_{n-k}(\,(X_1, \ldots, \ldots, X_n)_{[i]}\,) \quad (8)$$

instead of (1), where now $1 \le i \le n_b$ and $X_{[i]}$ means the sample $X = (X_1, X_2, \ldots, X_n)$ with the i^{th} *block* of k values — that is, with indices j in the range $ik + 1 \le j \le ik + k$ — removed.

For example, if $n = 200$, we might set $k = 20$ and $n_b = 10$. Each of the $n_b = 10$ pseudovalues (8) would be defined in terms of ϕ_{200} on the full sample and ϕ_{180} on a subsample of size 180. We then treat the 10 pseudovalues (8) as a sample of 10 independent values with mean θ and proceed as before.

7. A Warning and Another Example: The jackknife should NOT be applied if the estimator $\phi_n(X)$ is too discontinuous as a function of the X_i, or if $\phi_n(X)$ depends on one or a few values in X.

For example, suppose that $\phi_n(X)$ is the *sample median* of $X = (X_1, X_2, \ldots, X_n)$ where $X_1, \ldots, X_n$ are distinct. Then

Exercise: Suppose that $n = 2m$ is even. Prove that

(a) There exists two numbers a, b depending on $X_1, \ldots, X_n$ such that each value $\phi_{n-1}(X_{[i]})$ is either a or b but

(b) the bias-corrected mean $ps(X) = \phi_n(X)$.

How do these results change if $n = 2m + 1$ is odd?

8. Coverage Frequencies for Jackknife Confidence Intervals. As a test of the jackknife confidence interval (3), we generate 10,000 samples of size $n = 20$ from the probability distribution $12x(1-x)^2$ on the unit interval $(0, 1)$. (This is a beta density with parameters $\alpha = 2$ and $\beta = 3$.)

For each sample of size 20, we compute the jackknife 95% confidence (3) for the variance, and count the number of samples out of 10,000 for which the jackknife 95% confidence interval contains the true variance, which is 0.048 in this case.

The *coverage probability* of a (putative) confidence interval is the probability that it actually contains the true value. If it is supposed to be a 95% confidence interval, then the coverage probability should be as close to 0.95 as possible. If the coverage probability is higher, then the confidence intervals are too conservative. If the coverage probability is lower, then they are too small and we may be misled.

We do the same calculation for $\phi_n(X)$ replaced by the logarithm of the variance, and also for the jackknife 99% confidence interval (3) with 1.96 replaced by 2.576.

As a third test, we generate 10,000 samples of $n = 20$ pairs of standard normal variables (X_i, Y_i) with a theoretical Pearson correlation coefficient of $\rho = 0.50$, and apply the same procedures for the sample (Pearson) correlation coefficient.

Some typical jackknife confidence intervals in these cases are

	Beta: variance	Beta: log variance	Normal: ρ
95%:	(0.0221, 0.0655)	(-4.2420, -2.9155)	(0.1976, 0.8387)
99%:	(0.0153, 0.0723)	(-4.4505, -2.7071)	(0.0968, 0.9395)

The estimated coverage probabilities were

95%:	0.9025	0.9390	0.9017
99%:	0.9503	0.9779	0.9556

Thus the confidence intervals tend to be a bit small, but are approximately correct. Psuedovalues for the logarithm of variance are better behaved than pseudovalues for the variance itself, presumably because the effect of large sample variances is smaller.

We might expect the results to be worse if the sample were smaller or if the distribution was more heavy tailed, for example having an exponential instead of a beta distribution, but this remains to be checked.

Bootstrap: A Statistical Method

Kesar Singh and Minge Xie

Rutgers University

Abstract

This paper attempts to introduce readers with the concept and methodology of bootstrap in Statistics, which is placed under a larger umbrella of resampling. Major portion of the discussions should be accessible to any one who has had a couple of college level applied statistics courses. Towards the end, we attempt to provide glimpses of the vast literature published on the topic, which should be helpful to someone aspiring to go into the depth of the methodology. A section is dedicated to illustrate real data examples. We think the selected set of references cover the greater part of the developments on this subject matter.

1. Introduction and the Idea

B. Efron (1979) introduced the Bootstrap method. It spread like brush fire in statistical sciences within a couple of decades. Now if one conducts a "Google search" for the above title, an astounding 1.86 million records will be mentioned; scanning through even a fraction of these records is a daunting task. We attempt first to explain the idea behind the method and the purpose of it at a rather rudimentary level. The primary task of a statistician is to summarize a sample based study and generalize the finding to the parent population in a scientific manner. A technical term for a sample summary number is (sample) statistic. Some basic sample statistics are sample mean, sample median, sample standard deviation etc. Of course, a summary statistic like the sample mean will fluctuate from sample to sample and a statistician would like to know the magnitude of these fluctuations around the corresponding population parameter in an overall sense. This is then used in assessing Margin of Errors. The entire picture of all possible values of a sample statistics presented in the form of a probability distribution is called a sampling distribution. There is a plenty of theoretical knowledge of sampling distributions, which can be found in any text books of mathematical statistics. A general intuitive method applicable to just

about any kind of sample statistic that keeps the user away from the technical tedium has got its own special appeal. Bootstrap is such a method.

To understand bootstrap, suppose it were possible to draw repeated samples (of the same size) from the population of interest, a large number of times. Then, one would get a fairly good idea about the sampling distribution of a particular statistic from the collection of its values arising from these repeated samples. But, that does not make sense as it would be too expensive and defeat the purpose of a sample study. The purpose of a sample study is to gather information cheaply in a timely fashion. The idea behind bootstrap is to use the data of a sample study at hand as a "surrogate population", for the purpose of approximating the sampling distribution of a statistic; i.e. to resample (with replacement) from the sample data at hand and create a large number of "phantom samples" known as bootstrap samples. The sample summary is then computed on each of the bootstrap samples (usually a few thousand). A histogram of the set of these computed values is referred to as the bootstrap distribution of the statistic.

In bootstrap's most elementary application, one produces a large number of "copies" of a sample statistic, computed from these phantom bootstrap samples. Then, a small percentage, say $100(\alpha/2)\%$ (usually $\alpha = 0.05$), is trimmed off from the lower as well as from the upper end of these numbers. The range of remaining $100(1-\alpha)\%$ values is declared as the confidence limits of the corresponding unknown population summary number of interest, with level of confidence $100(1-\alpha)\%$. The above method is referred to as bootstrap percentile method. We shall return to it later in the article.

2. The Theoretical Support

Let us develop some mathematical notations for convenience. Suppose a population parameter θ is the target of a study; say for example, θ is the household median income of a chosen community. A random sample of size n yields the data $(X_1, X_2, ..., X_n)$. Suppose, the corresponding sample statistic computed from this data set is $\hat{\theta}$ (sample median in the case of the example). For most sample statistics, the sampling distribution of $\hat{\theta}$ for large n ($n \geq 30$ is generally accepted as large sample size), is bell shaped with center θ and standard deviation

($a/\sqrt{n}$**)**, where the positive number a depends on the population and the type of statistic $\hat{\theta}$. This phenomenon is the celebrated Central Limit Theorem (CLT). Often, there are serious technical complexities in approximating the required standard deviation from the data. Such is the case when $\hat{\theta}$ is sample median or sample correlation. Then bootstrap offers a bypass. Let $\hat{\theta}_B$ stand for a random quantity which represents the same statistic computed on a bootstrap sample drawn out of $(X_1, X_2, ..., X_n)$. What can we say about the sampling distribution of $\hat{\theta}_B$ (w.r.t. all possible bootstrap samples), while the original sample $(X_1, X_2, ..., X_n)$ is held fixed? The first two articles dealing with the theory of bootstrap – Bickel and Freedman (1981) and Singh (1981) provided large sample answers for most of the commonly used statistics. In limit, as ($n \to \infty$), the sampling distribution of $\hat{\theta}_B$ is also bell shaped with $\hat{\theta}$ as the center and the same standard deviation ($a/\sqrt{n}$). Thus, bootstrap distribution of $\hat{\theta}_B - \hat{\theta}$ approximates (fairly well) the sampling distribution of $\hat{\theta} - \theta$. Note that, as we go from one bootstrap sample to another, only $\hat{\theta}_B$ in the expression $\hat{\theta}_B - \hat{\theta}$ changes as $\hat{\theta}$ is computed on the original data $(X_1, X_2, ..., X_n)$. This is the bootstrap Central Limit Theorem. For a proof of bootstrap CLT for the mean, see Singh (1981).

Furthermore, it has been found that if the limiting sampling distribution of a statistical function does not involve population unknowns, bootstrap distribution offers a better approximation to the sampling distribution than the CLT. Such is the case when the statistical function is of the form $(\hat{\theta}_B - \hat{\theta})/SE$ where SE stands for true or sample estimate of the standard error of $\hat{\theta}$, in which case the limiting sampling distribution is usually standard normal. This phenomenon is referred to as the second order correction by bootstrap. A caution is warranted in designing bootstrap, for second order correction. For illustration, let $\theta = \mu$, the population mean, and $\hat{\theta} = \bar{X}$, the sample mean; σ = population standard deviation, s= sample standard deviation computed from original data and s_B is the sample standard deviation computed on a bootstrap sample. Then, the sampling distribution of $(\bar{X} - \mu)/SE$, with $SE = \sigma/\sqrt{n}$, will be approximated by the bootstrap distribution of $(\bar{X}_B - \bar{X})/\widehat{SE}$, with $\bar{X}_B$ = bootstrap sample mean and $\widehat{SE} = s/\sqrt{n}$.

Similarly, the sampling distribution of $(\bar{X}-\mu)/\widehat{SE}$, with $\widehat{SE}=s/\sqrt{n}$, will be approximated by the bootstrap distribution of $(\bar{X}_B-\bar{X})/SE_B$, with $SE_B=s_B/\sqrt{n}$. The earliest results on second order correction were reported in Singh (1981) and Babu and Singh (1983). In the subsequent years, a flood of large sample results on bootstrap with substantially higher depth, followed. A name among the researchers in this area that stands out is Peter Hall of Australian National University.

3. Primary Applications of Bootstrap

3.1 Approximating Standard Error of a Sample Estimate:

Let us suppose, information is sought about a population parameter θ. Suppose $\hat{\theta}$ is a sample estimator of θ based on a random sample of size n, i.e. $\hat{\theta}$ is a function of the data $(X_1, X_2, ..., X_n)$. In order to estimate standard error of $\hat{\theta}$, as the sample varies over the class of all possible samples, one has the following simple bootstrap approach:

Compute $(\theta_1^*, \theta_2^*, ..., \theta_N^*)$, using the same computing formula as the one used for $\hat{\theta}$, but now base it on N different bootstrap samples (each of size n). A crude recommendation for the size N could be $N=n^2$ (in our judgment), unless n^2 is too large. In that case, it could be reduced to an acceptable size, say $n\log_e n$. One defines $SE_B(\hat{\theta})=[(1/N)\sum_{i=1}^{N}(\theta_i^*-\hat{\theta})^2]^{1/2}$ following the philosophy of bootstrap: replace the population by the empirical population.

An older resampling technique used for this purpose is Jackknife, though bootstrap is more widely applicable. The famous example where Jackknife fails while bootstrap is still useful is that of $\hat{\theta}$ = the sample median.

3.2 Bias correction by bootstrap:

The mean of sampling distribution of $\hat{\theta}$ often differs from θ, usually by an amount = c/n for large n. In statistical language, one writes

$$Bias(\hat{\theta})=E(\hat{\theta})-\theta\approx O(1/n)\ .$$

A bootstrap based approximation to this bias is

$$\frac{1}{N}\sum_{i=1}^{N}\theta_i^* - \hat{\theta} = \widehat{Bias}_B(\hat{\theta}) \text{ (say),}$$

Where θ_i^* are bootstrap copies of $\hat{\theta}$, as defined in the earlier subsection. Clearly, this construction is also based on the standard bootstrap thinking: replace the population by the empirical population of the sample. The bootstrap bias corrected estimator is $\hat{\theta}_c = \hat{\theta} - \widehat{Bias}_B(\hat{\theta})$. It needs to be pointed out that the older resampling technique called Jackknife is more popular with statisticians for the purpose of bias estimation.

3.3 Bootstrap Confidence Intervals:

Confidence intervals for a given population parameter θ are sample based range $[\hat{\theta}_1, \hat{\theta}_2]$ given out for the unknown number θ. The range possesses the property that θ would lie within its bounds with a high (specified) probability. The latter is referred to as confidence level. Of course this probability is with respect to all possible samples, each sample giving rise to a confidence interval which thus depends on the chance mechanism involved in drawing the samples. The two mostly used levels of confidence are 95% and 99%. We limit ourselves to the level 95% for our discussion here. Traditional confidence intervals rely on the knowledge of sampling distribution of $\hat{\theta}$, exact or asymptotic as $n \to \infty$. Here are some standard brands of confidence intervals constructed using bootstrap.

Bootstrap Percentile Method:

This method was mentioned in the introduction itself, because of its popularity which is primarily due to its simplicity and natural appeal. Suppose one settles for 1000 bootstrap replications of $\hat{\theta}$, denoted by $(\theta_1^*, \theta_2^*, ..., \theta_{1000}^*)$. After ranking from bottom to top, let us denote these bootstrap values as $(\theta_{(1)}^*, \theta_{(2)}^*, ..., \theta_{(1000)}^*)$. Then the bootstrap percentile confidence interval at 95% level of confidence would be $[\theta_{(25)}^*, \theta_{(975)}^*]$. Turning to the theoretical aspects of this method, it should be pointed out that the method requires the symmetry of the sampling distribution of $\hat{\theta}$ around θ. The reason is that the method approximates the sampling distribution of $\hat{\theta} - \theta$ by the

bootstrap distribution of $\hat{\theta}-\hat{\theta}_B$, which is contrary to the bootstrap thinking that the sampling distribution of $\hat{\theta}-\theta$ could be approximated by the bootstrap distribution of $\hat{\theta}_B-\hat{\theta}$. Interested readers may check out Hall (1988).

Centered Bootstrap Percentile Method:

Suppose the sampling distribution of $\hat{\theta}-\theta$ is approximated by the bootstrap distribution of $\hat{\theta}_B-\hat{\theta}$, which is what the bootstrap prescribes. Denote 100s-th percentile of $\hat{\theta}_B$ (in bootstrap replications) by B_s. Then, the statement that $\hat{\theta}-\theta$ lies within the range $B_{.025}-\hat{\theta}$, $B_{.975}-\hat{\theta}$ would carry a probability $\approx .95$. But, this statement easily translates to the statement that θ lies within the range $(2\hat{\theta}-B_{.975}, 2\hat{\theta}-B_{.025})$. The latter range is what is known as centered bootstrap percentile confidence interval (at coverage level 95%). In terms of 1000 bootstrap replications $B_{.025}=\theta^*_{(25)}$ and $B_{.975}=\theta^*_{(975)}$.

Bootstrap-t Methods:

As it was mentioned in section 2, bootstrapping a statistical function of the form $T=(\hat{\theta}-\theta)/SE$ where SE is a sample estimate of the standard error of $\hat{\theta}$, brings extra accuracy. This additional accuracy is due to so called one-term Edgeworth correction by the bootstrap. The reader could find essential details in Hall (1992). The basic example of T is the standard t-statistics (from which the name bootstrap-t is derived): $t=(\bar{X}-\mu)/(s/\sqrt{n})$, which is a special case with $\theta=\mu$ (the population mean), $\hat{\theta}=\bar{X}$ (the sample mean) and s standing for the sample standard deviation. The bootstrap counterpart of such a function T is $T_B=(\hat{\theta}_B-\hat{\theta})/SE_B$ where SE_B is exactly like SE but computed on a bootstrap sample. Denote the 100s-th bootstrap percentile of T_B by b_s and consider the statement: T lies within $[b_{.025}, b_{.975}]$. After the substitution $T=(\hat{\theta}-\theta)/SE$, the above statement translates to 'θ lies within $(\hat{\theta}-SE\ b_{.975},\ \hat{\theta}-SE\ b_{.025})$'. This range for θ is called bootstrap-t based confidence interval for θ at coverage level 95%. Such an

interval is known to achieve higher accuracy than the earlier method, which is referred to as "second order accuracy" in technical literature.

We end the section with a remark that B. Efron proposed correction to the rudimentary percentile method to bring in extra accuracy. These corrections are known as Efron's "bias-correction" and "accelerated bias-correction". The details could be found in Efron and Tibishirani (1993). The bootstrap-t automatically takes care of such corrections, although the bootstrapper needs to look for a formula for SE which is avoided in the percentile method.

4. Some Real Data Example

Example 1. (Skewed Univariate Data) In the first example, the data are taken from (Hollander and Wolfe, 1999, page 63), which represent the effect of illumination (difference between counts with and without illumination) on the rate of beak-clapping among chick-embryos; see the end of the section. The boxplot suggests lack of normality of the population. We have carried out bootstrap analysis on the median and on the mean. A noteworthy finding is the lack of symmetry of bootstrap-t histogram, which differs from limiting normal curve. The 95% level confidence intervals coming from our analysis for both mean and the median (centered bootstrap percentile method) cover the range [10, 30], roughly speaking. This range represents overall difference (increase) in the beak-clapping counts per minute due to illumination.

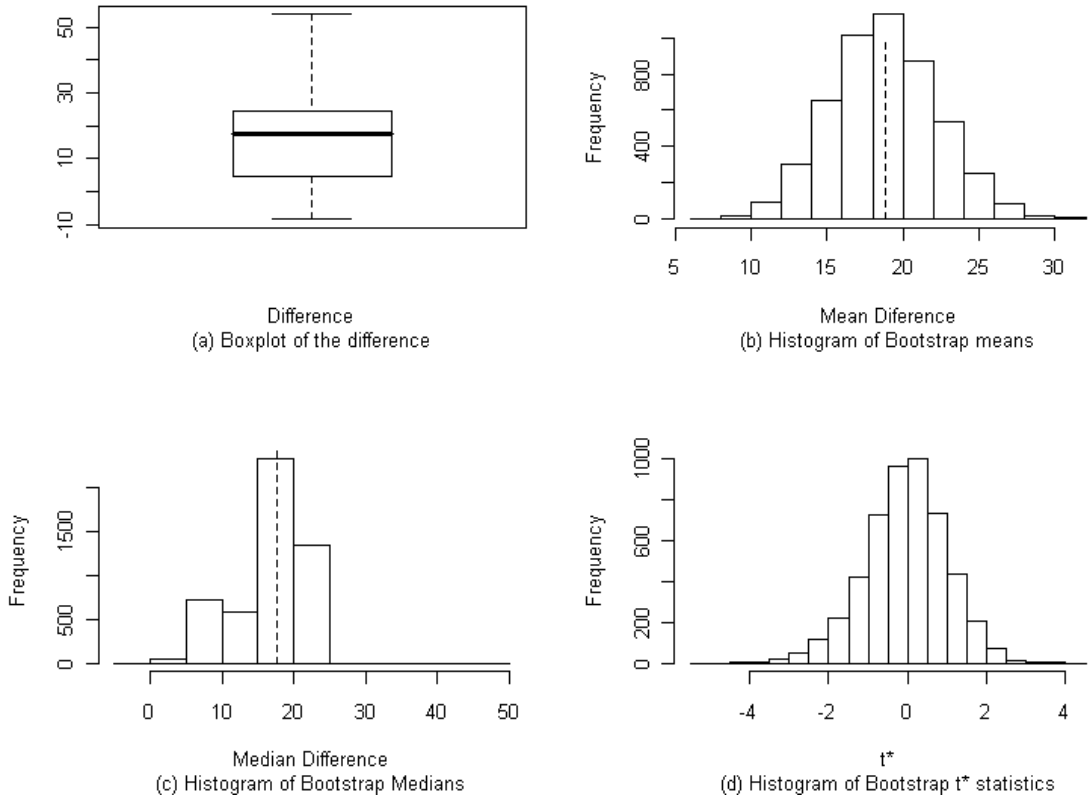
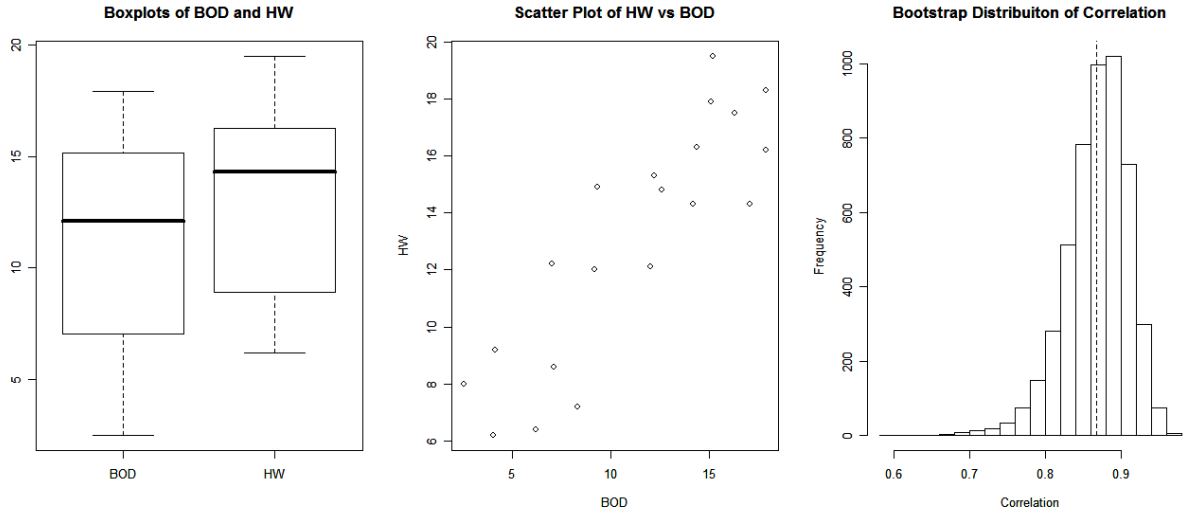


Figure 1. Boxplot of the measurement is presented in (a). Bootstrap distributions of the sample mean, sample median and t* statistic are plotted in (b)-(d), respectively. The dotted lines in (b) and (c) correspond respectively to the sample mean and sample median. Based the bootstrap distributions, the 95% confidence interval for the population median by the bootstrap percentile method is (4.7000, 24.7000), by the centered boostrap percentile is (10.5000, 30.5000). The 95% confidence interval for the population mean by the percentile bootstrap method is (10.0960, 28.1200), by the centered bootstrap method is (9.4880, 27.51200). The Bootstrap-t 95% CI for the population mean is (12.9413, 30.8147). Note that the bootstrap t on the mean show skewed histogram of the t-distribution.

Example 2. (Bivariate Data) In this example, the data are from Collins et al. (1999), which assess body fat in collegiate football players (Devore, 2003, page 553). We study correlation between the BOD and HW measurements; see the data at the end of this section. Here, BOD is BOD POD, a whole body air-displacement plethysmograph, and HW refers to hydrostatic weighing. The sample size is modest, but reasonable for bootstrap methods. As bivariate data consist of n pairs of data, say (X_i, Y_i), for $i = 1, ..., n$, one draws a pair of data randomly at a time in the bootstrap resampling. For instance, the first draw could be (X_7, Y_7) followed by (X_3, Y_3) etc.

The box plots of original data in Figure 1 (a) suggest lack of normality of the underlying populations. The histogram for correlations computed on bootstrap bivariate data is plotted in Figure 2 (c) which is asymmetric (skewed to the left). For this reason, the centered bootstrap percentile confidence interval appears more appropriate. According to our bootstrap analysis, the two measurements have at least a correlation of *0.78* in the population.



(a) Box plots (b) Scatter plot (c) Histogram

Figure 2. Boxplots of BOD and HW in (a) suggest somewhat non-normal data. Scatter plot in (b) indicates they are highly correlated. Bootstrap inference on the correlation between BOD and HW is presented in (c), which shows the Bootstrap distribution (in histogram) of correlation. IN particular, sample correlation of BOD and HW is 0.8679 which corresponds to the dotted vertical line in (c). The SE of the correlation is 0.0412 with an estimated bias of 0.0003. The 95% confidence interval of the correlation by the bootstrap percentile method is (0.7222, 0.9490) and the 95% confidence interval by the centered bootstrap percentile method is (0.7868, 1.0136).

Data for Example 1:

-8.5 -4.6 -1.8 -0.8 1.9 3.9 4.7 7.1 7.5 8.5 14.8 16.7 17.6 19.7 20.6 21.9 23.8 24.7 24.7 25.0 40.7 46.9 48.3 52.8 54.0

Data for Example 2:

BOD
2.5 4.0 4.1 6.2 7.1 7.0 8.3 9.2 9.3 12.0 12.2 12.6 14.2 14.4 15.1 15.2 16.3 17.1 17.9 17.9
HW
8.0 6.2 9.2 6.4 8.6 12.2 7.2 12.0 14.9 12.1 15.3 14.8 14.3 16.3 17.9 19.5 17.5 14.3 18.3 16.2

5. Engineering A Fitting Bootstrap

A sizable amount of journal literature on the topic is directed towards proposal and study of bootstrap schemes which will produce decent results in various statistical situations. The set up that has been the basis of forgoing discussion is basic and there are many types of departures from it. How to bootstrap in case of two stage sampling or a stratified sampling? Natural schemes are not hard to think of. Bootstrapping in case of data with regression models has attracted a lot of attention. There are two schemes which stand out: in one of which the covariate(s) and the response variable are resampled together (called paired bootstrap), and the other one bootstraps the "residuals" (=response – fitted model value) and then reconstructs the bootstrap regression data by plugging in the estimated regression parameters (called residual bootstrap). Paired bootstrap remains valid - in the sense of correct outcome in the limit as $n \to \infty$, even if the error variances in the model are unequal; a property which the residual bootstrap lacks. The shortcoming is compensated by the fact that the latter scheme brings additional accuracy in the estimation of standard error. This is the classic tug of war between efficiency and robustness in statistics (see Liu and Singh (1992)).

A lot harder to bootstrap are the time series data. Needless to say, time series analysis is of critical importance in several disciplines, especially in econometrics. The sources of difficulty are two-fold: (I) Time series data possess serial dependence i.e. X_{T+1} has dependence on X_T, X_{T-1} etc; (II)The statistical population changes with time, and that is known as non stationarity. It was noted very early on (see Singh (1981) for m-dependent data) that the classical bootstrap can not handle dependent data. A fair amount of research has been dedicated to modifying the bootstrap so that it could automatically bring in the dependence structure of the original sampling into bootstrap samples. The scheme of moving–block bootstrap has become quite well known (invented in Kunch (1989) and Liu and Singh (1992)). Potitis and Romano are well known authors on the topic, whose contributions have led to significant advancements on the topic of resampling, in general. In a moving block bootstrap scheme, one draws a block of data at a time, instead of one of the X_i's at a time, in order to preserve the underlying serial dependence structure that is present in the sample. There is plenty of ongoing research in the area of bootstrap methodology on econometric data.

6. The great m out n bootstrap with ($m/n \to 0$)

There are various types of conditions under which the straightforward bootstrap becomes inconsistent, meaning that the bootstrap estimate of sampling distribution and the true sampling distribution do not approach to the same limit, as the sample size n tends to ∞. That means, for large samples, one is bound to end up with an inaccurate statistical inference. The examples include, just to name a few, bootstrapping sample minimum or sample maximum which estimate end-point of a population distribution (Bickel and Freedman (1981)), the case of sample mean when the population variance is ∞ (Athreya (1981)), bootstrapping sample eigenvalues when population eigenvalues have multiplicity (Eaton and Tyler (1991)), the case of sample median when the population density is discontinuous at the population median (Huang, et.al. (1996)). Luckily, a general remedy exists and that is to keep the bootstrap sample size m much lower than the original size. Mathematically speaking, one requires $m \to \infty$ and $m/n \to 0$, as $n \to \infty$. In theory it fixes the problem, however for users, it is somewhat troublesome. How to choose m? An obvious suggestion would be settle for a fraction of n, say 20% or so. It should be pointed out that in good situations, where the regular bootstrap is fine, such a m is not advisable as it will result is loss of efficiency. See Bickel (2003), for a recent survey on the topic.

References:

Athreya, K.B. (1986). Bootstrap of the mean in the infinite variance case. *Ann. Stat*. 14, 724-731.

Azzalini, A. and Hall, P. (2000). Reducing variability using bootstrap methods with quantitative constraints. *Biometrika*, 87, 895-906.

Babu, G.J. (1984). Bootstrapping statistics with linear combination of Chi-square as weak limit. *Sankhya* A. 46, 85-93.

Babu, G.J. and Singh, K. (1983). Inference on means using the bootstrap. *Ann. Stat*. 11, 999-1003.

Beran, R. (1984). Prepivoting to reduce level errors of confidence sets. *Biometrika*. 74, 151-173.

Beran, R. (1990) Refining bootstrap simultaneous confidence sets. *Jour. Amer. Stat. Assoc*. 85, 417-428.

Bickel, P.J. (2003). Unorthodox bootstraps (invited papers). *J. of Korean Stat. Soc*. 32, 213-224.

Bickel, P.J. and Freedman, D. (1981). Some asymptotic theory for the bootstrap. *Ann. Stat*. 9, 1196- 1217.

Bickel, P.J. and Freedman, D (1984). Asymptotic normality and the bootstrap in stratified sampling. *Ann. Stat*. 12, 470-482.

Boos, D.D. and Brownie, C. (1989). Bootstrap methods for testing homogeneity of variances. Technometics. 31, 69-82.

Boos, D.D. and Munahan, J.F. (1986). Bootstrap methods using prior information. *Biometrika*. 73, 77-83.

Bose, A. (1988). Edgeworth correction by bootstrap in autoregressions. *Ann. Stat*. 16, 1709-1722.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 26, 123-140.

Buhlmann, P. (1994). Bootstrap empirical process for stationary sequences. *Ann. Stat*. 22, 995-1012.

Buhlmann, P. (2002). Sieve bootstrap with variable length – Markov chains for stationary categorical Time series (with discussions) *Jour. Amer. Stat. Assoc.* 97, 443-455.

Burr, D. (1994). A comparison of certain bootstrap confidence intervals in Cox model. *Jour. Amer. Stat. Assoc*. 89, 1290-1302.

Collins, M.A., Millard-Stafford, M.L., Sparling, P.B., Snow, T.K., Rosskopf, L.B., Webb, S.A., Omer, J. (1999). Evaluating BOD POD(R) for assessing body fat in collegiate football players. *Medicine and Science in Sports and Exercise*. 31,1350-56.

Davison, A.C. and Hinkley, D. V. (1988). Saddle point approximations in resampling method. *Biometrika*. 75, 417-431.

DiCiccio, T.J. and Romano, J.P. (1988). A review of bootstrap confidence intervals (with discussions). *J. R. Stat. Soc*. B. 50, 538-554.

Devore, L.J. (2003) PROBABILITY AND STATISTICS FOR ENGINEERING AND THE SCIENCE. Duxbury Press.

Eaton, M.L. and Tyler, D.E. (1991). On Wielandt's inequality and its application to the asymptotic distribution of the eigenvalues of a random symmetric matrix. *Ann. Stat.* **19**, 260–271.

Efron, B. (1979). Bootstrap methods: Another look at jackknife. *Ann. Stat*. 7, 1-26.

Efron, B. (1987). Better bootstrap confidence intervals (with discussions). *Jour. Amer. Stat. Assoc*. 82, 171-200.

Efron, B. (1992). Jackknife-after-bootstrap standard errors and influences functions (with discussions). *J.R. Stat. Soc*. B. 54, 83-127.

Efron, B. (1994). Missing data, imputation and the bootstrap (with discussions). *Jour. Amer. Stat. Assoc*. 89, 463-479.

Efron, B. and Tibshirani, R.J. (1993). *AN INTRODUCTION TO THE BOOTSTRAP*, Chapman and Hall New York.

Freedman, D.A. (1981) Bootstrapping Regression models. *Ann. Stat*. 9, 1281- 1228.

Hall, P. (1989). On efficient bootstrap simulation. *Boimetrika*. 76, 613-617.

Hall, P. (1992). Bootstrap confidence intervals in nonparametric regression. *Ann. Stat*. 20, 695-711.

Hall, P. (1992). THE *BOOTSTRAP AND EDGEWORTH EXPANSION*. Springer Verlag, N.Y.

Hall. P. (1988). Theoretical comparison of bootstrap confidence intervals (with discussions). *Ann. Stat*., 16, 927-953.

Hinkley, D.V. (1988). Bootstrap methods (with discussions*). J. Roy. Stat. Soc*. B, 50, 321-337.

Hollander, M. and Wolfe, D.A. (1999). *NONPARAMETRIC STATISTICAL METHODS* (2nd edition). John Wiley & Sons, N.Y.

Hwang, J.S., Sen, P.K. and Shao, J. (1996). Bootstrapping a sample quantile when the density has a jump. *Stat. Sinica*. 6, 1996.

Kunch, H.R. (1989). The jackknife and bootstrap for general stationary observations. *Ann. Stat*. 17, 1217-1241.

Lahiri, S.N. (1993). Bootstrapping the studentized sample mean of Lattice variables. *J. Mult. Analy*. 45, 247-256.

Lahiri, S.N. (1993). On the moving block bootstrap under long range dependence. *Stat. Prob. Letters*. 18, 405-413.

Liu, R.Y. and Singh, K. (1992). Efficiency and Robustness in re sampling. *Ann. Stat*. 20, 370-384.

Liu, R.Y. and Singh, K. (1992). Moving block jackknife and bootstrap capture weak dependence. *EXPLORING THE LIMITS OF BOOTSTRAP*, R. Lepage and L. Billard edited. Wiley, N.Y.

Lunneborg, EE. (2000). *DATA ANALYSIS BY RESAMPLING: CONCEPTS AND APLLICATIONS*. Duxbury Press.

Mamman, E. (1992). *WHEN DOES BOOTSTRAP WORK. ASYMPTOTOC RESULTS AND SIMULATIONS*. Springer Verlag, N.Y.

Politis, D.N. and Romano, J.P. (1994). The stationary bootstrap. *Jour. Amer. Stat. Assoc*. 89, 1303 – 1313.

Rubin, D.B. (1981). The Bayesian bootstrap. *Ann. Stat*. 9, 130-134.

Shao, J. and Tu, D. (1995). *THE JACKKNIFE AND BOOTSTRAP*, Springer, Verlag, N.Y.

Singh, K. (1981). On Asymptotic accuracy of Efron's bootstrap. *Ann. Stat*. 9, 1187-1195.

Singh, K (1998). Breakdown theory for bootstrap quantiles. *Ann. Stat*. 26, 1719-1732.

Singh, K. and Xie M. (2003). Bootlier-plot-Bootstrap based outlier detection plot. *Sankhya*, 65, 532-559.

Taylor, C.C. (1989). Bootstrap choice of smoothing parameter in kernel density estimation. *Biometrika*. 76, 705-712.

Tibshirani, R.J. (1988). Variance stabilization and the bootstrap. *Biometrika*. 75, 433-444.

Wu, C.F.J. (1986). Jackknife, bootstrap and other resampling procedures (with discussions). *Ann. Stat*. 14, 1261-1350.

Young, G.A.(1994) Bootstrap: More than a stab in the dark? (with discussion) *Stat. Scie*. 9, 382-415.

A Permutation Test

Its possible to take a completely different approach to hypothesis testing that assumes nothing about the distribution of the data. Rather than assuming a distribution for the data and using that to derive the distribution of the test statistic, we can instead work only with the experimental outcome data. Working with our algorithm running time example, we make a modification to hypothesis 2:

Null hypothesis 2A Assume that running times for algorithm A and B are identically distributed, independent random variables.

The distribution of running time could be anything at all. We didnt even say whether it is continuous or discrete. Even so, the null hypothesis does induce a probability distribution on the test statistic without any other information. Specifically, it implies that the samples are indistinguishable and exchangeable. Our test statistic is a mean of 6 values (algorithm B running times) minus the mean of 8 values (algorithm A running times). If the null hypothesis is true, these measurements are indistinguishable. So there is a test statistic distribution induced by all permutations of the 14 values with the first 8 labelled as "A" and the last 6 as "B". The resulting test statistic distribution is called the *permutation distribution.*

The permutation distribution has no simple analytic form, but it can be generated numerically by enumerating permutations and building a histogram. In practice the number of permutations is usually too large to enumerate. But we can approximate the permutation distribution to arbitrary accuracy by using enough *random permutations*. An approximate permutation distribution built from one million random permutations is shown in figure 1.

Like any other test statistic distribution, we can use this one to determine the p-value of the test statistic on our original data. The actual difference in mean times between algorithms A and B is 13.0. When we plot this on the test statistic distribution, we find that the fraction of permutations of the original data which gave an equal or higher test value was 0.0188 (the two-sided p-value is 0.0376). This p-value is remarkably close to the analytic value we derived for hypothesis 2 using a t-distribution.

The permutation test above is an example of a non-parametric test. We made no assumption about the distribution under the null hypothesis. Usually, when you weaken the assumptions in the null hypothesis it becomes harder to reject as our earlier example showed. But the permutation test rivaled the sensitivity of a parametric t-test assuming equal variances. This wasnt a fluke. Permutation tests often rival or even exceed the performance of parametric tests. They are remarkably simple to design and implement - the only "hard" part is generating enough random permutations to get the desired accuracy. But on today's machines the computing time needed is very small.

Monte-Carlo Permutation Tests

Our random permutation test was an approximation to a true permutation test: instead of enumerating all possible ($14! \approx 8.7 \times 10^{10}$) permutations, we sampled them using a Monte-Carlo

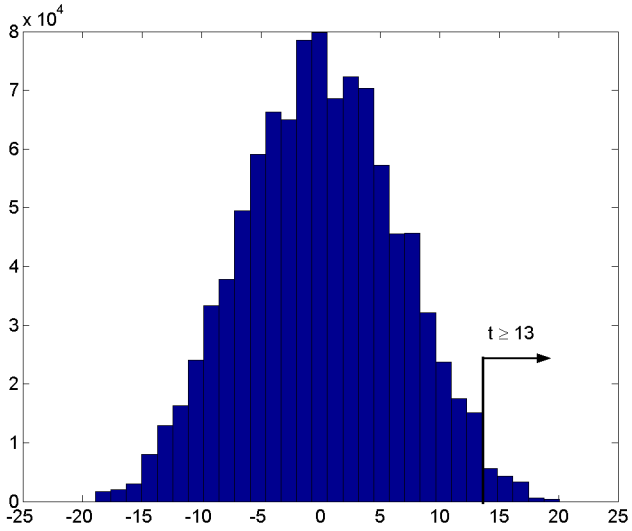


Figure 1: Permutation distribution on the observed data and the test statistic at $t \geq 13$

method. This gave us a p-value which was close but not exactly the same as the p-value for the true permutation distribution. We should estimate the variance of this p-value.

Let $v_1, \ldots, v_n$ be a sequence of observations which are iid under the null hypothesis. let $v_\pi = v_{\pi(1)}, \ldots, v_{\pi(n)}$ denote the sequence re-ordered by a permutation π. Let $t(v)$ denote the value of the test statistic on the given sequence v. Now assume π is a uniformly-chosen random permutation of $1, \ldots, n$. Given this distribution let p denote the probability $\Pr(t(v_\pi) \geq t(v))$. That is, p is the actual p-value for the test statistic on the ideal permutation distribution. Since there are $n!$ equally-likely permutations, the value of p will be rational and will have the form $m/n!$ for some integer m.

Now suppose we run a Monte-Carlo permutation test with N random permutations. Let π_k be one of these permutations, and X_k be an indicator variable define such that

$$X_k = \begin{cases} 1 & \text{if } t(v_{\pi_k}) \geq t(v) \\ 0 & \text{otherwise} \end{cases}$$

then X_k is a poisson trial with probability p. The variable $X = \sum_{k=1}^{N} X_k$ is a binomial variable with expected value Np and variance $Np(1-p)$. Our estimate of the actual p-value is found by dividing the observed X by the number of permutations:

$$p_{\text{est}} = X/N$$

so $\mathrm{E}[p_{\text{est}}] = p$ and $\mathrm{Var}[p_{\text{est}}] = p(1-p)/N$. A more useful representation of the variance in the estimate of p is its *relative error* which is its standard deviation divided by its value.

Theorem 1
Let $v_1, \ldots, v_n$ be a given sequence of n iid observed experimental values and $t(\cdot) \to \mathbb{R}$ be a test statistic. For the permutation test, N iid random permutations $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$ are chosen. Let p be the actual p-value for the test, which is the probability $\Pr(t(v_\pi) \geq t(v))$ for such random π. Then the relative error in the estimate p_{est} of p is bounded by:

$$\text{Rerr}(p_{\text{est}}) \leq \frac{1}{\sqrt{pN}}$$

Proof
The relative error in p_{est} is $\text{Std}[p_{\text{est}}]/\text{E}[p_{\text{est}}]$ which is $\sqrt{\text{Var}[p_{\text{est}}]}/p$. Since p is usually close to zero, $\text{Var}[p_{\text{est}}] = p(1-p)/N \leq p/N$. Substituting:

$$\text{Rerr}(p_{\text{est}}) \leq \sqrt{p/N}/p = \frac{1}{\sqrt{pN}}$$

Last lecture, we compared the running times of algorithms A and B using a million-sample permutation test which gave a p-value of 0.0188. From the above theorem, we see that the relative error in this estimate is about $1/\sqrt{0.02 \times 10^6} \approx 0.7$ percent. This is good enough for almost any practical purpose.

But note that the relative error depends on the p-value. If test yields a smaller value, e.g. 0.0001, the million-sample test would have a relative error of 10 percent. In this case, it would be wise to run more samples since the relative error decreases with N. Fortunately, the time needed to run even hundreds of millions of samples on today's computers is very small.

Since X and p_{est} are computed form poisson trials, we can use Chernoff bounds to bounded the probability of an error of a given size.

Theorem 2
Let $v_1, \ldots, v_n$, $t(\cdot)$, and π be defined as per theorem 1. Then

$$\Pr(p_{\text{est}} \geq (1+\delta)p) \leq \exp(-Np\delta^2/3) \qquad \text{and} \qquad \Pr(p_{\text{est}} \leq (1-\delta)p) \leq \exp(-Np\delta^2/2)$$

The proof is straightforward application of the Chernoff bounds for sums of poisson trials to the random variable X.

Bounds of this form are convenient for determining *confidence intervals*. We would like to say with some given confidence (i.e. with low probability of error), that an estimated value lies in a given range. For instance, a 98-percent CI can be built by limiting the probability for each of the above errors to 1 percent. If N is one million, we can solve for δ in both bounds. In the first (upper tail) bound, we find $\delta = 0.027$. For the second lower tail bound, we find $\delta = 0.022$. It follows that with 98% probability,

$$\frac{1}{1+0.027} p_{\text{est}} < p < \frac{1}{1-0.022} p_{\text{est}} \qquad \text{or} \qquad p \in [0.9736, 1.0226] \times p_{est}$$

And similarly to improving relative error, we can increase N to either reduce the error probability for the interval, or to derive tighter bounds for the interval.

The Power of the Permutation Test

Our running-time example from last lecture showed that the permutation test was very similar to the t-test in its p-value estimate. This was very encouraging. But it would be good to have a more systematic understanding of the relationship between permutation tests and classical statistical tests.

This is trickier than it sounds. Since the t-test is a parametric test and the permutation test is non-parametric, we are trying to compare apples and oranges. Actually, it is more like comparing apples and fruits. T-tests can only be applied when the null hypothesis is a normal distribution, while parametric tests are much more widely applicable. But more than that, we are interested now in the *power* of these tests, that is in how likely they are to reject the null hypothesis when an alternate hypothesis is true. To make such an evaluation, we have to assume specific distributions for the alternate hypothesis. For the t-test, the most natural alternate hypothesis is that the two samples come from *different* normal distributions. For these alternates, it has been shown that the t-test based on the difference in sample means is the most powerful test. For large samples, the power of the permutation test using the difference in sample means is equal to the t-test [1] for normally-distributed alternates.

So using the permutation test seems to give us the best of both worlds. It is non-parametric, and so can be applied even if we dont know what the population distributions are. If the populations are normally distributed, we lose very little power by applying the permutation test compared to an ideal test (the t-test).

1 Bootstrapping

Bootstrapping is the process of drawing a random sample *with replacement* from an experimental sample. Bootstrapping provides a non-parametric way to estimate the statistics of a population from which an experimental sample was drawn. It is particularly helpful for computing confidence intervals from data samples. For instance, our table of running times from last lecture was:

Algorithm A	95.7	83.3	101.2	102.9	88.5	111.9	112.0	99.6
Algorithm B	112.9	96.6	117.1	126.3	103.1	118.6		

Suppose we concluded that the two populations are different based on a permutation or t-test. Now we would like to know the range of means for each of the two algorithms.

A *bootstrap sample* from the algorithm A data set

$$A = \{95.7, 83.3, 101.2, 102.9, 88.5, 111.9, 112.0, 99.6\}$$

is a random sample of the *same size* drawn with repetition. e.g. the sample

$$A' = \{101.2, 111.9, 101.2, 88.5, 83.3, 99.6, 83.3, 102.9\}$$

is a boostrap sample from A. A histogram of the means of 1 million bootstrap samples is shown in figure 2.

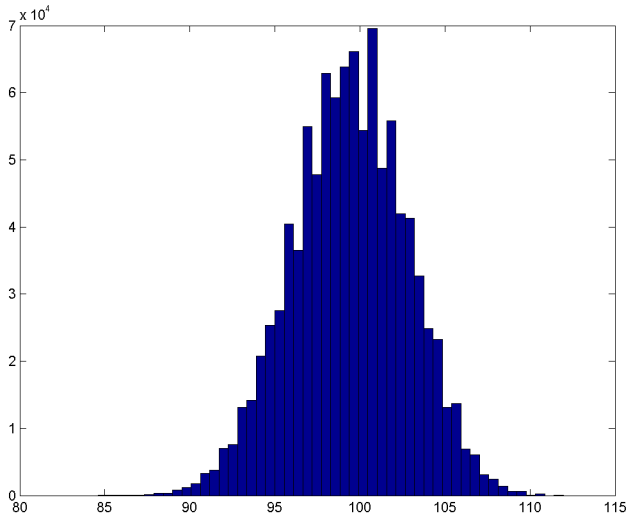


Figure 2: Histogram of means of 1 million bootstrap samples of algorithm A's running time data

Bootstrapping: Same-sized Samples

A bootstrap sample is exactly the same size as the original data sample. That is, if we have $n = 8$ samples for algorithm A, we should choose 8 (re)samples for any bootstrap sample for A. Since we are drawing with replacement, a bootstrap sample could have any size at all. But the goal is to produce a sample with the same statistics as the original sample. e.g. Bootstrap samples have approximately the same variance as samples from the original distribution. So to measure the variance of the *mean* of bootstrap samples (which is the sample variance divided by n), we should use exactly n samples in every bootstrap.

Bootstrap Confidence Intervals

To obtain a 90% confidence interval for a population mean from a sample, we find the values at the 5% and 95% percentiles of the histogram. These are respectively, the values of rank 50,000 and 950,000 in an ordered list of the re-sampled means. The resulting confidence interval for the running time of algorithm A is $[93.775, 104.8]$.

We can apply the same resampling to the data for algorithm B. In that case, the 90% confidence interval is $[105.6, 118.9]$.

This simple approach is a *primitive bootstrap confidence interval* or CI. If X is a sample mean, then ideally we want an interval $[l, u]$ such that $\Pr(X < l)$ and $\Pr(X > u)$ are exactly 5% or whatever our desired confidence is. In fact for the primitive bootstrap, $\Pr(X < l) = \alpha + O(n^{-1/2})$, where α is the desired confidence level.

Definition
A confidence interval $[l, u]$ for X is said to be first order accurate if $\Pr(X < l)$ and $\Pr(X > u)$ are $\alpha + O(n^{-1/2})$, where α is the desired confidence level.

A confidence interval $[l, u]$ for X is said to be second order accurate if $\Pr(X < l)$ and $\Pr(X > u)$ are $\alpha + O(n^{-1})$, where α is the desired confidence level.

i.e. the primitive bootstrap is first-order accurate. The main problem with the primitive bootstrap is that it exhibits *bias*. The actual percentiles of the population distribution differ from the resampled population. The error is minimized when the bootstrap distribution is close to normal. There is a standard transformation for doing this, which leads to the *bias-corrected bootstrap CI*. The bias-corrected 90% CI for the mean of algorithm A is $[93.56, 104.74]$, very close to our primitive bootstrap CI. However, it should be noted that the data for the example were generated from normal distributions and so bias was not significant.

References

[1] Bickel, P.M. and Van Zwet, W.R. Asymptotic expansion for the power of distribution-free tests in the two-sample problem. *Annals of Statistics*, 6, 987-1004, 1987.

Chapter 11

Simple Linear Regression and Correlation

11.1 Introduction to Linear Regression

Often, in practice, one is called upon to solve problems involving sets of variables when it is known that there exists some inherent relationship among the variables. For example, in an industrial situation it may be known that the tar content in the outlet stream in a chemical process is related to the inlet temperature. It may be of interest to develop a method of prediction, that is, a procedure for estimating the tar content for various levels of the inlet temperature from experimental information. Now, of course, it is highly likely that for many example runs in which the inlet temperature is the same, say 130°C, the outlet tar content will not be the same. This is much like what happens when we study several automobiles with the same engine volume. They will not all have the same gas mileage. Houses in the same part of the country that have the same square footage of living space will not all be sold for the same price. Tar content, gas mileage (mpg), and the price of houses (in thousands of dollars) are natural **dependent variables**, or responses, in these three scenarios. Inlet temperature, engine volume (cubic feet), and square feet of living space are, respectively, natural **independent variables**, or **regressors**. A reasonable form of a relationship between the **response** $\boldsymbol{Y}$ and the regressor x is the linear relationship

$$Y = \beta_0 + \beta_1 x,$$

where, of course, β_0 is the **intercept** and β_1 is the **slope**. The relationship is illustrated in Figure 11.1.

If the relationship is exact, then it is a **deterministic** relationship between two scientific variables and there is no random or probabilistic component to it. However, in the examples listed above, as well as in countless other scientific and engineering phenomena, the relationship is not deterministic (i.e., a given x does not always give the same value for Y). As a result, important problems here are probabilistic in nature since the relationship above cannot be viewed as being exact. The concept of **regression analysis** deals with finding the best relationship

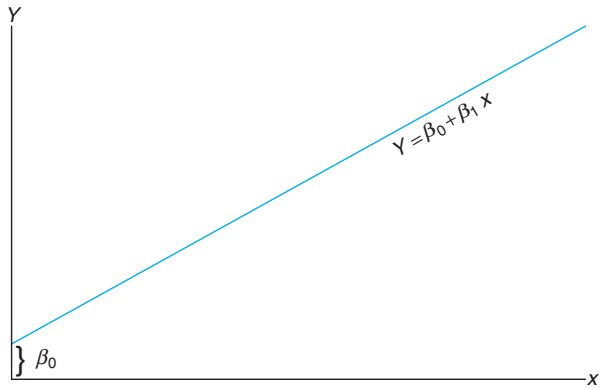


Figure 11.1: A linear relationship; β_0: intercept; β_1: slope.

between Y and x, quantifying the strength of that relationship, and using methods that allow for prediction of the response values given values of the regressor x.

In many applications, there will be more than one regressor (i.e., more than one independent variable **that helps to explain Y**). For example, in the case where the response is the price of a house, one would expect the age of the house to contribute to the explanation of the price, so in this case the multiple regression structure might be written

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

where Y is price, x_1 is square footage, and x_2 is age in years. In the next chapter, we will consider problems with multiple regressors. The resulting analysis is termed **multiple regression**, while the analysis of the single regressor case is called **simple regression**. As a second illustration of multiple regression, a chemical engineer may be concerned with the amount of hydrogen lost from samples of a particular metal when the material is placed in storage. In this case, there may be two inputs, storage time x_1 in hours and storage temperature x_2 in degrees centigrade. The response would then be hydrogen loss Y in parts per million.

In this chapter, we deal with the topic of **simple linear regression**, treating only the case of a single regressor variable in which the relationship between y and x is linear. For the case of more than one regressor variable, the reader is referred to Chapter 12. Denote a random sample of size n by the set $\{(x_i, y_i);\ i = 1, 2, \ldots, n\}$. If additional samples were taken using exactly the same values of x, we should expect the y values to vary. Hence, the value y_i in the ordered pair (x_i, y_i) is a value of some random variable Y_i.

11.2 The Simple Linear Regression (SLR) Model

We have already confined the terminology *regression analysis* to situations in which relationships among variables are not deterministic (i.e., not exact). In other words, there must be a **random component** to the equation that relates the variables.

This random component takes into account considerations that are not being measured or, in fact, are not understood by the scientists or engineers. Indeed, in most applications of regression, the linear equation, say $Y = \beta_0 + \beta_1 x$, is an approximation that is a simplification of something unknown and much more complicated. For example, in our illustration involving the response $Y =$ tar content and $x =$ inlet temperature, $Y = \beta_0 + \beta_1 x$ is likely a reasonable approximation that may be operative within a confined range on x. More often than not, the models that are simplifications of more complicated and unknown structures are linear in nature (i.e., linear in the **parameters** β_0 and β_1 or, in the case of the model involving the price, size, and age of the house, linear in the **parameters** β_0, β_1, and β_2). These linear structures are simple and empirical in nature and are thus called **empirical models**.

An analysis of the relationship between Y and x requires the statement of a **statistical model**. A model is often used by a statistician as a representation of an **ideal** that essentially defines how we perceive that the data were generated by the system in question. The model must include the set $\{(x_i, y_i);\ i = 1, 2, \ldots, n\}$ of data involving n pairs of (x, y) values. One must bear in mind that the value y_i depends on x_i via a linear structure that also has the random component involved. The basis for the use of a statistical model relates to how the random variable Y moves with x and the random component. The model also includes what is assumed about the statistical properties of the random component. The statistical model for simple linear regression is given below. The response Y is related to the independent variable x through the equation

Simple Linear Regression Model

$$Y = \beta_0 + \beta_1 x + \epsilon.$$

In the above, β_0 and β_1 are unknown intercept and slope parameters, respectively, and ϵ is a random variable that is assumed to be distributed with $E(\epsilon) = 0$ and $\text{Var}(\epsilon) = \sigma^2$. The quantity σ^2 is often called the error variance or residual variance.

From the model above, several things become apparent. The quantity Y is a random variable since ϵ is random. The value x of the regressor variable is not random and, in fact, is measured with negligible error. The quantity ϵ, often called a **random error** or **random disturbance**, has constant variance. This portion of the assumptions is often called the **homogeneous variance assumption**. The presence of this random error, ϵ, keeps the model from becoming simply a deterministic equation. Now, the fact that $E(\epsilon) = 0$ implies that at a specific x the y-values are distributed around the **true**, or population, **regression line** $y = \beta_0 + \beta_1 x$. If the model is well chosen (i.e., there are no additional important regressors and the linear approximation is good within the ranges of the data), then positive and negative errors around the true regression are reasonable. We must keep in mind that in practice β_0 and β_1 are not known and must be estimated from data. In addition, the model described above is conceptual in nature. As a result, we never observe the actual ϵ values in practice and thus we can never draw the true regression line (but we assume it is there). We can only draw an estimated line. Figure 11.2 depicts the nature of hypothetical (x, y) data scattered around a true regression line for a case in which only $n = 5$ observations are available. Let us emphasize that what we see in Figure 11.2 is not the line that is used by the

scientist or engineer. Rather, the picture merely describes what the assumptions mean! The regression that the user has at his or her disposal will now be described.

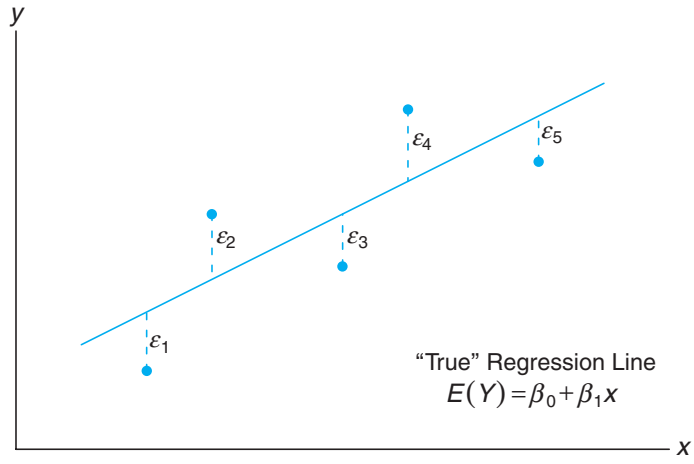


Figure 11.2: Hypothetical (x, y) data scattered around the true regression line for $n = 5$.

The Fitted Regression Line

An important aspect of regression analysis is, very simply, to estimate the parameters β_0 and β_1 (i.e., estimate the so-called **regression coefficients**). The method of estimation will be discussed in the next section. Suppose we denote the estimates b_0 for β_0 and b_1 for β_1. Then the estimated or **fitted regression** line is given by

$$\hat{y} = b_0 + b_1 x,$$

where $\hat{y}$ is the predicted or fitted value. Obviously, the fitted line is an estimate of the true regression line. We expect that the fitted line should be closer to the true regression line when a large amount of data are available. In the following example, we illustrate the fitted line for a real-life pollution study.

One of the more challenging problems confronting the water pollution control field is presented by the tanning industry. Tannery wastes are chemically complex. They are characterized by high values of chemical oxygen demand, volatile solids, and other pollution measures. Consider the experimental data in Table 11.1, which were obtained from 33 samples of chemically treated waste in a study conducted at Virginia Tech. Readings on x, the percent reduction in total solids, and y, the percent reduction in chemical oxygen demand, were recorded.

The data of Table 11.1 are plotted in a **scatter diagram** in Figure 11.3. From an inspection of this scatter diagram, it can be seen that the points closely follow a straight line, indicating that the assumption of linearity between the two variables appears to be reasonable.

Table 11.1: Measures of Reduction in Solids and Oxygen Demand

Solids Reduction, x (%)	Oxygen Demand Reduction, y (%)	Solids Reduction, x (%)	Oxygen Demand Reduction, y (%)
3	5	36	34
7	11	37	36
11	21	38	38
15	16	39	37
18	16	39	36
27	28	39	45
29	27	40	39
30	25	41	41
30	35	42	40
31	30	42	44
31	40	43	37
32	32	44	44
33	34	45	46
33	32	46	46
34	34	47	49
36	37	50	51
36	38		

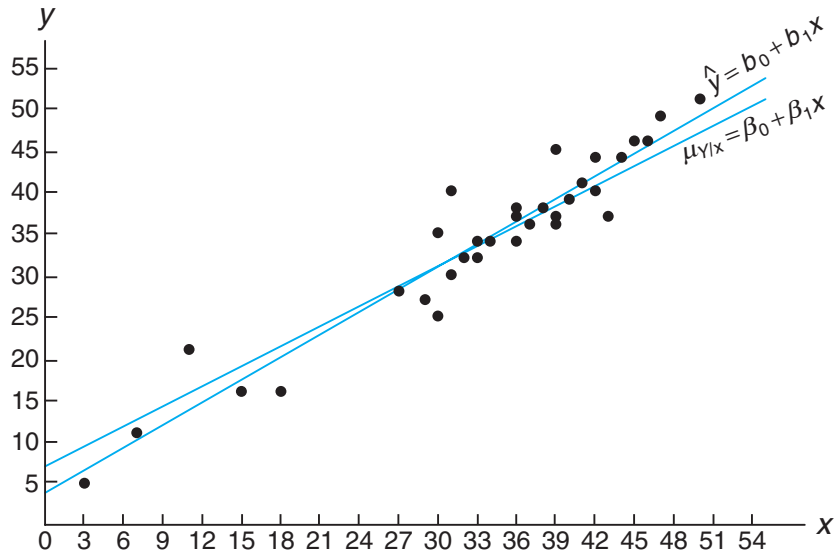


Figure 11.3: Scatter diagram with regression lines.

The fitted regression line and a *hypothetical true regression line* are shown on the scatter diagram of Figure 11.3. This example will be revisited as we move on to the method of estimation, discussed in Section 11.3.

Another Look at the Model Assumptions

It may be instructive to revisit the simple linear regression model presented previously and discuss in a graphical sense how it relates to the so-called true regression. Let us expand on Figure 11.2 by illustrating not merely where the ϵ_i fall on a graph but also what the implication is of the normality assumption on the ϵ_i.

Suppose we have a simple linear regression with $n = 6$ evenly spaced values of x and a single y-value at each x. Consider the graph in Figure 11.4. This illustration should give the reader a clear representation of the model and the assumptions involved. The line in the graph is the true regression line. The points plotted are actual (y, x) points which are scattered about the line. Each point is on its own normal distribution with the center of the distribution (i.e., the mean of y) falling on the line. This is certainly expected since $E(Y) = \beta_0 + \beta_1 x$. As a result, the true regression line **goes through the means of the response**, and the actual observations are on the distribution around the means. Note also that all distributions have the same variance, which we referred to as σ^2. Of course, the deviation between an individual y and the point on the line will be its individual ϵ value. This is clear since

$$y_i - E(Y_i) = y_i - (\beta_0 + \beta_1 x_i) = \epsilon_i.$$

Thus, at a given x, Y and the corresponding ϵ both have variance σ^2.

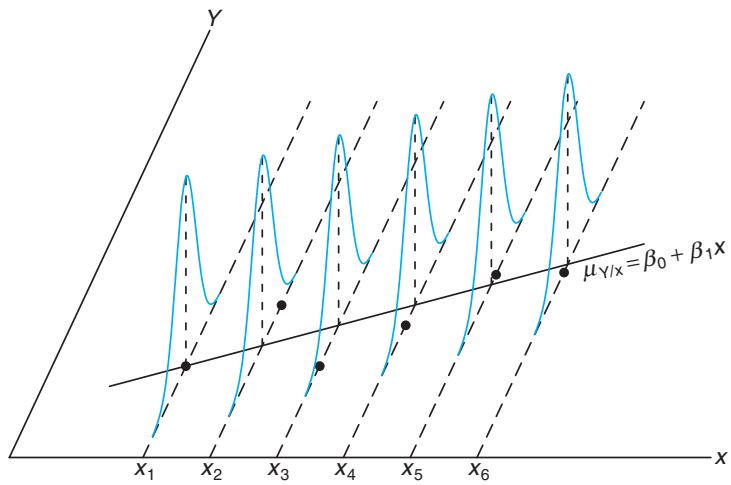


Figure 11.4: Individual observations around true regression line.

Note also that we have written the true regression line here as $\mu_{Y|x} = \beta_0 + \beta_1 x$ in order to reaffirm that the line goes through the mean of the Y random variable.

11.3 Least Squares and the Fitted Model

In this section, we discuss the method of fitting an estimated regression line to the data. This is tantamount to the determination of estimates b_0 for β_0 and b_1

for β_1. This of course allows for the computation of predicted values from the fitted line $\hat{y} = b_0 + b_1 x$ and other types of analyses and diagnostic information that will ascertain the strength of the relationship and the adequacy of the fitted model. Before we discuss the method of least squares estimation, it is important to introduce the concept of a **residual**. A residual is essentially an error in the fit of the model $\hat{y} = b_0 + b_1 x$.

Residual: Error in Fit

Given a set of regression data $\{(x_i, y_i); i = 1, 2, \ldots, n\}$ and a fitted model, $\hat{y}_i = b_0 + b_1 x_i$, the ith residual e_i is given by

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \ldots, n.$$

Obviously, if a set of n residuals is large, then the fit of the model is not good. Small residuals are a sign of a good fit. Another interesting relationship which is useful at times is the following:

$$y_i = b_0 + b_1 x_i + e_i.$$

The use of the above equation should result in clarification of the distinction between the residuals, e_i, and the conceptual model errors, ϵ_i. One must bear in mind that whereas the ϵ_i are not observed, the e_i not only are observed but also play an important role in the total analysis.

Figure 11.5 depicts the line fit to this set of data, namely $\hat{y} = b_0 + b_1 x$, and the line reflecting the model $\mu_{Y|x} = \beta_0 + \beta_1 x$. Now, of course, β_0 and β_1 are unknown parameters. The fitted line is an estimate of the line produced by the statistical model. Keep in mind that the line $\mu_{Y|x} = \beta_0 + \beta_1 x$ is not known.

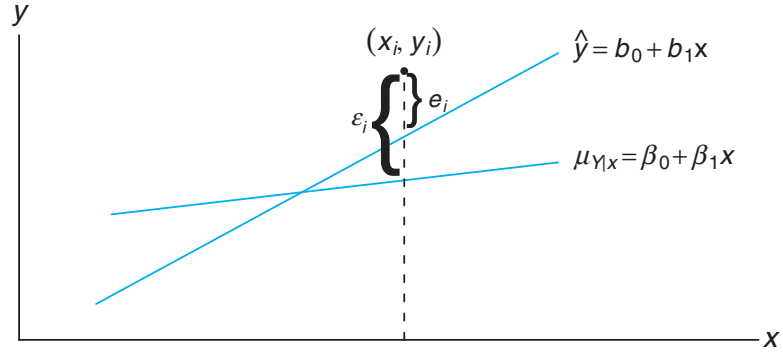


Figure 11.5: Comparing ϵ_i with the residual, e_i.

The Method of Least Squares

We shall find b_0 and b_1, the estimates of β_0 and β_1, so that the sum of the squares of the residuals is a minimum. The residual sum of squares is often called the sum of squares of the errors about the regression line and is denoted by SSE. This

minimization procedure for estimating the parameters is called the **method of least squares**. Hence, we shall find a and b so as to minimize

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2.$$

Differentiating SSE with respect to b_0 and b_1, we have

$$\frac{\partial(SSE)}{\partial b_0} = -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i), \quad \frac{\partial(SSE)}{\partial b_1} = -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)x_i.$$

Setting the partial derivatives equal to zero and rearranging the terms, we obtain the equations (called the **normal equations**)

$$nb_0 + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \quad b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i,$$

which may be solved simultaneously to yield computing formulas for b_0 and b_1.

Estimating the Regression Coefficients

Given the sample $\{(x_i, y_i);\ i = 1, 2, \ldots, n\}$, the least squares estimates b_0 and b_1 of the regression coefficients β_0 and β_1 are computed from the formulas

$$b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \text{ and}$$

$$b_0 = \frac{\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i}{n} = \bar{y} - b_1 \bar{x}.$$

The calculations of b_0 and b_1, using the data of Table 11.1, are illustrated by the following example.

Example 11.1: Estimate the regression line for the pollution data of Table 11.1.

***Solution*:**

$$\sum_{i=1}^{33} x_i = 1104, \quad \sum_{i=1}^{33} y_i = 1124, \quad \sum_{i=1}^{33} x_i y_i = 41{,}355, \quad \sum_{i=1}^{33} x_i^2 = 41{,}086$$

Therefore,

$$b_1 = \frac{(33)(41{,}355) - (1104)(1124)}{(33)(41{,}086) - (1104)^2} = 0.903643 \text{ and}$$

$$b_0 = \frac{1124 - (0.903643)(1104)}{33} = 3.829633.$$

Thus, the estimated regression line is given by

$$\hat{y} = 3.8296 + 0.9036x.$$

Using the regression line of Example 11.1, we would predict a 31% reduction in the chemical oxygen demand when the reduction in the total solids is 30%. The

31% reduction in the chemical oxygen demand may be interpreted as an estimate of the population mean $\mu_{Y|30}$ or as an estimate of a new observation when the reduction in total solids is 30%. Such estimates, however, are subject to error. Even if the experiment were controlled so that the reduction in total solids was 30%, it is unlikely that we would measure a reduction in the chemical oxygen demand exactly equal to 31%. In fact, the original data recorded in Table 11.1 show that measurements of 25% and 35% were recorded for the reduction in oxygen demand when the reduction in total solids was kept at 30%.

What Is Good about Least Squares?

It should be noted that the least squares criterion is designed to provide a fitted line that results in a "closeness" between the line and the plotted points. There are many ways of measuring closeness. For example, one may wish to determine b_0 and b_1 for which $\sum_{i=1}^{n} |y_i - \hat{y}_i|$ is minimized or for which $\sum_{i=1}^{n} |y_i - \hat{y}_i|^{1.5}$ is minimized. These are both viable and reasonable methods. Note that both of these, as well as the least squares procedure, result in forcing residuals to be "small" in some sense. One should remember that the residuals are the empirical counterpart to the ϵ values. Figure 11.6 illustrates a set of residuals. One should note that the fitted line has predicted values as points on the line and hence the residuals are vertical deviations from points to the line. As a result, the least squares procedure produces a line that **minimizes the sum of squares of vertical deviations** from the points to the line.

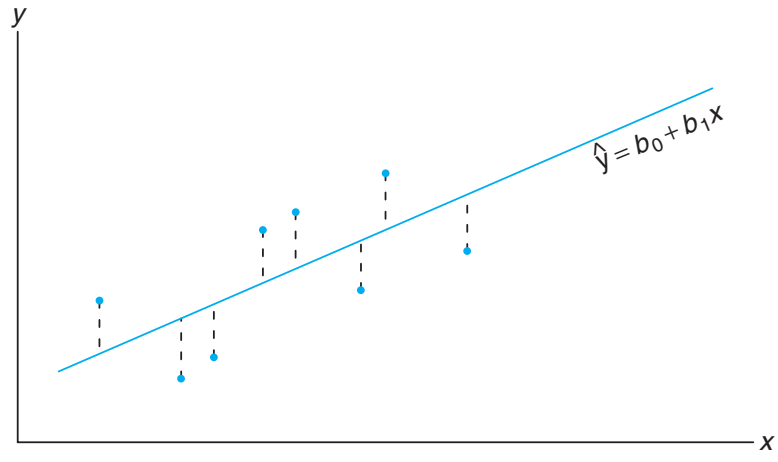


Figure 11.6: Residuals as vertical deviations.

Exercises

11.1 A study was conducted at Virginia Tech to determine if certain static arm-strength measures have an influence on the "dynamic lift" characteristics of an individual. Twenty-five individuals were subjected to strength tests and then were asked to perform a weight-lifting test in which weight was dynamically lifted overhead. The data are given here.

Individual	Arm Strength, x	Dynamic Lift, y
1	17.3	71.7
2	19.3	48.3
3	19.5	88.3
4	19.7	75.0
5	22.9	91.7
6	23.1	100.0
7	26.4	73.3
8	26.8	65.0
9	27.6	75.0
10	28.1	88.3
11	28.2	68.3
12	28.7	96.7
13	29.0	76.7
14	29.6	78.3
15	29.9	60.0
16	29.9	71.7
17	30.3	85.0
18	31.3	85.0
19	36.0	88.3
20	39.5	100.0
21	40.4	100.0
22	44.3	100.0
23	44.6	91.7
24	50.4	100.0
25	55.9	71.7

(a) Estimate β_0 and β_1 for the linear regression curve $\mu_{Y|x} = \beta_0 + \beta_1 x$.

(b) Find a point estimate of $\mu_{Y|30}$.

(c) Plot the residuals versus the x's (arm strength). Comment.

11.2 The grades of a class of 9 students on a midterm report (x) and on the final examination (y) are as follows:

x	77	50	71	72	81	94	96	99	67
y	82	66	78	34	47	85	99	99	68

(a) Estimate the linear regression line.

(b) Estimate the final examination grade of a student who received a grade of 85 on the midterm report.

11.3 The amounts of a chemical compound y that dissolved in 100 grams of water at various temperatures x were recorded as follows:

x (°C)	y (grams)		
0	8	6	8
15	12	10	14
30	25	21	24
45	31	33	28
60	44	39	42
75	48	51	44

(a) Find the equation of the regression line.

(b) Graph the line on a scatter diagram.

(c) Estimate the amount of chemical that will dissolve in 100 grams of water at 50°C.

11.4 The following data were collected to determine the relationship between pressure and the corresponding scale reading for the purpose of calibration.

Pressure, x (lb/sq in.)	Scale Reading, y
10	13
10	18
10	16
10	15
10	20
50	86
50	90
50	88
50	88
50	92

(a) Find the equation of the regression line.

(b) The purpose of calibration in this application is to estimate pressure from an observed scale reading. Estimate the pressure for a scale reading of 54 using $\hat{x} = (54 - b_0)/b_1$.

11.5 A study was made on the amount of converted sugar in a certain process at various temperatures. The data were coded and recorded as follows:

Temperature, x	Converted Sugar, y
1.0	8.1
1.1	7.8
1.2	8.5
1.3	9.8
1.4	9.5
1.5	8.9
1.6	8.6
1.7	10.2
1.8	9.3
1.9	9.2
2.0	10.5

(a) Estimate the linear regression line.

(b) Estimate the mean amount of converted sugar produced when the coded temperature is 1.75.

(c) Plot the residuals versus temperature. Comment.

11.6 In a certain type of metal test specimen, the normal stress on a specimen is known to be functionally related to the shear resistance. The following is a set of coded experimental data on the two variables:

Normal Stress, x	Shear Resistance, y
26.8	26.5
25.4	27.3
28.9	24.2
23.6	27.1
27.7	23.6
23.9	25.9
24.7	26.3
28.1	22.5
26.9	21.7
27.4	21.4
22.6	25.8
25.6	24.9

(a) Estimate the regression line $\mu_{Y|x} = \beta_0 + \beta_1 x$.

(b) Estimate the shear resistance for a normal stress of 24.5.

11.7 The following is a portion of a classic data set called the "pilot plot data" in *Fitting Equations to Data* by Daniel and Wood, published in 1971. The response y is the acid content of material produced by titration, whereas the regressor x is the organic acid content produced by extraction and weighing.

y	x	y	x
76	123	70	109
62	55	37	48
66	100	82	138
58	75	88	164
88	159	43	28

(a) Plot the data; does it appear that a simple linear regression will be a suitable model?

(b) Fit a simple linear regression; estimate a slope and intercept.

(c) Graph the regression line on the plot in (a).

11.8 A mathematics placement test is given to all entering freshmen at a small college. A student who receives a grade below 35 is denied admission to the regular mathematics course and placed in a remedial class. The placement test scores and the final grades for 20 students who took the regular course were recorded.

(a) Plot a scatter diagram.

(b) Find the equation of the regression line to predict course grades from placement test scores.

(c) Graph the line on the scatter diagram.

(d) If 60 is the minimum passing grade, below which placement test score should students in the future be denied admission to this course?

Placement Test	Course Grade
50	53
35	41
35	61
40	56
55	68
65	36
35	11
60	70
90	79
35	59
90	54
80	91
60	48
60	71
60	71
40	47
55	53
50	68
65	57
50	79

11.9 A study was made by a retail merchant to determine the relation between weekly advertising expenditures and sales.

Advertising Costs ($)	Sales ($)
40	385
20	400
25	395
20	365
30	475
50	440
40	490
20	420
50	560
40	525
25	480
50	510

(a) Plot a scatter diagram.

(b) Find the equation of the regression line to predict weekly sales from advertising expenditures.

(c) Estimate the weekly sales when advertising costs are $35.

(d) Plot the residuals versus advertising costs. Comment.

11.10 The following data are the selling prices z of a certain make and model of used car w years old. Fit a curve of the form $\mu_{z|w} = \gamma\delta^w$ by means of the nonlinear sample regression equation $\hat{z} = cd^w$. [*Hint*: Write $\ln \hat{z} = \ln c + (\ln d)w = b_0 + b_1 w$.]

w (years)	z (dollars)	w (years)	z (dollars)
1	6350	3	5395
2	5695	5	4985
2	5750	5	4895

11.11 The thrust of an engine (y) is a function of exhaust temperature (x) in °F when other important variables are held constant. Consider the following data.

y	x	y	x
4300	1760	4010	1665
4650	1652	3810	1550
3200	1485	4500	1700
3150	1390	3008	1270
4950	1820		

(a) Plot the data.

(b) Fit a simple linear regression to the data and plot the line through the data.

11.12 A study was done to study the effect of ambient temperature x on the electric power consumed by a chemical plant y. Other factors were held constant, and the data were collected from an experimental pilot plant.

y (BTU)	x (°F)	y (BTU)	x (°F)
250	27	265	31
285	45	298	60
320	72	267	34
295	58	321	74

(a) Plot the data.

(b) Estimate the slope and intercept in a simple linear regression model.

(c) Predict power consumption for an ambient temperature of 65°F.

11.13 A study of the amount of rainfall and the quantity of air pollution removed produced the following data:

Daily Rainfall, x (0.01 cm)	Particulate Removed, y (μg/m^3)
4.3	126
4.5	121
5.9	116
5.6	118
6.1	114
5.2	118
3.8	132
2.1	141
7.5	108

(a) Find the equation of the regression line to predict the particulate removed from the amount of daily rainfall.

(b) Estimate the amount of particulate removed when the daily rainfall is $x = 4.8$ units.

11.14 A professor in the School of Business in a university polled a dozen colleagues about the number of professional meetings they attended in the past five years (x) and the number of papers they submitted to refereed journals (y) during the same period. The summary data are given as follows:

$$n = 12, \quad \bar{x} = 4, \quad \bar{y} = 12,$$
$$\sum_{i=1}^{n} x_i^2 = 232, \quad \sum_{i=1}^{n} x_i y_i = 318.$$

Fit a simple linear regression model between x and y by finding out the estimates of intercept and slope. Comment on whether attending more professional meetings would result in publishing more papers.

11.4 Properties of the Least Squares Estimators

In addition to the assumptions that the error term in the model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

is a random variable with mean 0 and constant variance σ^2, suppose that we make the further assumption that $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are independent from run to run in the experiment. This provides a foundation for finding the means and variances for the estimators of β_0 and β_1.

It is important to remember that our values of b_0 and b_1, based on a given sample of n observations, are only estimates of true parameters β_0 and β_1. If the experiment is repeated over and over again, each time using the same fixed values of x, the resulting estimates of β_0 and β_1 will most likely differ from experiment to experiment. These different estimates may be viewed as values assumed by the random variables B_0 and B_1, while b_0 and b_1 are specific realizations.

Since the values of x remain fixed, the values of B_0 and B_1 depend on the variations in the values of y or, more precisely, on the values of the random variables,

$Y_1, Y_2, \ldots, Y_n$. The distributional assumptions imply that the Y_i, $i = 1, 2, \ldots, n$, are also independently distributed, with mean $\mu_{Y|x_i} = \beta_0 + \beta_1 x_i$ and equal variances σ^2; that is,

$$\sigma^2_{Y|x_i} = \sigma^2 \quad \text{for} \quad i = 1, 2, \ldots, n.$$

Mean and Variance of Estimators

In what follows, we show that the estimator B_1 is unbiased for β_1 and demonstrate the variances of both B_0 and B_1. This will begin a series of developments that lead to hypothesis testing and confidence interval estimation on the intercept and slope.

Since the estimator

$$B_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

is of the form $\sum_{i=1}^{n} c_i Y_i$, where

$$c_i = \frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad i = 1, 2, \ldots, n,$$

we may conclude from Theorem 7.11 that B_1 has a $n(\mu_{B_1}, \sigma_{B_1})$ distribution with

$$\mu_{B_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \beta_1 \text{ and } \sigma^2_{B_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sigma^2_{Y_i}}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

It can also be shown (Review Exercise 11.60 on page 438) that the random variable B_0 is normally distributed with

$$\text{mean } \mu_{B_0} = \beta_0 \text{ and variance } \sigma^2_{B_0} = \frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}\sigma^2.$$

From the foregoing results, it is apparent that the **least squares estimators for β_0 and β_1 are both unbiased estimators**.

Partition of Total Variability and Estimation of σ^2

To draw inferences on β_0 and β_1, it becomes necessary to arrive at an estimate of the parameter σ^2 appearing in the two preceding variance formulas for B_0 and B_1. The parameter σ^2, the model error variance, reflects random variation or

experimental error variation around the regression line. In much of what follows, it is advantageous to use the notation

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2, \quad S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2, \quad S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

Now we may write the error sum of squares as follows:

$$\begin{aligned} SSE &= \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2 = \sum_{i=1}^{n}[(y_i - \bar{y}) - b_1(x_i - \bar{x})]^2 \\ &= \sum_{i=1}^{n}(y_i - \bar{y})^2 - 2b_1 \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) + b_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= S_{yy} - 2b_1 S_{xy} + b_1^2 S_{xx} = S_{yy} - b_1 S_{xy}, \end{aligned}$$

the final step following from the fact that $b_1 = S_{xy}/S_{xx}$.

Theorem 11.1: An unbiased estimate of σ^2 is

$$s^2 = \frac{SSE}{n-2} = \sum_{i=1}^{n}\frac{(y_i - \hat{y}_i)^2}{n-2} = \frac{S_{yy} - b_1 S_{xy}}{n-2}.$$

The proof of Theorem 11.1 is left as an exercise (see Review Exercise 11.59).

The Estimator of σ^2 as a Mean Squared Error

One should observe the result of Theorem 11.1 in order to gain some intuition about the estimator of σ^2. The parameter σ^2 measures variance or squared deviations between Y values and their mean given by $\mu_{Y|x}$ (i.e., squared deviations between Y and $\beta_0 + \beta_1 x$). Of course, $\beta_0 + \beta_1 x$ is estimated by $\hat{y} = b_0 + b_1 x$. Thus, it would make sense that the variance σ^2 is best depicted as a squared deviation of the typical observation y_i from the estimated mean, $\hat{y}_i$, which is the corresponding point on the fitted line. Thus, $(y_i - \hat{y}_i)^2$ values reveal the appropriate variance, much like the way $(y_i - \bar{y})^2$ values measure variance when one is sampling in a nonregression scenario. In other words, $\bar{y}$ estimates the mean in the latter simple situation, whereas $\hat{y}_i$ estimates the mean of y_i in a regression structure. Now, what about the divisor $n-2$? In future sections, we shall note that these are the degrees of freedom associated with the estimator s^2 of σ^2. Whereas in the standard normal i.i.d. scenario, one degree of freedom is subtracted from n in the denominator and a reasonable explanation is that one parameter is estimated, namely the mean μ by, say, $\bar{y}$, but in the regression problem, **two parameters are estimated**, namely β_0 and β_1 by b_0 and b_1. Thus, the important parameter σ^2, estimated by

$$s^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-2),$$

is called a **mean squared error**, depicting a type of mean (division by $n-2$) of the squared residuals.

11.5 Inferences Concerning the Regression Coefficients

Aside from merely estimating the linear relationship between x and Y for purposes of prediction, the experimenter may also be interested in drawing certain inferences about the slope and intercept. In order to allow for the testing of hypotheses and the construction of confidence intervals on β_0 and β_1, one must be willing to make the further assumption that each ϵ_i, $i = 1, 2, \ldots, n$, is normally distributed. This assumption implies that $Y_1, Y_2, \ldots, Y_n$ are also normally distributed, each with probability distribution $n(y_i; \beta_0 + \beta_1 x_i, \sigma)$.

From Section 11.4 we know that B_1 follows a normal distribution. It turns out that under the normality assumption, a result very much analogous to that given in Theorem 8.4 allows us to conclude that $(n-2)S^2/\sigma^2$ is a chi-squared variable with $n-2$ degrees of freedom, independent of the random variable B_1. Theorem 8.5 then assures us that the statistic

$$T = \frac{(B_1 - \beta_1)/(\sigma/\sqrt{S_{xx}})}{S/\sigma} = \frac{B_1 - \beta_1}{S/\sqrt{S_{xx}}}$$

has a t-distribution with $n-2$ degrees of freedom. The statistic T can be used to construct a $100(1-\alpha)\%$ confidence interval for the coefficient β_1.

Confidence Interval for β_1 A $100(1-\alpha)\%$ confidence interval for the parameter β_1 in the regression line $\mu_{Y|x} = \beta_0 + \beta_1 x$ is

$$b_1 - t_{\alpha/2}\frac{s}{\sqrt{S_{xx}}} < \beta_1 < b_1 + t_{\alpha/2}\frac{s}{\sqrt{S_{xx}}},$$

where $t_{\alpha/2}$ is a value of the t-distribution with $n-2$ degrees of freedom.

Example 11.2: Find a 95% confidence interval for β_1 in the regression line $\mu_{Y|x} = \beta_0 + \beta_1 x$, based on the pollution data of Table 11.1.

***Solution*:** From the results given in Example 11.1 we find that $S_{xx} = 4152.18$ and $S_{xy} = 3752.09$. In addition, we find that $S_{yy} = 3713.88$. Recall that $b_1 = 0.903643$. Hence,

$$s^2 = \frac{S_{yy} - b_1 S_{xy}}{n-2} = \frac{3713.88 - (0.903643)(3752.09)}{31} = 10.4299.$$

Therefore, taking the square root, we obtain $s = 3.2295$. Using Table A.4, we find $t_{0.025} \approx 2.045$ for 31 degrees of freedom. Therefore, a 95% confidence interval for β_1 is

$$0.903643 - \frac{(2.045)(3.2295)}{\sqrt{4152.18}} < \beta < 0.903643 + \frac{(2.045)(3.2295)}{\sqrt{4152.18}},$$

which simplifies to

$$0.8012 < \beta_1 < 1.0061.$$

Hypothesis Testing on the Slope

To test the null hypothesis H_0 that $\beta_1 = \beta_{10}$ against a suitable alternative, we again use the t-distribution with $n-2$ degrees of freedom to establish a critical region and then base our decision on the value of

$$t = \frac{b_1 - \beta_{10}}{s/\sqrt{S_{xx}}}.$$

The method is illustrated by the following example.

Example 11.3: Using the estimated value $b_1 = 0.903643$ of Example 11.1, test the hypothesis that $\beta_1 = 1.0$ against the alternative that $\beta_1 < 1.0$.

***Solution*:** The hypotheses are H_0: $\beta_1 = 1.0$ and H_1: $\beta_1 < 1.0$. So

$$t = \frac{0.903643 - 1.0}{3.2295/\sqrt{4152.18}} = -1.92,$$

with $n - 2 = 31$ degrees of freedom ($P \approx 0.03$).

Decision: The t-value is significant at the 0.03 level, suggesting strong evidence that $\beta_1 < 1.0$.

One important t-test on the slope is the test of the hypothesis

$$H_0\text{: } \beta_1 = 0 \text{ versus } H_1\text{: } \beta_1 \neq 0.$$

When the null hypothesis is not rejected, the conclusion is that there is no significant linear relationship between $E(y)$ and the independent variable x. The plot of the data for Example 11.1 would suggest that a linear relationship exists. However, in some applications in which σ^2 is large and thus considerable "noise" is present in the data, a plot, while useful, may not produce clear information for the researcher. Rejection of H_0 above implies that a significant linear regression exists.

Figure 11.7 displays a *MINITAB* printout showing the t-test for

$$H_0\text{: } \beta_1 = 0 \text{ versus } H_1\text{: } \beta_1 \neq 0,$$

for the data of Example 11.1. Note the regression coefficient (Coef), standard error (SE Coef), t-value (T), and P-value (P). The null hypothesis is rejected. Clearly, there is a significant linear relationship between mean chemical oxygen demand reduction and solids reduction. Note that the t-statistic is computed as

$$t = \frac{\text{coefficient}}{\text{standard error}} = \frac{b_1}{s/\sqrt{S_{xx}}}.$$

The failure to reject H_0: $\beta_1 = 0$ suggests that there is no linear relationship between Y and x. Figure 11.8 is an illustration of the implication of this result. It may mean that changing x has little impact on changes in Y, as seen in (a). However, it may also indicate that the true relationship is nonlinear, as indicated by (b).

When H_0: $\beta_1 = 0$ is rejected, there is an implication that the linear term in x residing in the model explains a significant portion of variability in Y. The two

```
Regression Analysis: COD versus Per_Red
The regression equation is COD = 3.83 + 0.904 Per_Red

Predictor      Coef  SE Coef      T      P
Constant      3.830    1.768   2.17  0.038
Per_Red     0.90364  0.05012  18.03  0.000

S = 3.22954   R-Sq = 91.3%   R-Sq(adj) = 91.0%
Analysis of Variance
Source           DF      SS      MS       F      P
Regression        1  3390.6  3390.6  325.08  0.000
Residual Error   31   323.3    10.4
Total            32  3713.9
```

Figure 11.7: *MINITAB* printout for t-test for data of Example 11.1.

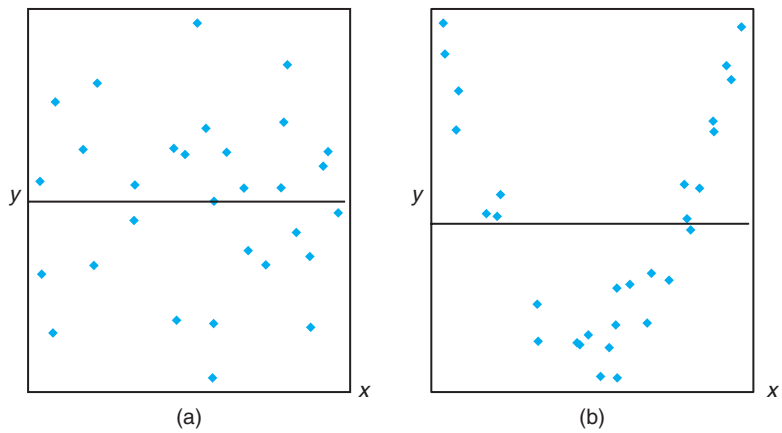


Figure 11.8: The hypothesis H_0: $\beta_1 = 0$ is not rejected.

plots in Figure 11.9 illustrate possible scenarios. As depicted in (a) of the figure, rejection of H_0 may suggest that the relationship is, indeed, linear. As indicated in (b), it may suggest that while the model does contain a linear effect, a better representation may be found by including a polynomial (perhaps quadratic) term (i.e., terms that supplement the linear term).

Statistical Inference on the Intercept

Confidence intervals and hypothesis testing on the coefficient β_0 may be established from the fact that B_0 is also normally distributed. It is not difficult to show that

$$T = \frac{B_0 - \beta_0}{S\sqrt{\sum_{i=1}^{n} x_i^2/(nS_{xx})}}$$

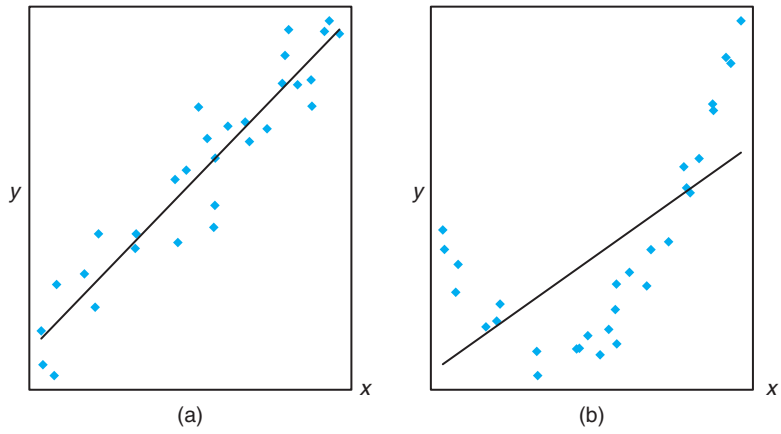


Figure 11.9: The hypothesis H_0: $\beta_1 = 0$ is rejected.

has a t-distribution with $n-2$ degrees of freedom from which we may construct a $100(1-\alpha)\%$ confidence interval for α.

Confidence Interval for β_0

A $100(1-\alpha)\%$ confidence interval for the parameter β_0 in the regression line $\mu_{Y|x} = \beta_0 + \beta_1 x$ is

$$b_0 - t_{\alpha/2}\frac{s}{\sqrt{nS_{xx}}}\sqrt{\sum_{i=1}^{n} x_i^2} < \beta_0 < b_0 + t_{\alpha/2}\frac{s}{\sqrt{nS_{xx}}}\sqrt{\sum_{i=1}^{n} x_i^2},$$

where $t_{\alpha/2}$ is a value of the t-distribution with $n-2$ degrees of freedom.

Example 11.4: Find a 95% confidence interval for β_0 in the regression line $\mu_{Y|x} = \beta_0 + \beta_1 x$, based on the data of Table 11.1.

***Solution*:** In Examples 11.1 and 11.2, we found that

$$S_{xx} = 4152.18 \quad \text{and} \quad s = 3.2295.$$

From Example 11.1 we had

$$\sum_{i=1}^{n} x_i^2 = 41{,}086 \quad \text{and} \quad b_0 = 3.829633.$$

Using Table A.4, we find $t_{0.025} \approx 2.045$ for 31 degrees of freedom. Therefore, a 95% confidence interval for β_0 is

$$3.829633 - \frac{(2.045)(3.2295)\sqrt{41{,}086}}{\sqrt{(33)(4152.18)}} < \beta_0 < 3.829633 + \frac{(2.045)(3.2295)\sqrt{41{,}086}}{\sqrt{(33)(4152.18)}},$$

which simplifies to $0.2132 < \beta_0 < 7.4461$.

To test the null hypothesis H_0 that $\beta_0 = \beta_{00}$ against a suitable alternative, we can use the t-distribution with $n-2$ degrees of freedom to establish a critical region and then base our decision on the value of

$$t = \frac{b_0 - \beta_{00}}{s\sqrt{\sum_{i=1}^{n} x_i^2/(nS_{xx})}}.$$

Example 11.5: Using the estimated value $b_0 = 3.829633$ of Example 11.1, test the hypothesis that $\beta_0 = 0$ at the 0.05 level of significance against the alternative that $\beta_0 \neq 0$.

***Solution*:** The hypotheses are H_0: $\beta_0 = 0$ and H_1: $\beta_0 \neq 0$. So

$$t = \frac{3.829633 - 0}{3.2295\sqrt{41{,}086/[(33)(4152.18)]}} = 2.17,$$

with 31 degrees of freedom. Thus, $P = P$-value ≈ 0.038 and we conclude that $\beta_0 \neq 0$. Note that this is merely Coef/StDev, as we see in the *MINITAB* printout in Figure 11.7. The SE Coef is the standard error of the estimated intercept. ∎

A Measure of Quality of Fit: Coefficient of Determination

Note in Figure 11.7 that an item denoted by R-Sq is given with a value of 91.3%. This quantity, R^2, is called the **coefficient of determination**. This quantity is a measure of the **proportion of variability explained by the fitted model**. In Section 11.8, we shall introduce the notion of an analysis-of-variance approach to hypothesis testing in regression. The analysis-of-variance approach makes use of the error sum of squares $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ and the **total corrected sum of squares** $SST = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2$. The latter represents the variation in the response values that *ideally* would be explained by the model. The SSE value is the variation due to error, or **variation unexplained**. Clearly, if $SSE = 0$, all variation is explained. The quantity that represents variation explained is $SST - SSE$. The R^2 is

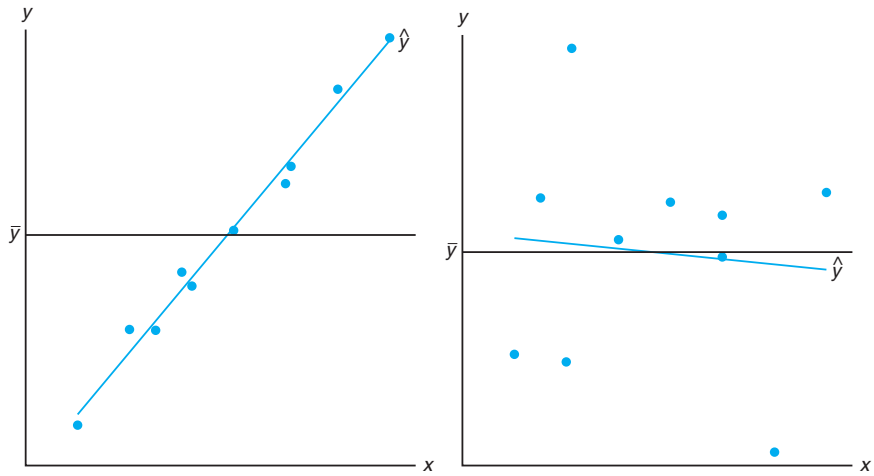
$$\text{Coeff. of determination:} \quad R^2 = 1 - \frac{SSE}{SST}.$$

Note that if the fit is perfect, *all residuals are zero*, and thus $R^2 = 1.0$. But if SSE is only slightly smaller than SST, $R^2 \approx 0$. Note from the printout in Figure 11.7 that the coefficient of determination suggests that the model fit to the data explains 91.3% of the variability observed in the response, the reduction in chemical oxygen demand.

Figure 11.10 provides an illustration of a good fit ($R^2 \approx 1.0$) in plot (a) and a poor fit ($R^2 \approx 0$) in plot (b).

Pitfalls in the Use of R^2

Analysts quote values of R^2 quite often, perhaps due to its simplicity. However, there are pitfalls in its interpretation. The reliability of R^2 is a function of the



(a) $R^2 \approx 1.0$ (b) $R^2 \approx 0$

Figure 11.10: Plots depicting a very good fit and a poor fit.

size of the regression data set and the type of application. Clearly, $0 \leq R^2 \leq 1$ and the upper bound is achieved when the fit to the data is perfect (i.e., all of the residuals are zero). What is an acceptable value for R^2? This is a difficult question to answer. A chemist, charged with doing a linear calibration of a high-precision piece of equipment, certainly expects to experience a very high R^2-value (perhaps exceeding 0.99), while a behavioral scientist, dealing in data impacted by variability in human behavior, may feel fortunate to experience an R^2 as large as 0.70. An experienced model fitter senses when a value is large enough, given the situation confronted. Clearly, some scientific phenomena lend themselves to modeling with more precision than others.

The R^2 criterion is dangerous to use for comparing *competing models* for the same data set. Adding additional terms to the model (e.g., an additional regressor) decreases SSE and thus increases R^2 (or at least does not decrease it). This implies that R^2 can be made artificially high by an unwise practice of **overfitting** (i.e., the inclusion of too many model terms). Thus, the inevitable increase in R^2 enjoyed by adding an additional term does not imply the additional term was needed. In fact, the simple model may be superior for predicting response values. The role of overfitting and its influence on prediction capability will be discussed at length in Chapter 12 as we visit the notion of models involving **more than a single regressor**. Suffice it to say at this point that one *should not subscribe to a model selection process that solely involves the consideration of* R^2.

11.6 Prediction

There are several reasons for building a linear regression. One, of course, is to predict response values at one or more values of the independent variable. In this

section, the focus is on errors associated with prediction.

The equation $\hat{y} = b_0 + b_1 x$ may be used to predict or estimate the **mean response** $\mu_{Y|x_0}$ at $x = x_0$, where x_0 is not necessarily one of the prechosen values, or it may be used to predict a single value y_0 of the variable Y_0, when $x = x_0$. We would expect the error of prediction to be higher in the case of a single predicted value than in the case where a mean is predicted. This, then, will affect the width of our intervals for the values being predicted.

Suppose that the experimenter wishes to construct a confidence interval for $\mu_{Y|x_0}$. We shall use the point estimator $\hat{Y}_0 = B_0 + B_1 x_0$ to estimate $\mu_{Y|x_0} = \beta_0 + \beta_1 x$. It can be shown that the sampling distribution of $\hat{Y}_0$ is normal with mean

$$\mu_{Y|x_0} = E(\hat{Y}_0) = E(B_0 + B_1 x_0) = \beta_0 + \beta_1 x_0 = \mu_{Y|x_0}$$

and variance

$$\sigma^2_{\hat{Y}_0} = \sigma^2_{B_0+B_1 x_0} = \sigma^2_{\bar{Y}+B_1(x_0-\bar{x})} = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right],$$

the latter following from the fact that $\text{Cov}(\bar{Y}, B_1) = 0$ (see Review Exercise 11.61 on page 438). Thus, a $100(1-\alpha)\%$ confidence interval on the mean response $\mu_{Y|x_0}$ can now be constructed from the statistic

$$T = \frac{\hat{Y}_0 - \mu_{Y|x_0}}{S\sqrt{1/n + (x_0 - \bar{x})^2/S_{xx}}},$$

which has a t-distribution with $n - 2$ degrees of freedom.

Confidence Interval for $\mu_{Y|x_0}$

A $100(1 - \alpha)\%$ confidence interval for the mean response $\mu_{Y|x_0}$ is

$$\hat{y}_0 - t_{\alpha/2} s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} < \mu_{Y|x_0} < \hat{y}_0 + t_{\alpha/2} s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}},$$

where $t_{\alpha/2}$ is a value of the t-distribution with $n - 2$ degrees of freedom.

Example 11.6: Using the data of Table 11.1, construct 95% confidence limits for the mean response $\mu_{Y|x_0}$.

***Solution*:** From the regression equation we find for $x_0 = 20\%$ solids reduction, say,

$$\hat{y}_0 = 3.829633 + (0.903643)(20) = 21.9025.$$

In addition, $\bar{x} = 33.4545$, $S_{xx} = 4152.18$, $s = 3.2295$, and $t_{0.025} \approx 2.045$ for 31 degrees of freedom. Therefore, a 95% confidence interval for $\mu_{Y|20}$ is

$$21.9025 - (2.045)(3.2295)\sqrt{\frac{1}{33} + \frac{(20 - 33.4545)^2}{4152.18}} < \mu_{Y|20}$$
$$< 21.9025 + (2.045)(3.2295)\sqrt{\frac{1}{33} + \frac{(20 - 33.4545)^2}{4152.18}},$$

or simply $20.1071 < \mu_{Y|20} < 23.6979$.

Repeating the previous calculations for each of several different values of x_0, one can obtain the corresponding confidence limits on each $\mu_{Y|x_0}$. Figure 11.11 displays the data points, the estimated regression line, and the upper and lower confidence limits on the mean of $Y|x$.

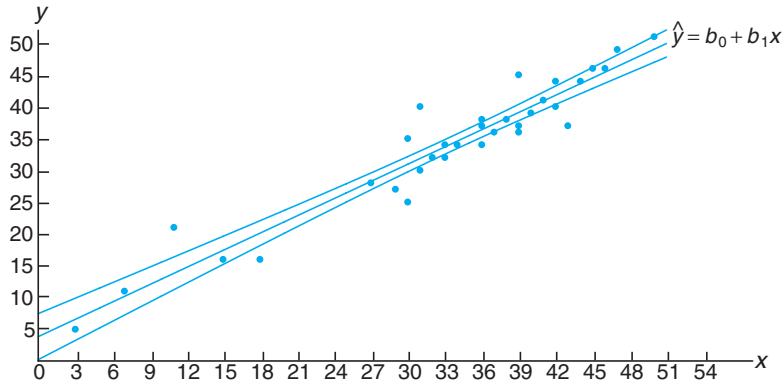


Figure 11.11: Confidence limits for the mean value of $Y|x$.

In Example 11.6, we are 95% confident that the population mean reduction in chemical oxygen demand is between 20.1071% and 23.6979% when solid reduction is 20%.

Prediction Interval

Another type of interval that is often misinterpreted and confused with that given for $\mu_{Y|x}$ is the prediction interval for a future observed response. Actually in many instances, the prediction interval is more relevant to the scientist or engineer than the confidence interval on the mean. In the tar content and inlet temperature example cited in Section 11.1, there would certainly be interest not only in estimating the mean tar content at a specific temperature but also in constructing an interval that reflects the error in predicting a future observed amount of tar content at the given temperature.

To obtain a **prediction interval** for any single value y_0 of the variable Y_0, it is necessary to estimate the variance of the differences between the ordinates $\hat{y}_0$, obtained from the computed regression lines in repeated sampling when $x = x_0$, and the corresponding true ordinate y_0. We can think of the difference $\hat{y}_0 - y_0$ as a value of the random variable $\hat{Y}_0 - Y_0$, whose sampling distribution can be shown to be normal with mean

$$\mu_{\hat{Y}_0 - Y_0} = E(\hat{Y}_0 - Y_0) = E[B_0 + B_1 x_0 - (\beta_0 + \beta_1 x_0 + \epsilon_0)] = 0$$

and variance

$$\sigma^2_{\hat{Y}_0 - Y_0} = \sigma^2_{B_0 + B_1 x_0 - \epsilon_0} = \sigma^2_{\bar{Y} + B_1(x_0 - \bar{x}) - \epsilon_0} = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right].$$

Thus, a $100(1-\alpha)\%$ prediction interval for a single predicted value y_0 can be constructed from the statistic

$$T = \frac{\hat{Y}_0 - Y_0}{S\sqrt{1 + 1/n + (x_0 - \bar{x})^2/S_{xx}}},$$

which has a t-distribution with $n-2$ degrees of freedom.

Prediction Interval for y_0

A $100(1-\alpha)\%$ prediction interval for a single response y_0 is given by

$$\hat{y}_0 - t_{\alpha/2}s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} < y_0 < \hat{y}_0 + t_{\alpha/2}s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}},$$

where $t_{\alpha/2}$ is a value of the t-distribution with $n-2$ degrees of freedom.

Clearly, there is a distinction between the concept of a confidence interval and the prediction interval described previously. The interpretation of the confidence interval is identical to that described for all confidence intervals on population parameters discussed throughout the book. Indeed, $\mu_{Y|x_0}$ is a population parameter. The computed prediction interval, however, represents an interval that has a probability equal to $1-\alpha$ of containing not a parameter but a future value y_0 of the random variable Y_0.

Example 11.7: Using the data of Table 11.1, construct a 95% prediction interval for y_0 when $x_0 = 20\%$.

***Solution*:** We have $n = 33$, $x_0 = 20$, $\bar{x} = 33.4545$, $\hat{y}_0 = 21.9025$, $S_{xx} = 4152.18$, $s = 3.2295$, and $t_{0.025} \approx 2.045$ for 31 degrees of freedom. Therefore, a 95% prediction interval for y_0 is

$$21.9025 - (2.045)(3.2295)\sqrt{1 + \frac{1}{33} + \frac{(20 - 33.4545)^2}{4152.18}} < y_0$$
$$< 21.9025 + (2.045)(3.2295)\sqrt{1 + \frac{1}{33} + \frac{(20 - 33.4545)^2}{4152.18}},$$

which simplifies to $15.0585 < y_0 < 28.7464$.

Figure 11.12 shows another plot of the chemical oxygen demand reduction data, with both the confidence interval on the mean response and the prediction interval on an individual response plotted. The plot reflects a much tighter interval around the regression line in the case of the mean response.

Exercises

11.15 With reference to Exercise 11.1 on page 398,
(a) evaluate s^2;
(b) test the hypothesis that $\beta_1 = 0$ against the alternative that $\beta_1 \neq 0$ at the 0.05 level of significance and interpret the resulting decision.

11.16 With reference to Exercise 11.2 on page 398,
(a) evaluate s^2;
(b) construct a 95% confidence interval for β_0;
(c) construct a 95% confidence interval for β_1.

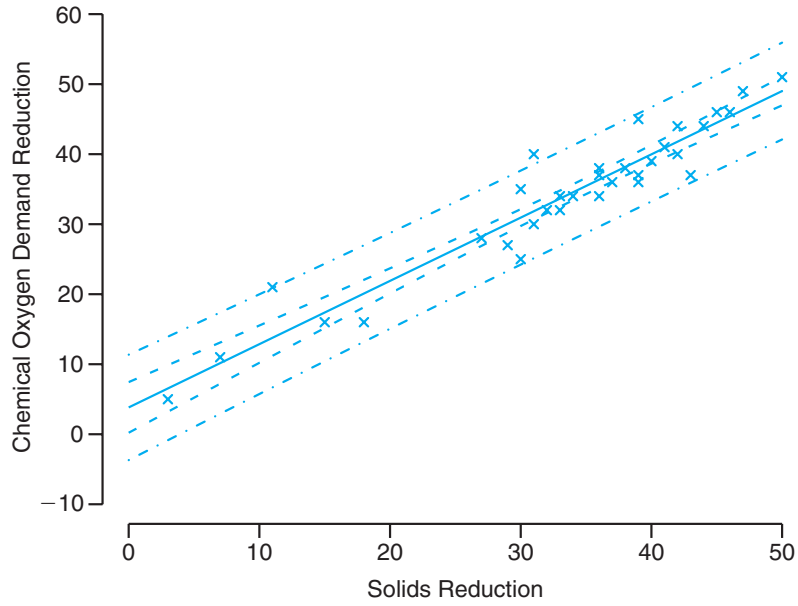


Figure 11.12: Confidence and prediction intervals for the chemical oxygen demand reduction data; inside bands indicate the confidence limits for the mean responses and outside bands indicate the prediction limits for the future responses.

11.17 With reference to Exercise 11.5 on page 398,
(a) evaluate s^2;
(b) construct a 95% confidence interval for β_0;
(c) construct a 95% confidence interval for β_1.

11.18 With reference to Exercise 11.6 on page 399,
(a) evaluate s^2;
(b) construct a 99% confidence interval for β_0;
(c) construct a 99% confidence interval for β_1.

11.19 With reference to Exercise 11.3 on page 398,
(a) evaluate s^2;
(b) construct a 99% confidence interval for β_0;
(c) construct a 99% confidence interval for β_1.

11.20 Test the hypothesis that $\beta_0 = 10$ in Exercise 11.8 on page 399 against the alternative that $\beta_0 < 10$. Use a 0.05 level of significance.

11.21 Test the hypothesis that $\beta_1 = 6$ in Exercise 11.9 on page 399 against the alternative that $\beta_1 < 6$. Use a 0.025 level of significance.

11.22 Using the value of s^2 found in Exercise 11.16(a), construct a 95% confidence interval for $\mu_{Y|85}$ in Exercise 11.2 on page 398.

11.23 With reference to Exercise 11.6 on page 399, use the value of s^2 found in Exercise 11.18(a) to compute
(a) a 95% confidence interval for the mean shear resistance when $x = 24.5$;
(b) a 95% prediction interval for a single predicted value of the shear resistance when $x = 24.5$.

11.24 Using the value of s^2 found in Exercise 11.17(a), graph the regression line and the 95% confidence bands for the mean response $\mu_{Y|x}$ for the data of Exercise 11.5 on page 398.

11.25 Using the value of s^2 found in Exercise 11.17(a), construct a 95% confidence interval for the amount of converted sugar corresponding to $x = 1.6$ in Exercise 11.5 on page 398.

11.26 With reference to Exercise 11.3 on page 398, use the value of s^2 found in Exercise 11.19(a) to compute
(a) a 99% confidence interval for the average amount

of chemical that will dissolve in 100 grams of water at 50°C;

(b) a 99% prediction interval for the amount of chemical that will dissolve in 100 grams of water at 50°C.

11.27 Consider the regression of mileage for certain automobiles, measured in miles per gallon (mpg) on their weight in pounds (wt). The data are from *Consumer Reports* (April 1997). Part of the *SAS* output from the procedure is shown in Figure 11.13.

(a) Estimate the mileage for a vehicle weighing 4000 pounds.

(b) Suppose that Honda engineers claim that, on average, the Civic (or any other model weighing 2440 pounds) gets more than 30 mpg. Based on the results of the regression analysis, would you believe that claim? Why or why not?

(c) The design engineers for the Lexus ES300 targeted 18 mpg as being ideal for this model (or any other model weighing 3390 pounds), although it is expected that some variation will be experienced. Is it likely that this target value is realistic? Discuss.

11.28 There are important applications in which, due to known scientific constraints, the regression line **must go through the origin** (i.e., the intercept must be zero). In other words, the model should read

$$Y_i = \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

and only a simple parameter requires estimation. The model is often called the **regression through the origin model**.

(a) Show that the least squares estimator of the slope is

$$b_1 = \left(\sum_{i=1}^{n} x_i y_i\right) \Big/ \left(\sum_{i=1}^{n} x_i^2\right).$$

(b) Show that $\sigma^2_{B_1} = \sigma^2 \Big/ \left(\sum_{i=1}^{n} x_i^2\right)$.

(c) Show that b_1 in part (a) is an unbiased estimator for β_1. That is, show $E(B_1) = \beta_1$.

11.29 Use the data set

y	x
7	2
50	15
100	30
40	10
70	20

(a) Plot the data.

(b) Fit a regression line through the origin.

(c) Plot the regression line on the graph with the data.

(d) Give a general formula (in terms of the y_i and the slope b_1) for the estimator of σ^2.

(e) Give a formula for $\text{Var}(\hat{y}_i)$, $i = 1, 2, \ldots, n$, for this case.

(f) Plot 95% confidence limits for the mean response on the graph around the regression line.

11.30 For the data in Exercise 11.29, find a 95% prediction interval at $x = 25$.

```
                Root MSE              1.48794    R-Square     0.9509
                Dependent Mean       21.50000    Adj R-Sq     0.9447

                              Parameter Estimates
                            Parameter       Standard
Variable      DF             Estimate          Error    t Value    Pr > |t|
Intercept      1             44.78018        1.92919      23.21      <.0001
WT             1             -0.00686     0.00055133     -12.44      <.0001

MODEL        WT     MPG    Predict     LMean      UMean      Lpred      Upred     Residual
GMC        4520      15    13.7720   11.9752    15.5688     9.8988    17.6451     1.22804
Geo        2065      29    30.6138   28.6063    32.6213    26.6385    34.5891    -1.61381
Honda      2440      31    28.0412   26.4143    29.6681    24.2439    31.8386     2.95877
Hyundai    2290      28    29.0703   27.2967    30.8438    25.2078    32.9327    -1.07026
Infiniti   3195      23    22.8618   21.7478    23.9758    19.2543    26.4693     0.13825
Isuzu      3480      21    20.9066   19.8160    21.9972    17.3062    24.5069     0.09341
Jeep       4090      15    16.7219   15.3213    18.1224    13.0158    20.4279    -1.72185
Land       4535      13    13.6691   11.8570    15.4811     9.7888    17.5493    -0.66905
Lexus      3390      22    21.5240   20.4390    22.6091    17.9253    25.1227     0.47599
Lincoln    3930      18    17.8195   16.5379    19.1011    14.1568    21.4822     0.18051
```

Figure 11.13: *SAS* printout for Exercise 11.27.

11.7 Choice of a Regression Model

Much of what has been presented thus far on regression involving a single independent variable depends on the assumption that the model chosen is correct, the presumption that $\mu_{Y|x}$ is related to x linearly in the parameters. Certainly, one cannot expect the prediction of the response to be good if there are several independent variables, not considered in the model, that are affecting the response and are varying in the system. In addition, the prediction will certainly be inadequate if the true structure relating $\mu_{Y|x}$ to x is extremely nonlinear in the range of the variables considered.

Often the simple linear regression model is used even though it is known that the model is something other than linear or that the true structure is unknown. This approach is often sound, particularly when the range of x is narrow. Thus, the model used becomes an approximating function that one hopes is an adequate representation of the true picture in the region of interest. One should note, however, the effect of an inadequate model on the results presented thus far. For example, if the true model, unknown to the experimenter, is linear in more than one x, say

$$\mu_{Y|x_1,x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

then the ordinary least squares estimate $b_1 = S_{xy}/S_{xx}$, calculated by only considering x_1 in the experiment, is, under general circumstances, a biased estimate of the coefficient β_1, the bias being a function of the additional coefficient β_2 (see Review Exercise 11.65 on page 438). Also, the estimate s^2 for σ^2 is biased due to the additional variable.

11.8 Analysis-of-Variance Approach

Often the problem of analyzing the quality of the estimated regression line is handled by an **analysis-of-variance** (ANOVA) approach: a procedure whereby the total variation in the dependent variable is subdivided into meaningful components that are then observed and treated in a systematic fashion. The analysis of variance, discussed in Chapter 13, is a powerful resource that is used for many applications.

Suppose that we have n experimental data points in the usual form (x_i, y_i) and that the regression line is estimated. In our estimation of σ^2 in Section 11.4, we established the identity

$$S_{yy} = b_1 S_{xy} + SSE.$$

An alternative and perhaps more informative formulation is

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

We have achieved a partitioning of the **total corrected sum of squares of y** into two components that should reflect particular meaning to the experimenter. We shall indicate this partitioning symbolically as

$$SST = SSR + SSE.$$

The first component on the right, SSR, is called the **regression sum of squares**, and it reflects the amount of variation in the y-values **explained by the model**, in this case the postulated straight line. The second component is the familiar error sum of squares, which reflects variation about the regression line.

Suppose that we are interested in testing the hypothesis

$$H_0\text{: } \beta_1 = 0 \text{ versus } H_1\text{: } \beta_1 \neq 0,$$

where the null hypothesis says essentially that the model is $\mu_{Y|x} = \beta_0$. That is, the variation in Y results from chance or random fluctuations which are independent of the values of x. This condition is reflected in Figure 11.10(b). Under the conditions of this null hypothesis, it can be shown that SSR/σ^2 and SSE/σ^2 are values of independent chi-squared variables with 1 and $n-2$ degrees of freedom, respectively, and then by Theorem 7.12 it follows that SST/σ^2 is also a value of a chi-squared variable with $n-1$ degrees of freedom. To test the hypothesis above, we compute

$$f = \frac{SSR/1}{SSE/(n-2)} = \frac{SSR}{s^2}$$

and reject H_0 at the α-level of significance when $f > f_\alpha(1, n-2)$.

The computations are usually summarized by means of an **analysis-of-variance table**, as in Table 11.2. It is customary to refer to the various sums of squares divided by their respective degrees of freedom as the **mean squares**.

Table 11.2: Analysis of Variance for Testing $\beta_1 = 0$

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	**Computed f**
Regression	SSR	1	SSR	$\frac{SSR}{s^2}$
Error	SSE	$n-2$	$s^2 = \frac{SSE}{n-2}$	
Total	SST	$n-1$		

When the null hypothesis is rejected, that is, when the computed F-statistic exceeds the critical value $f_\alpha(1, n-2)$, we conclude that **there is a significant amount of variation in the response accounted for by the postulated model, the straight-line function**. If the F-statistic is in the fail to reject region, we conclude that the data did not reflect sufficient evidence to support the model postulated.

In Section 11.5, a procedure was given whereby the statistic

$$T = \frac{B_1 - \beta_{10}}{S/\sqrt{S_{xx}}}$$

is used to test the hypothesis

$$H_0\text{: } \beta_1 = \beta_{10} \text{ versus } H_1\text{: } \beta_1 \neq \beta_{10},$$

where T follows the t-distribution with $n-2$ degrees of freedom. The hypothesis is rejected if $|t| > t_{\alpha/2}$ for an α-level of significance. It is interesting to note that

in the special case in which we are testing

$$H_0: \ \beta_1 = 0 \text{ versus } H_1: \ \beta_1 \neq 0,$$

the value of our T-statistic becomes

$$t = \frac{b_1}{s/\sqrt{S_{xx}}},$$

and the hypothesis under consideration is identical to that being tested in Table 11.2. Namely, the null hypothesis states that the variation in the response is due merely to chance. The analysis of variance uses the F-distribution rather than the t-distribution. For the two-sided alternative, the two approaches are identical. This we can see by writing

$$t^2 = \frac{b_1^2 S_{xx}}{s^2} = \frac{b_1 S_{xy}}{s^2} = \frac{SSR}{s^2},$$

which is identical to the f-value used in the analysis of variance. The basic relationship between the t-distribution with v degrees of freedom and the F-distribution with 1 and v degrees of freedom is

$$t^2 = f(1, v).$$

Of course, the t-test allows for testing against a one-sided alternative while the F-test is restricted to testing against a two-sided alternative.

Annotated Computer Printout for Simple Linear Regression

Consider again the chemical oxygen demand reduction data of Table 11.1. Figures 11.14 and 11.15 show more complete annotated computer printouts. Again we illustrate it with *MINITAB* software. The t-ratio column indicates tests for null hypotheses of zero values on the parameter. The term "Fit" denotes $\hat{y}$-values, often called **fitted values**. The term "SE Fit" is used in computing confidence intervals on mean response. The item R^2 is computed as $(SSR/SST) \times 100$ and signifies the proportion of variation in y explained by the straight-line regression. Also shown are confidence intervals on the mean response and prediction intervals on a new observation.

11.9 Test for Linearity of Regression: Data with Repeated Observations

In certain kinds of experimental situations, the researcher has the capability of obtaining repeated observations on the response for each value of x. Although it is not necessary to have these repetitions in order to estimate β_0 and β_1, nevertheless repetitions enable the experimenter to obtain quantitative information concerning the appropriateness of the model. In fact, if repeated observations are generated, the experimenter can make a significance test to aid in determining whether or not the model is adequate.

```
 The regression equation is COD = 3.83 + 0.904 Per_Red
 Predictor     Coef  SE Coef       T      P
  Constant   3.830    1.768    2.17  0.038
   Per_Red 0.90364  0.05012   18.03  0.000
S = 3.22954   R-Sq = 91.3%   R-Sq(adj) = 91.0%
             Analysis of Variance
 Source           DF      SS      MS       F    P
 Regression        1  3390.6  3390.6  325.08  0.000
 Residual Error 31   323.3    10.4
 Total            32  3713.9

Obs  Per_Red     COD     Fit  SE Fit  Residual  St Resid
  1      3.0   5.000   6.541   1.627    -1.541     -0.55
  2     36.0  34.000  36.361   0.576    -2.361     -0.74
  3      7.0  11.000  10.155   1.440     0.845      0.29
  4     37.0  36.000  37.264   0.590    -1.264     -0.40
  5     11.0  21.000  13.770   1.258     7.230      2.43
  6     38.0  38.000  38.168   0.607    -0.168     -0.05
  7     15.0  16.000  17.384   1.082    -1.384     -0.45
  8     39.0  37.000  39.072   0.627    -2.072     -0.65
  9     18.0  16.000  20.095   0.957    -4.095     -1.33
 10     39.0  36.000  39.072   0.627    -3.072     -0.97
 11     27.0  28.000  28.228   0.649    -0.228     -0.07
 12     39.0  45.000  39.072   0.627     5.928      1.87
 13     29.0  27.000  30.035   0.605    -3.035     -0.96
 14     40.0  39.000  39.975   0.651    -0.975     -0.31
 15     30.0  25.000  30.939   0.588    -5.939     -1.87
 16     41.0  41.000  40.879   0.678     0.121      0.04
 17     30.0  35.000  30.939   0.588     4.061      1.28
 18     42.0  40.000  41.783   0.707    -1.783     -0.57
 19     31.0  30.000  31.843   0.575    -1.843     -0.58
 20     42.0  44.000  41.783   0.707     2.217      0.70
 21     31.0  40.000  31.843   0.575     8.157      2.57
 22     43.0  37.000  42.686   0.738    -5.686     -1.81
 23     32.0  32.000  32.746   0.567    -0.746     -0.23
 24     44.0  44.000  43.590   0.772     0.410      0.13
 25     33.0  34.000  33.650   0.563     0.350      0.11
 26     45.0  46.000  44.494   0.807     1.506      0.48
 27     33.0  32.000  33.650   0.563    -1.650     -0.52
 28     46.0  46.000  45.397   0.843     0.603      0.19
 29     34.0  34.000  34.554   0.563    -0.554     -0.17
 30     47.0  49.000  46.301   0.881     2.699      0.87
 31     36.0  37.000  36.361   0.576     0.639      0.20
 32     50.0  51.000  49.012   1.002     1.988      0.65
 33     36.0  38.000  36.361   0.576     1.639      0.52
```

Figure 11.14: *MINITAB* printout of simple linear regression for chemical oxygen demand reduction data; part I.

Let us select a random sample of n observations using k distinct values of x, say $x_1, x_2, \ldots, x_n$, such that the sample contains n_1 observed values of the random variable Y_1 corresponding to x_1, n_2 observed values of Y_2 corresponding to $x_2, \ldots,$ n_k observed values of Y_k corresponding to x_k. Of necessity, $n = \sum_{i=1}^{k} n_i$.

Obs	Fit	SE Fit	95% CI	95% PI
1	6.541	1.627	(3.223, 9.858)	(-0.834, 13.916)
2	36.361	0.576	(35.185, 37.537)	(29.670, 43.052)
3	10.155	1.440	(7.218, 13.092)	(2.943, 17.367)
4	37.264	0.590	(36.062, 38.467)	(30.569, 43.960)
5	13.770	1.258	(11.204, 16.335)	(6.701, 20.838)
6	38.168	0.607	(36.931, 39.405)	(31.466, 44.870)
7	17.384	1.082	(15.177, 19.592)	(10.438, 24.331)
8	39.072	0.627	(37.793, 40.351)	(32.362, 45.781)
9	20.095	0.957	(18.143, 22.047)	(13.225, 26.965)
10	39.072	0.627	(37.793, 40.351)	(32.362, 45.781)
11	28.228	0.649	(26.905, 29.551)	(21.510, 34.946)
12	39.072	0.627	(37.793, 40.351)	(32.362, 45.781)
13	30.035	0.605	(28.802, 31.269)	(23.334, 36.737)
14	39.975	0.651	(38.648, 41.303)	(33.256, 46.694)
15	30.939	0.588	(29.739, 32.139)	(24.244, 37.634)
16	40.879	0.678	(39.497, 42.261)	(34.149, 47.609)
17	30.939	0.588	(29.739, 32.139)	(24.244, 37.634)
18	41.783	0.707	(40.341, 43.224)	(35.040, 48.525)
19	31.843	0.575	(30.669, 33.016)	(25.152, 38.533)
20	41.783	0.707	(40.341, 43.224)	(35.040, 48.525)
21	31.843	0.575	(30.669, 33.016)	(25.152, 38.533)
22	42.686	0.738	(41.181, 44.192)	(35.930, 49.443)
23	32.746	0.567	(31.590, 33.902)	(26.059, 39.434)
24	43.590	0.772	(42.016, 45.164)	(36.818, 50.362)
25	33.650	0.563	(32.502, 34.797)	(26.964, 40.336)
26	44.494	0.807	(42.848, 46.139)	(37.704, 51.283)
27	33.650	0.563	(32.502, 34.797)	(26.964, 40.336)
28	45.397	0.843	(43.677, 47.117)	(38.590, 52.205)
29	34.554	0.563	(33.406, 35.701)	(27.868, 41.239)
30	46.301	0.881	(44.503, 48.099)	(39.473, 53.128)
31	36.361	0.576	(35.185, 37.537)	(29.670, 43.052)
32	49.012	1.002	(46.969, 51.055)	(42.115, 55.908)
33	36.361	0.576	(35.185, 37.537)	(29.670, 43.052)

Figure 11.15: *MINITAB* printout of simple linear regression for chemical oxygen demand reduction data; part II.

We define

$$\begin{aligned} y_{ij} &= \text{the } j\text{th value of the random variable } Y_i, \\ y_{i.} &= T_{i.} = \sum_{j=1}^{n_i} y_{ij}, \\ \bar{y}_{i.} &= \frac{T_{i.}}{n_i}. \end{aligned}$$

Hence, if $n_4 = 3$ measurements of Y were made corresponding to $x = x_4$, we would indicate these observations by y_{41}, y_{42}, and y_{43}. Then

$$T_{i.} = y_{41} + y_{42} + y_{43}.$$

Concept of Lack of Fit

The error sum of squares consists of two parts: the amount due to the variation between the values of Y within given values of x and a component that is normally

called the **lack-of-fit** contribution. The first component reflects mere random variation, or **pure experimental error**, while the second component is a measure of the systematic variation brought about by higher-order terms. In our case, these are terms in x other than the linear, or first-order, contribution. Note that in choosing a linear model we are essentially assuming that this second component does not exist and hence our error sum of squares is completely due to random errors. If this should be the case, then $s^2 = SSE/(n-2)$ is an unbiased estimate of σ^2. However, if the model does not adequately fit the data, then the error sum of squares is inflated and produces a biased estimate of σ^2. Whether or not the model fits the data, an unbiased estimate of σ^2 can always be obtained when we have repeated observations simply by computing

$$s_i^2 = \frac{\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2}{n_i - 1}, \quad i = 1, 2, \ldots, k,$$

for each of the k distinct values of x and then pooling these variances to get

$$s^2 = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{n-k} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2}{n-k}.$$

The numerator of s^2 is a **measure of the pure experimental error**. A computational procedure for separating the error sum of squares into the two components representing pure error and lack of fit is as follows:

Computation of Lack-of-Fit Sum of Squares

1. Compute the pure error sum of squares

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2.$$

This sum of squares has $n-k$ degrees of freedom associated with it, and the resulting mean square is our unbiased estimate s^2 of σ^2.

2. Subtract the pure error sum of squares from the error sum of squares SSE, thereby obtaining the sum of squares due to lack of fit. The degrees of freedom for lack of fit are obtained by simply subtracting $(n-2) - (n-k) = k-2$.

The computations required for testing hypotheses in a regression problem with repeated measurements on the response may be summarized as shown in Table 11.3.

Figures 11.16 and 11.17 display the sample points for the "correct model" and "incorrect model" situations. In Figure 11.16, where the $\mu_{Y|x}$ fall on a straight line, there is no lack of fit when a linear model is assumed, so the sample variation around the regression line is a pure error resulting from the variation that occurs among repeated observations. In Figure 11.17, where the $\mu_{Y|x}$ clearly do not fall on a straight line, the lack of fit from erroneously choosing a linear model accounts for a large portion of the variation around the regression line, supplementing the pure error.

Table 11.3: Analysis of Variance for Testing Linearity of Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Computed f
Regression	SSR	1	SSR	$\frac{SSR}{s^2}$
Error	SSE	$n-2$		
Lack of fit	$SSE - SSE$ (pure)	$k-2$	$\frac{SSE-SSE(\text{pure})}{k-2}$	$\frac{SSE-SSE(\text{pure})}{s^2(k-2)}$
Pure error	SSE (pure)	$n-k$	$s^2 = \frac{SSE(\text{pure})}{n-k}$	
Total	SST	$n-1$		

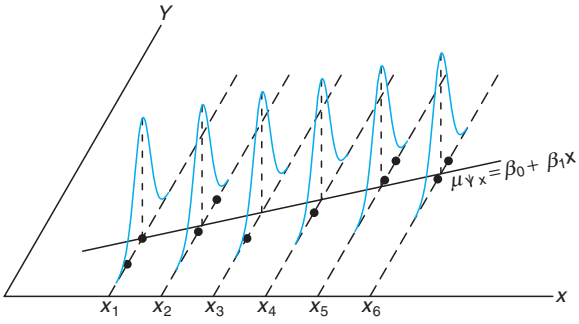


Figure 11.16: Correct linear model with no lack-of-fit component.

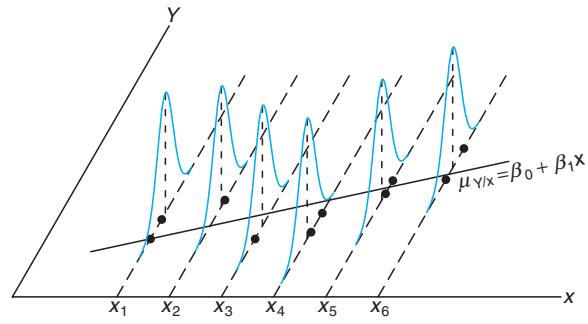


Figure 11.17: Incorrect linear model with lack-of-fit component.

What Is the Importance in Detecting Lack of Fit?

The concept of lack of fit is extremely important in applications of regression analysis. In fact, the need to construct or design an experiment that will account for lack of fit becomes more critical as the problem and the underlying mechanism involved become more complicated. Surely, one cannot always be certain that his or her postulated structure, in this case the linear regression model, is correct or even an adequate representation. The following example shows how the error sum of squares is partitioned into the two components representing pure error and lack of fit. The adequacy of the model is tested at the α-level of significance by comparing the lack-of-fit mean square divided by s^2 with $f_\alpha(k-2, n-k)$.

Example 11.8: Observations of the yield of a chemical reaction taken at various temperatures were recorded in Table 11.4. Estimate the linear model $\mu_{Y|x} = \beta_0 + \beta_1 x$ and test for lack of fit.

***Solution*:** Results of the computations are shown in Table 11.5.

Conclusion: The partitioning of the total variation in this manner reveals a significant variation accounted for by the linear model and an insignificant amount of variation due to lack of fit. Thus, the experimental data do not seem to suggest the need to consider terms higher than first order in the model, and the null hypothesis is not rejected. ∎

Table 11.4: Data for Example 11.8

y (%)	x (°C)	y (%)	x (°C)
77.4	150	88.9	250
76.7	150	89.2	250
78.2	150	89.7	250
84.1	200	94.8	300
84.5	200	94.7	300
83.7	200	95.9	300

Table 11.5: Analysis of Variance on Yield-Temperature Data

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Computed f	P-Values
Regression	509.2507	1	509.2507	1531.58	<0.0001
Error	3.8660	10			
Lack of fit	1.2060	2	0.6030	1.81	0.2241
Pure error	2.6600	8	0.3325		
Total	513.1167	11			

Annotated Computer Printout for Test for Lack of Fit

Figure 11.18 is an annotated computer printout showing analysis of the data of Example 11.8 with *SAS*. Note the "LOF" with 2 degrees of freedom, representing the quadratic and cubic contribution to the model, and the P-value of 0.22, suggesting that the linear (first-order) model is adequate.

```
Dependent Variable: yield
                                      Sum of
Source                     DF         Squares     Mean Square    F Value    Pr > F
Model                       3     510.4566667     170.1522222     511.74    <.0001
Error                       8       2.6600000       0.3325000
Corrected Total            11     513.1166667
              R-Square     Coeff Var      Root MSE    yield Mean
              0.994816      0.666751      0.576628      86.48333
Source                     DF       Type I SS     Mean Square    F Value    Pr > F
temperature                 1     509.2506667     509.2506667    1531.58    <.0001
LOF                         2       1.2060000       0.6030000       1.81    0.2241
```

Figure 11.18: *SAS* printout, showing analysis of data of Example 11.8.

Exercises

11.31 Test for linearity of regression in Exercise 11.3 on page 398. Use a 0.05 level of significance. Comment.

11.32 Test for linearity of regression in Exercise 11.8 on page 399. Comment.

11.33 Suppose we have a linear equation through the origin (Exercise 11.28) $\mu_{Y|x} = \beta x$.

(a) Estimate the regression line passing through the origin for the following data:

x	0.5	1.5	3.2	4.2	5.1	6.5
y	1.3	3.4	6.7	8.0	10.0	13.2

(b) Suppose it is not known whether the true regression should pass through the origin. Estimate the linear model $\mu_{Y|x} = \beta_0 + \beta_1 x$ and test the hypothesis that $\beta_0 = 0$, at the 0.10 level of significance, against the alternative that $\beta_0 \neq 0$.

11.34 Use an analysis-of-variance approach to test the hypothesis that $\beta_1 = 0$ against the alternative hypothesis $\beta_1 \neq 0$ in Exercise 11.5 on page 398 at the 0.05 level of significance.

11.35 The following data are a result of an investigation as to the effect of reaction temperature x on percent conversion of a chemical process y. (See Myers, Montgomery and Anderson-Cook, 2009.) Fit a simple linear regression, and use a lack-of-fit test to determine if the model is adequate. Discuss.

Observation	Temperature (°C), x	Conversion (%), y
1	200	43
2	250	78
3	200	69
4	250	73
5	189.65	48
6	260.35	78
7	225	65
8	225	74
9	225	76
10	225	79
11	225	83
12	225	81

11.36 Transistor gain between emitter and collector in an integrated circuit device (hFE) is related to two variables (Myers, Montgomery and Anderson-Cook, 2009) that can be controlled at the deposition process, emitter drive-in time (x_1, in minutes) and emitter dose (x_2, in ions $\times 10^{14}$). Fourteen samples were observed following deposition, and the resulting data are shown in the table below. We will consider linear regression models using gain as the response and emitter drive-in time or emitter dose as the regressor variable.

Obs.	x_1 (drive-in time, min)	x_2 (dose, ions $\times 10^{14}$)	y (gain, or hFE)
1	195	4.00	1004
2	255	4.00	1636
3	195	4.60	852
4	255	4.60	1506
5	255	4.20	1272
6	255	4.10	1270
7	255	4.60	1269
8	195	4.30	903
9	255	4.30	1555
10	255	4.00	1260
11	255	4.70	1146
12	255	4.30	1276
13	255	4.72	1225
14	340	4.30	1321

(a) Determine if emitter drive-in time influences gain in a linear relationship. That is, test H_0: $\beta_1 = 0$, where β_1 is the slope of the regressor variable.

(b) Do a lack-of-fit test to determine if the linear relationship is adequate. Draw conclusions.

(c) Determine if emitter dose influences gain in a linear relationship. Which regressor variable is the better predictor of gain?

11.37 Organophosphate (OP) compounds are used as pesticides. However, it is important to study their effect on species that are exposed to them. In the laboratory study *Some Effects of Organophosphate Pesticides on Wildlife Species*, by the Department of Fisheries and Wildlife at Virginia Tech, an experiment was conducted in which different dosages of a particular OP pesticide were administered to 5 groups of 5 mice (*peromysius leucopus*). The 25 mice were females of similar age and condition. One group received no chemical. The basic response y was a measure of activity in the brain. It was postulated that brain activity would decrease with an increase in OP dosage. The data are as follows:

Animal	Dose, x (mg/kg body weight)	Activity, y (moles/liter/min)
1	0.0	10.9
2	0.0	10.6
3	0.0	10.8
4	0.0	9.8
5	0.0	9.0
6	2.3	11.0
7	2.3	11.3
8	2.3	9.9
9	2.3	9.2
10	2.3	10.1
11	4.6	10.6
12	4.6	10.4
13	4.6	8.8
14	4.6	11.1
15	4.6	8.4
16	9.2	9.7
17	9.2	7.8
18	9.2	9.0
19	9.2	8.2
20	9.2	2.3
21	18.4	2.9
22	18.4	2.2
23	18.4	3.4
24	18.4	5.4
25	18.4	8.2

(a) Using the model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \ldots, 25,$$

find the least squares estimates of β_0 and β_1.

(b) Construct an analysis-of-variance table in which the lack of fit and pure error have been separated.

Determine if the lack of fit is significant at the 0.05 level. Interpret the results.

11.38 Heat treating is often used to carburize metal parts such as gears. The thickness of the carburized layer is considered an important feature of the gear, and it contributes to the overall reliability of the part. Because of the critical nature of this feature, a lab test is performed on each furnace load. The test is a destructive one, where an actual part is cross sectioned and soaked in a chemical for a period of time. This test involves running a carbon analysis on the surface of both the gear pitch (top of the gear tooth) and the gear root (between the gear teeth). The data below are the results of the pitch carbon-analysis test for 19 parts.

Soak Time	Pitch	Soak Time	Pitch
0.58	0.013	1.17	0.021
0.66	0.016	1.17	0.019
0.66	0.015	1.17	0.021
0.66	0.016	1.20	0.025
0.66	0.015	2.00	0.025
0.66	0.016	2.00	0.026
1.00	0.014	2.20	0.024
1.17	0.021	2.20	0.025
1.17	0.018	2.20	0.024
1.17	0.019		

(a) Fit a simple linear regression relating the pitch carbon analysis y against soak time. Test H_0: $\beta_1 = 0$.

(b) If the hypothesis in part (a) is rejected, determine if the linear model is adequate.

11.39 A regression model is desired relating temperature and the proportion of impurities passing through solid helium. Temperature is listed in degrees centigrade. The data are as follows:

Temperature (°C)	Proportion of Impurities
−260.5	0.425
−255.7	0.224
−264.6	0.453
−265.0	0.475
−270.0	0.705
−272.0	0.860
−272.5	0.935
−272.6	0.961
−272.8	0.979
−272.9	0.990

(a) Fit a linear regression model.

(b) Does it appear that the proportion of impurities passing through helium increases as the temperature approaches −273 degrees centigrade?

(c) Find R^2.

(d) Based on the information above, does the linear model seem appropriate? What additional information would you need to better answer that question?

11.40 It is of interest to study the effect of population size in various cities in the United States on ozone concentrations. The data consist of the 1999 population in millions and the amount of ozone present per hour in ppb (parts per billion). The data are as follows.

Ozone (ppb/hour), y	Population, x
126	0.6
135	4.9
124	0.2
128	0.5
130	1.1
128	0.1
126	1.1
128	2.3
128	0.6
129	2.3

(a) Fit the linear regression model relating ozone concentration to population. Test H_0: $\beta_1 = 0$ using the ANOVA approach.

(b) Do a test for lack of fit. Is the linear model appropriate based on the results of your test?

(c) Test the hypothesis of part (a) using the pure mean square error in the F-test. Do the results change? Comment on the advantage of each test.

11.41 Evaluating nitrogen deposition from the atmosphere is a major role of the National Atmospheric Deposition Program (NADP), a partnership of many agencies. NADP is studying atmospheric deposition and its effect on agricultural crops, forest surface waters, and other resources. Nitrogen oxides may affect the ozone in the atmosphere and the amount of pure nitrogen in the air we breathe. The data are as follows:

Year	Nitrogen Oxide
1978	0.73
1979	2.55
1980	2.90
1981	3.83
1982	2.53
1983	2.77
1984	3.93
1985	2.03
1986	4.39
1987	3.04
1988	3.41
1989	5.07
1990	3.95
1991	3.14
1992	3.44
1993	3.63
1994	4.50
1995	3.95
1996	5.24
1997	3.30
1998	4.36
1999	3.33

(a) Plot the data.

(b) Fit a linear regression model and find R^2.

(c) What can you say about the trend in nitrogen oxide across time?

11.42 For a particular variety of plant, researchers wanted to develop a formula for predicting the quantity of seeds (in grams) as a function of the density of plants. They conducted a study with four levels of the factor x, the number of plants per plot. Four replications were used for each level of x. The data are shown as follows:

Plants per Plot, x	**Quantity of Seeds, y (grams)**			
10	12.6	11.0	12.1	10.9
20	15.3	16.1	14.9	15.6
30	17.9	18.3	18.6	17.8
40	19.2	19.6	18.9	20.0

Is a simple linear regression model adequate for analyzing this data set?

11.10 Data Plots and Transformations

In this chapter, we deal with building regression models where there is one independent, or regressor, variable. In addition, we are assuming, through model formulation, that both x and y enter the model in a *linear fashion*. Often it is advisable to work with an alternative model in which either x or y (or both) enters in a nonlinear way. A **transformation** of the data may be indicated because of theoretical considerations inherent in the scientific study, or a simple plotting of the data may suggest the need to *reexpress* the variables in the model. The need to perform a transformation is rather simple to diagnose in the case of simple linear regression because two-dimensional plots give a true pictorial display of how each variable enters the model.

A model in which x or y is transformed should not be viewed as a *nonlinear regression model*. We normally refer to a regression model as linear when it is **linear in the parameters**. In other words, suppose the complexion of the data or other scientific information suggests that we should **regress y^* against x^***, where each is a transformation on the natural variables x and y. Then the model of the form

$$y_i^* = \beta_0 + \beta_1 x_i^* + \epsilon_i$$

is a linear model since it is linear in the parameters β_0 and β_1. The material given in Sections 11.2 through 11.9 remains intact, with y_i^* and x_i^* replacing y_i and x_i. A simple and useful example is the log-log model

$$\log y_i = \beta_0 + \beta_1 \log x_i + \epsilon_i.$$

Although this model is not linear in x and y, it is linear in the parameters and is thus treated as a linear model. On the other hand, an example of a truly nonlinear model is

$$y_i = \beta_0 + \beta_1 x^{\beta_2} + \epsilon_i,$$

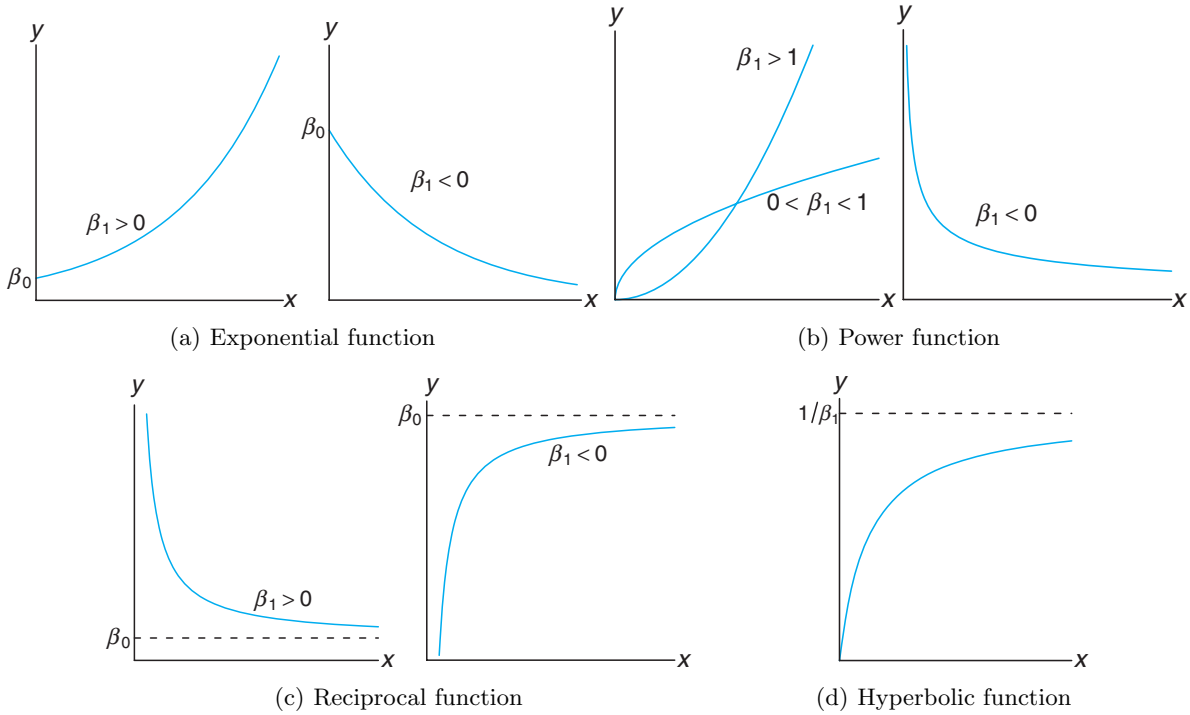
where the parameter β_2 (as well as β_0 and β_1) is to be estimated. The model is not linear in β_2.

Transformations that may enhance the fit and predictability of a model are many in number. For a thorough discussion of transformations, the reader is referred to Myers (1990, see the Bibliography). We choose here to indicate a few of them and show the appearance of the graphs that serve as a diagnostic tool. Consider Table 11.6. Several functions are given describing relationships between y and x that can produce a *linear regression* through the transformation indicated.

In addition, for the sake of completeness the reader is given the dependent and independent variables to use in the resulting *simple linear regression*. Figure 11.19 depicts functions listed in Table 11.6. These serve as a guide for the analyst in choosing a transformation from the observation of the plot of y against x.

Table 11.6: Some Useful Transformations to Linearize

Functional Form Relating y to x	**Proper Transformation**	**Form of Simple Linear Regression**
Exponential: $y = \beta_0 e^{\beta_1 x}$	$y^* = \ln y$	Regress y^* against x
Power: $y = \beta_0 x^{\beta_1}$	$y^* = \log y; \quad x^* = \log x$	Regress y^* against x^*
Reciprocal: $y = \beta_0 + \beta_1\left(\frac{1}{x}\right)$	$x^* = \frac{1}{x}$	Regress y against x^*
Hyperbolic: $y = \frac{x}{\beta_0 + \beta_1 x}$	$y^* = \frac{1}{y}; \quad x^* = \frac{1}{x}$	Regress y^* against x^*



(a) Exponential function

(b) Power function

(c) Reciprocal function

(d) Hyperbolic function

Figure 11.19: Diagrams depicting functions listed in Table 11.6.

What Are the Implications of a Transformed Model?

The foregoing is intended as an aid for the analyst when it is apparent that a transformation will provide an improvement. However, before we provide an example, two important points should be made. The first one revolves around the formal writing of the model when the data are transformed. Quite often the analyst does not think about this. He or she merely performs the transformation without any

concern about the model form *before* and *after* the transformation. The exponential model serves as a good illustration. The model in the natural (untransformed) variables that produces an *additive error model* in the transformed variables is given by

$$y_i = \beta_0 e^{\beta_1 x_i} \cdot \epsilon_i,$$

which is a *multiplicative error model*. Clearly, taking logs produces

$$\ln y_i = \ln \beta_0 + \beta_1 x_i + \ln \epsilon_i.$$

As a result, it is on $\ln \epsilon_i$ that the basic assumptions are made. The purpose of this presentation is merely to remind the reader that one should not view a transformation as merely an algebraic manipulation with an error added. Often a model in the transformed variables that has a proper *additive error structure* is a result of a model in the natural variables with a different type of error structure.

The second important point deals with the notion of measures of improvement. Obvious measures of comparison are, of course, R^2 and the residual mean square, s^2. (Other measures of performance used to compare competing models are given in Chapter 12.) Now, if the response y is not transformed, then clearly s^2 and R^2 can be used in measuring the utility of the transformation. The residuals will be in the same units for both the transformed and the untransformed models. But when y is transformed, performance criteria for the transformed model should be based on values of the residuals in the metric of the untransformed response so that comparisons that are made are proper. The example that follows provides an illustration.

Example 11.9: The pressure P of a gas corresponding to various volumes V is recorded, and the data are given in Table 11.7.

Table 11.7: Data for Example 11.9

V (cm^3)	50	60	70	90	100
P (kg/cm^2)	64.7	51.3	40.5	25.9	7.8

The ideal gas law is given by the functional form $PV^\gamma = C$, where γ and C are constants. Estimate the constants C and γ.

***Solution*:** Let us take natural logs of both sides of the model

$$P_i V^\gamma = C \cdot \epsilon_i, \quad i = 1, 2, 3, 4, 5.$$

As a result, a linear model can be written

$$\ln P_i = \ln C - \gamma \ln V_i + \epsilon_i^*, \quad i = 1, 2, 3, 4, 5,$$

where $\epsilon_i^* = \ln \epsilon_i$. The following represents results of the simple linear regression:

$$\text{Intercept: } \widehat{\ln C} = 14.7589, \ \widehat{C} = 2,568,862.88, \ \text{Slope: } \hat{\gamma} = 2.65347221.$$

The following represents information taken from the regression analysis.

$\boldsymbol{P_i}$	$\boldsymbol{V_i}$	$\boldsymbol{\ln P_i}$	$\boldsymbol{\ln V_i}$	$\boldsymbol{\widehat{\ln P_i}}$	$\boldsymbol{\widehat{P_i}}$	$\boldsymbol{e_i = P_i - \widehat{P_i}}$
64.7	50	4.16976	3.91202	4.37853	79.7	−15.0
51.3	60	3.93769	4.09434	3.89474	49.1	2.2
40.5	70	3.70130	4.24850	3.48571	32.6	7.9
25.9	90	3.25424	4.49981	2.81885	16.8	9.1
7.8	100	2.05412	4.60517	2.53921	12.7	−4.9

It is instructive to plot the data and the regression equation. Figure 11.20 shows a plot of the data in the untransformed pressure and volume and the curve representing the regression equation.

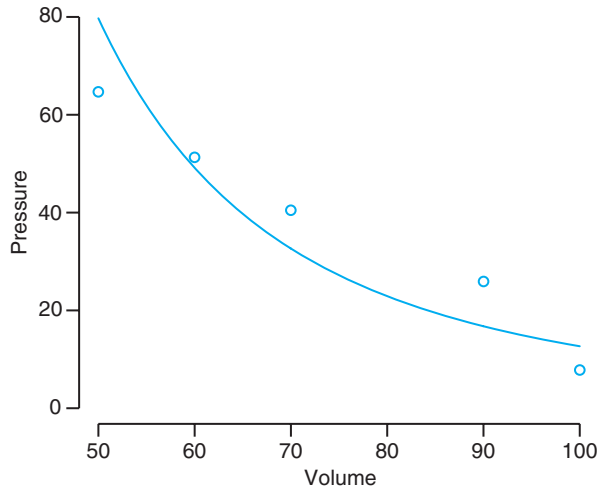


Figure 11.20: Pressure and volume data and fitted regression.

Diagnostic Plots of Residuals: Graphical Detection of Violation of Assumptions

Plots of the raw data can be extremely helpful in determining the nature of the model that should be fit to the data when there is a single independent variable. We have attempted to illustrate this in the foregoing. Detection of proper model form is, however, not the only benefit gained from diagnostic plotting. As in much of the material associated with significance testing in Chapter 10, plotting methods can illustrate and detect violation of assumptions. The reader should recall that much of what is illustrated in this chapter requires assumptions made on the model errors, the ϵ_i. In fact, we assume that the ϵ_i are independent $N(0, \sigma)$ random variables. Now, of course, the ϵ_i are not observed. However, the $e_i = y_i - \hat{y}_i$, the *residuals*, are the error in the fit of the regression line and thus serve to mimic the ϵ_i. Thus, the general complexion of these residuals can often highlight difficulties. Ideally, of course, the plot of the residuals is as depicted in Figure 11.21. That is, they should truly show random fluctuations around a value of zero.

Nonhomogeneous Variance

Homogeneous variance is an important assumption made in regression analysis. Violations can often be detected through the appearance of the residual plot. Increasing error variance with an increase in the regressor variable is a common condition in scientific data. Large error variance produces large residuals, and hence a residual plot like the one in Figure 11.22 is a signal of nonhomogeneous variance. More discussion regarding these residual plots and information regard-

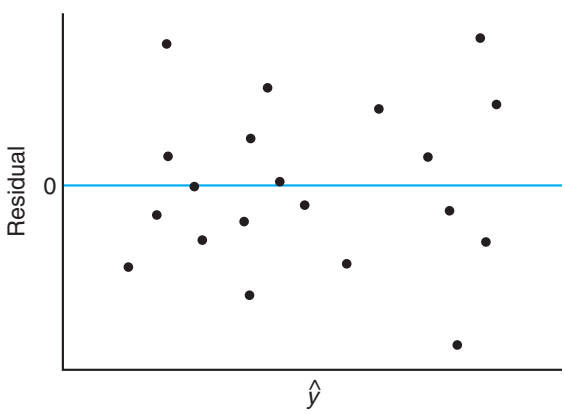


Figure 11.21: Ideal residual plot.

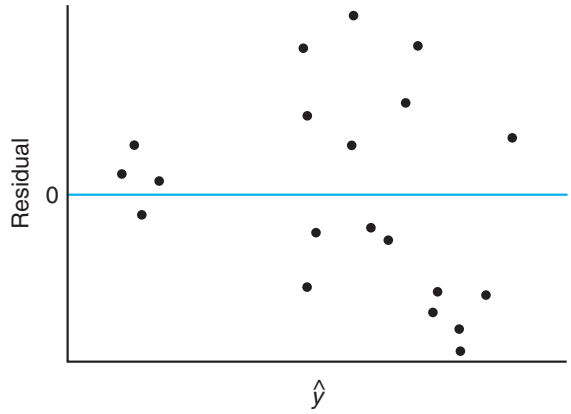


Figure 11.22: Residual plot depicting heterogeneous error variance.

ing different types of residuals appears in Chapter 12, where we deal with multiple linear regression.

Normal Probability Plotting

The assumption that the model errors are normal is made when the data analyst deals in either hypothesis testing or confidence interval estimation. Again, the numerical counterpart to the ϵ_i, namely the residuals, are subjects of diagnostic plotting to detect any extreme violations. In Chapter 8, we introduced normal quantile-quantile plots and briefly discussed normal probability plots. These plots on residuals are illustrated in the case study introduced in the next section.

11.11 Simple Linear Regression Case Study

In the manufacture of commercial wood products, it is important to estimate the relationship between the density of a wood product and its stiffness. A relatively new type of particleboard is being considered that can be formed with considerably more ease than the accepted commercial product. It is necessary to know at what density the stiffness is comparable to that of the well-known, well-documented commercial product. A study was done by Terrance E. Conners, *Investigation of Certain Mechanical Properties of a Wood-Foam Composite* (M.S. Thesis, Department of Forestry and Wildlife Management, University of Massachusetts). Thirty particleboards were produced at densities ranging from roughly 8 to 26 pounds per cubic foot, and the stiffness was measured in pounds per square inch. Table 11.8 shows the data.

It is necessary for the data analyst to focus on an appropriate fit to the data and use inferential methods discussed in this chapter. Hypothesis testing on the slope of the regression, as well as confidence or prediction interval estimation, may well be appropriate. We begin by demonstrating a simple scatter plot of the raw data with a simple linear regression superimposed. Figure 11.23 shows this plot.

The simple linear regression fit to the data produced the fitted model

$$\hat{y} = -25{,}433.739 + 3884.976x \quad (R^2 = 0.7975),$$

Table 11.8: Density and Stiffness for 30 Particleboards

Density, x	Stiffness, y	Density, x	Stiffness, y
9.50	14,814.00	8.40	17,502.00
9.80	14,007.00	11.00	19,443.00
8.30	7573.00	9.90	14,191.00
8.60	9714.00	6.40	8076.00
7.00	5304.00	8.20	10,728.00
17.40	43,243.00	15.00	25,319.00
15.20	28,028.00	16.40	41,792.00
16.70	49,499.00	15.40	25,312.00
15.00	26,222.00	14.50	22,148.00
14.80	26,751.00	13.60	18,036.00
25.60	96,305.00	23.40	104,170.00
24.40	72,594.00	23.30	49,512.00
19.50	32,207.00	21.20	48,218.00
22.80	70,453.00	21.70	47,661.00
19.80	38,138.00	21.30	53,045.00

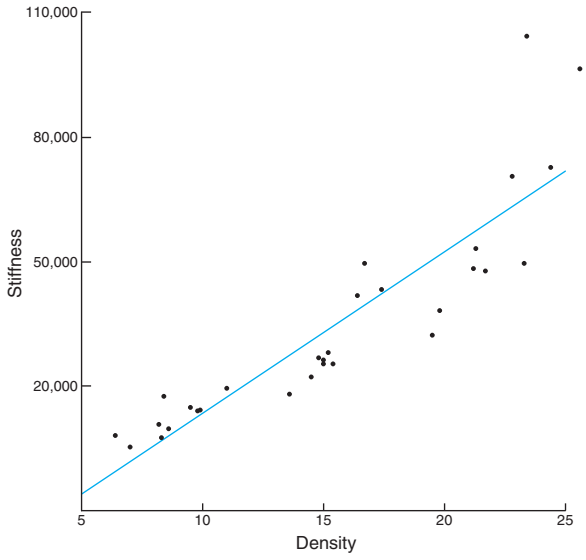


Figure 11.23: Scatter plot of the wood density data.

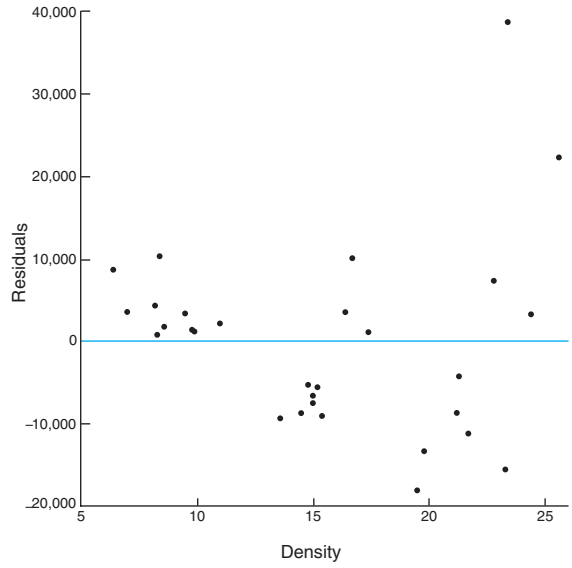


Figure 11.24: Residual plot for the wood density data.

and the residuals were computed. Figure 11.24 shows the residuals plotted against the measurements of density. This is hardly an ideal or healthy set of residuals. They do not show a random scatter around a value of zero. In fact, clusters of positive and negative values suggest that a curvilinear trend in the data should be investigated.

To gain some type of idea regarding the normal error assumption, a normal probability plot of the residuals was generated. This is the type of plot discussed in

Section 8.8 in which the horizontal axis represents the empirical normal distribution function on a scale that produces a straight-line plot when plotted against the residuals. Figure 11.25 shows the normal probability plot of the residuals. The normal probability plot does not reflect the straight-line appearance that one would like to see. This is another symptom of a faulty, perhaps overly simplistic choice of a regression model.

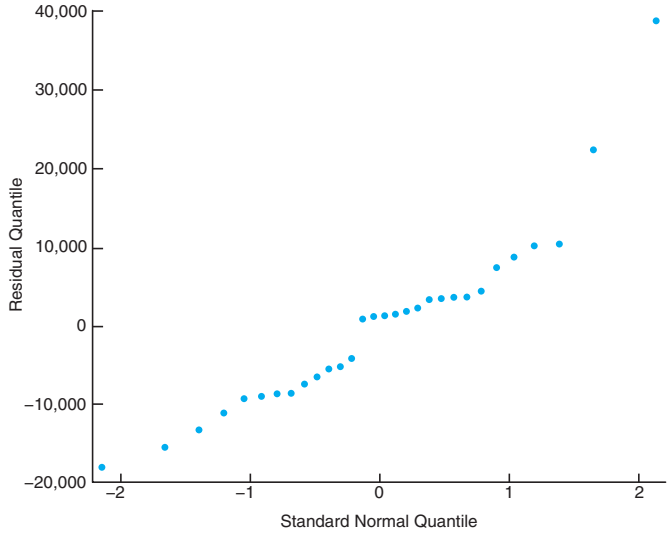


Figure 11.25: Normal probability plot of residuals for wood density data.

Both types of residual plots and, indeed, the scatter plot itself suggest here that a somewhat complicated model would be appropriate. One possible approach is to use a natural log transformation. In other words, one might choose to regress $\ln y$ against x. This produces the regression

$$\widehat{\ln y} = 8.257 + 0.125x \quad (R^2 = 0.9016).$$

To gain some insight into whether the transformed model is more appropriate, consider Figures 11.26 and 11.27, which reveal plots of the residuals in stiffness [i.e., y_i-antilog $(\widehat{\ln y})$] against density. Figure 11.26 appears to be closer to a random pattern around zero, while Figure 11.27 is certainly closer to a straight line. This in addition to the higher R^2-value would suggest that the transformed model is more appropriate.

11.12 Correlation

Up to this point we have assumed that the independent regressor variable x is a physical or scientific variable but not a random variable. In fact, in this context, x is often called a **mathematical variable**, which, in the sampling process, is measured with negligible error. In many applications of regression techniques, it is more realistic to assume that both X and Y are random variables and the measurements $\{(x_i, y_i); i = 1, 2, \ldots, n\}$ are observations from a population having

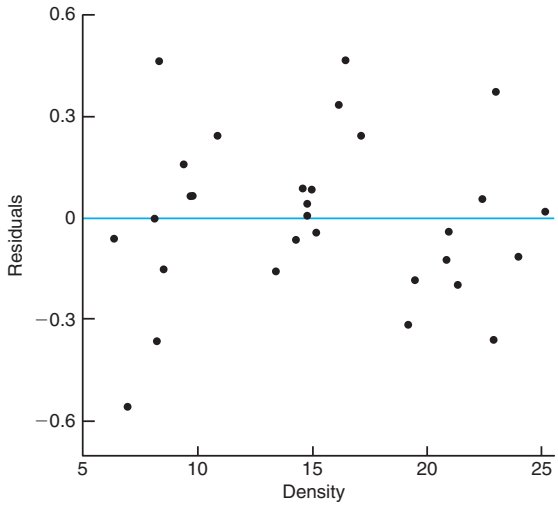


Figure 11.26: Residual plot using the log transformation for the wood density data.

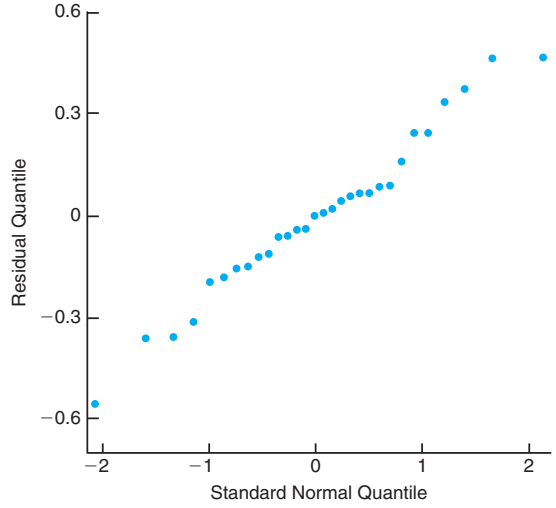


Figure 11.27: Normal probability plot of residuals using the log transformation for the wood density data.

the joint density function $f(x, y)$. We shall consider the problem of measuring the relationship between the two variables X and Y. For example, if X and Y represent the length and circumference of a particular kind of bone in the adult body, we might conduct an anthropological study to determine whether large values of X are associated with large values of Y, and vice versa.

On the other hand, if X represents the age of a used automobile and Y represents the retail book value of the automobile, we would expect large values of X to correspond to small values of Y and small values of X to correspond to large values of Y. **Correlation analysis** attempts to measure the strength of such relationships between two variables by means of a single number called a **correlation coefficient**.

In theory, it is often assumed that the conditional distribution $f(y|x)$ of Y, for fixed values of X, is normal with mean $\mu_{Y|x} = \beta_0 + \beta_1 x$ and variance $\sigma^2_{Y|x} = \sigma^2$ and that X is likewise normally distributed with mean μ and variance σ^2_x. The joint density of X and Y is then

$$\begin{aligned} f(x, y) = \; & n(y|x; \beta_0 + \beta_1 x, \sigma) n(x; \mu_X, \sigma_X) \\ = \; & \frac{1}{2\pi\sigma_x\sigma} \exp\left\{ -\frac{1}{2} \left[\left(\frac{y - \beta_0 - \beta_1 x}{\sigma} \right)^2 + \left(\frac{x - \mu_X}{\sigma_X} \right)^2 \right] \right\}, \end{aligned}$$

for $-\infty < x < \infty$ and $-\infty < y < \infty$.

Let us write the random variable Y in the form

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

where X is now a random variable independent of the random error ϵ. Since the mean of the random error ϵ is zero, it follows that

$$\mu_Y = \beta_0 + \beta_1 \mu_X \quad \text{and} \quad \sigma^2_Y = \sigma^2 + \beta_1^2 \sigma^2_X.$$

Substituting for α and σ^2 into the preceding expression for $f(x, y)$, we obtain the **bivariate normal distribution**

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\},$$

for $-\infty < x < \infty$ and $-\infty < y < \infty$, where

$$\rho^2 = 1 - \frac{\sigma^2}{\sigma_Y^2} = \beta_1^2\frac{\sigma_X^2}{\sigma_Y^2}.$$

The constant ρ (rho) is called the **population correlation coefficient** and plays a major role in many bivariate data analysis problems. It is important for the reader to understand the physical interpretation of this correlation coefficient and the distinction between correlation and regression. The term *regression* still has meaning here. In fact, the straight line given by $\mu_{Y|x} = \beta_0 + \beta_1 x$ is still called the regression line as before, and the estimates of β_0 and β_1 are identical to those given in Section 11.3. The value of ρ is 0 when $\beta_1 = 0$, which results when there essentially is no linear regression; that is, the regression line is horizontal and any knowledge of X is useless in predicting Y. Since $\sigma_Y^2 \geq \sigma^2$, we must have $\rho^2 \leq 1$ and hence $-1 \leq \rho \leq 1$. Values of $\rho = \pm 1$ only occur when $\sigma^2 = 0$, in which case we have a perfect linear relationship between the two variables. Thus, a value of ρ equal to $+1$ implies a perfect linear relationship with a positive slope, while a value of ρ equal to -1 results from a perfect linear relationship with a negative slope. It might be said, then, that sample estimates of ρ close to unity in magnitude imply good correlation, or **linear association**, between X and Y, whereas values near zero indicate little or no correlation.

To obtain a sample estimate of ρ, recall from Section 11.4 that the error sum of squares is

$$SSE = S_{yy} - b_1 S_{xy}.$$

Dividing both sides of this equation by S_{yy} and replacing S_{xy} by $b_1 S_{xx}$, we obtain the relation

$$b_1^2\frac{S_{xx}}{S_{yy}} = 1 - \frac{SSE}{S_{yy}}.$$

The value of $b_1^2 S_{xx}/S_{yy}$ is zero when $b_1 = 0$, which will occur when the sample points show no linear relationship. Since $S_{yy} \geq SSE$, we conclude that $b_1^2 S_{xx}/S_{xy}$ must be between 0 and l. Consequently, $b_1\sqrt{S_{xx}/S_{yy}}$ must range from -1 to $+1$, negative values corresponding to lines with negative slopes and positive values to lines with positive slopes. A value of -1 or $+1$ will occur when $SSE = 0$, but this is the case where all sample points lie in a straight line. Hence, a perfect linear relationship appears in the sample data when $b_1\sqrt{S_{xx}/S_{yy}} = \pm 1$. Clearly, the quantity $b_1\sqrt{S_{xx}/S_{yy}}$, which we shall henceforth designate as r, can be used as an estimate of the population correlation coefficient ρ. It is customary to refer to the estimate r as the **Pearson product-moment correlation coefficient** or simply the **sample correlation coefficient**.

Correlation Coefficient

The measure ρ of linear association between two variables X and Y is estimated by the **sample correlation coefficient** r, where

$$r = b_1\sqrt{\frac{S_{xx}}{S_{yy}}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}.$$

For values of r between -1 and $+1$ we must be careful in our interpretation. For example, values of r equal to 0.3 and 0.6 only mean that we have two positive correlations, one somewhat stronger than the other. It is wrong to conclude that $r = 0.6$ indicates a linear relationship twice as good as that indicated by the value $r = 0.3$. On the other hand, if we write

$$r^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \frac{SSR}{S_{yy}},$$

then r^2, which is usually referred to as the **sample coefficient of determination**, represents the proportion of the variation of S_{yy} explained by the regression of Y on x, namely SSR. That is, r^2 expresses the proportion of the total variation in the values of the variable Y that can be accounted for or explained by a linear relationship with the values of the random variable X. Thus, a correlation of 0.6 means that 0.36, or 36%, of the total variation of the values of Y in our sample is accounted for by a linear relationship with values of X.

Example 11.10: It is important that scientific researchers in the area of forest products be able to study correlation among the anatomy and mechanical properties of trees. For the study *Quantitative Anatomical Characteristics of Plantation Grown Loblolly Pine (Pinus Taeda L.) and Cottonwood (Populus deltoides Bart. Ex Marsh.) and Their Relationships to Mechanical Properties*, conducted by the Department of Forestry and Forest Products at Virginia Tech, 29 loblolly pines were randomly selected for investigation. Table 11.9 shows the resulting data on the specific gravity in grams/cm^3 and the modulus of rupture in kilopascals (kPa). Compute and interpret the sample correlation coefficient.

Table 11.9: Data on 29 Loblolly Pines for Example 11.10

Specific Gravity, x (g/cm^3)	**Modulus of Rupture, y (kPa)**	**Specific Gravity, x (g/cm^3)**	**Modulus of Rupture, y (kPa)**
0.414	29,186	0.581	85,156
0.383	29,266	0.557	69,571
0.399	26,215	0.550	84,160
0.402	30,162	0.531	73,466
0.442	38,867	0.550	78,610
0.422	37,831	0.556	67,657
0.466	44,576	0.523	74,017
0.500	46,097	0.602	87,291
0.514	59,698	0.569	86,836
0.530	67,705	0.544	82,540
0.569	66,088	0.557	81,699
0.558	78,486	0.530	82,096
0.577	89,869	0.547	75,657
0.572	77,369	0.585	80,490
0.548	67,095		

***Solution*:** From the data we find that

$$S_{xx} = 0.11273, \quad S_{yy} = 11{,}807{,}324{,}805, \quad S_{xy} = 34{,}422.27572.$$

Therefore,

$$r = \frac{34{,}422.27572}{\sqrt{(0.11273)(11{,}807{,}324{,}805)}} = 0.9435.$$

A correlation coefficient of 0.9435 indicates a good linear relationship between X and Y. Since $r^2 = 0.8902$, we can say that approximately 89% of the variation in the values of Y is accounted for by a linear relationship with X.

A test of the special hypothesis $\rho = 0$ versus an appropriate alternative is equivalent to testing $\beta_1 = 0$ for the simple linear regression model, and therefore the procedures of Section 11.8 using either the t-distribution with $n-2$ degrees of freedom or the F-distribution with 1 and $n-2$ degrees of freedom are applicable. However, if one wishes to avoid the analysis-of-variance procedure and compute only the sample correlation coefficient, it can be verified (see Review Exercise 11.66 on page 438) that the t-value

$$t = \frac{b_1}{s/\sqrt{S_{xx}}}$$

can also be written as

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},$$

which, as before, is a value of the statistic T having a t-distribution with $n-2$ degrees of freedom.

Example 11.11: For the data of Example 11.10, test the hypothesis that there is no linear association among the variables.

Solution:

1. H_0: $\rho = 0$.
2. H_1: $\rho \neq 0$.
3. $\alpha = 0.05$.
4. Critical region: $t < -2.052$ or $t > 2.052$.
5. Computations: $t = \frac{0.9435\sqrt{27}}{\sqrt{1-0.9435^2}} = 14.79$, $P < 0.0001$.
6. Decision: Reject the hypothesis of no linear association.

A test of the more general hypothesis $\rho = \rho_0$ against a suitable alternative is easily conducted from the sample information. If X and Y follow the bivariate normal distribution, the quantity

$$\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

is the value of a random variable that follows approximately the normal distribution with mean $\frac{1}{2}\ln\frac{1+\rho}{1-\rho}$ and variance $1/(n-3)$. Thus, the test procedure is to compute

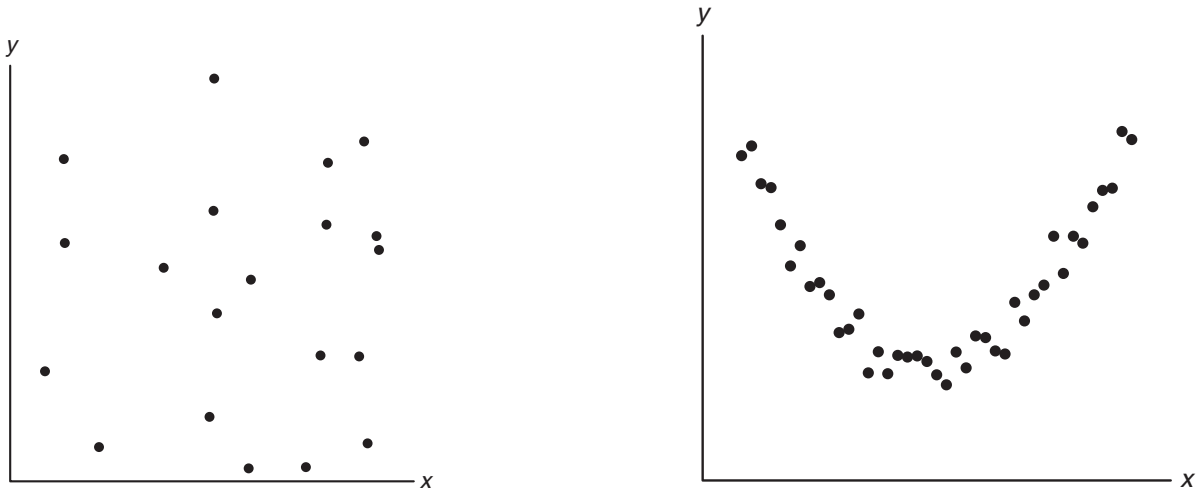
$$z = \frac{\sqrt{n-3}}{2}\left[\ln\left(\frac{1+r}{1-r}\right) - \ln\left(\frac{1+\rho_0}{1-\rho_0}\right)\right] = \frac{\sqrt{n-3}}{2}\ln\left[\frac{(1+r)(1-\rho_0)}{(1-r)(1+\rho_0)}\right]$$

and compare it with the critical points of the standard normal distribution.

Example 11.12: For the data of Example 11.10, test the null hypothesis that $\rho = 0.9$ against the alternative that $\rho > 0.9$. Use a 0.05 level of significance.

Solution:

1. H_0: $\rho = 0.9$.
2. H_1: $\rho > 0.9$.
3. $\alpha = 0.05$.
4. Critical region: $z > 1.645$.



(a) No Association (b) Causal Relationship

Figure 11.28: Scatter diagram showing zero correlation.

5. Computations:

$$z = \frac{\sqrt{26}}{2} \ln \left[\frac{(1 + 0.9435)(0.1)}{(1 - 0.9435)(1.9)} \right] = 1.51, \quad P = 0.0655.$$

6. Decision: There is certainly some evidence that the correlation coefficient does not exceed 0.9.

It should be pointed out that in correlation studies, as in linear regression problems, the results obtained are only as good as the model that is assumed. In the correlation techniques studied here, a bivariate normal density is assumed for the variables X and Y, with the mean value of Y at each x-value being linearly related to x. To observe the suitability of the linearity assumption, a preliminary plotting of the experimental data is often helpful. A value of the sample correlation coefficient close to zero will result from data that display a strictly random effect as in Figure 11.28(a), thus implying little or no causal relationship. It is important to remember that the correlation coefficient between two variables is a measure of their linear relationship and that a value of $r = 0$ implies *a lack of linearity and not a lack of association*. Hence, if a strong quadratic relationship exists between X and Y, as indicated in Figure 11.28(b), we can still obtain a zero correlation indicating a nonlinear relationship.

Exercises

11.43 Compute and interpret the correlation coefficient for the following grades of 6 students selected at random:

Mathematics grade	70	92	80	74	65	83
English grade	74	84	63	87	78	90

11.44 With reference to Exercise 11.1 on page 398, assume that x and y are random variables with a bivariate normal distribution.

(a) Calculate r.

(b) Test the hypothesis that $\rho = 0$ against the alternative that $\rho \neq 0$ at the 0.05 level of significance.

11.45 With reference to Exercise 11.13 on page 400, assume a bivariate normal distribution for x and y.

(a) Calculate r.

(b) Test the null hypothesis that $\rho = -0.5$ against the alternative that $\rho < -0.5$ at the 0.025 level of significance.

(c) Determine the percentage of the variation in the amount of particulate removed that is due to changes in the daily amount of rainfall.

11.46 Test the hypothesis that $\rho = 0$ in Exercise 11.43 against the alternative that $\rho \neq 0$. Use a 0.05 level of significance.

11.47 The following data were obtained in a study of the relationship between the weight and chest size of infants at birth.

Weight (kg)	Chest Size (cm)
2.75	29.5
2.15	26.3
4.41	32.2
5.52	36.5
3.21	27.2
4.32	27.7
2.31	28.3
4.30	30.3
3.71	28.7

(a) Calculate r.

(b) Test the null hypothesis that $\rho = 0$ against the alternative that $\rho > 0$ at the 0.01 level of significance.

(c) What percentage of the variation in infant chest sizes is explained by difference in weight?

Review Exercises

11.48 With reference to Exercise 11.8 on page 399, construct

(a) a 95% confidence interval for the average course grade of students who make a 35 on the placement test;

(b) a 95% prediction interval for the course grade of a student who made a 35 on the placement test.

11.49 The Statistics Consulting Center at Virginia Tech analyzed data on normal woodchucks for the Department of Veterinary Medicine. The variables of interest were body weight in grams and heart weight in grams. It was desired to develop a linear regression equation in order to determine if there is a significant linear relationship between heart weight and total body weight. Use heart weight as the independent variable and body weight as the dependent variable and fit a simple linear regression using the following data. In addition, test the hypothesis H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$. Draw conclusions.

Body Weight (grams)	Heart Weight (grams)
4050	11.2
2465	12.4
3120	10.5
5700	13.2
2595	9.8
3640	11.0
2050	10.8
4235	10.4
2935	12.2
4975	11.2
3690	10.8
2800	14.2
2775	12.2
2170	10.0
2370	12.3
2055	12.5
2025	11.8
2645	16.0
2675	13.8

11.50 The amounts of solids removed from a particular material when exposed to drying periods of different lengths are as shown.

x (hours)	y (grams)	
4.4	13.1	14.2
4.5	9.0	11.5
4.8	10.4	11.5
5.5	13.8	14.8
5.7	12.7	15.1
5.9	9.9	12.7
6.3	13.8	16.5
6.9	16.4	15.7
7.5	17.6	16.9
7.8	18.3	17.2

(a) Estimate the linear regression line.

(b) Test at the 0.05 level of significance whether the linear model is adequate.

11.51 With reference to Exercise 11.9 on page 399, construct

(a) a 95% confidence interval for the average weekly sales when \$45 is spent on advertising;

(b) a 95% prediction interval for the weekly sales when \$45 is spent on advertising.

11.52 An experiment was designed for the Department of Materials Engineering at Virginia Tech to study hydrogen embrittlement properties based on electrolytic hydrogen pressure measurements. The so-

lution used was 0.1 N NaOH, and the material was a certain type of stainless steel. The cathodic charging current density was controlled and varied at four levels. The effective hydrogen pressure was observed as the response. The data follow.

Run	Charging Current Density, x (mA/cm^2)	Effective Hydrogen Pressure, y (atm)
1	0.5	86.1
2	0.5	92.1
3	0.5	64.7
4	0.5	74.7
5	1.5	223.6
6	1.5	202.1
7	1.5	132.9
8	2.5	413.5
9	2.5	231.5
10	2.5	466.7
11	2.5	365.3
12	3.5	493.7
13	3.5	382.3
14	3.5	447.2
15	3.5	563.8

(a) Run a simple linear regression of y against x.

(b) Compute the pure error sum of squares and make a test for lack of fit.

(c) Does the information in part (b) indicate a need for a model in x beyond a first-order regression? Explain.

11.53 The following data represent the chemistry grades for a random sample of 12 freshmen at a certain college along with their scores on an intelligence test administered while they were still seniors in high school.

Student	Test Score, x	Chemistry Grade, y
1	65	85
2	50	74
3	55	76
4	65	90
5	55	85
6	70	87
7	65	94
8	70	98
9	55	81
10	70	91
11	50	76
12	55	74

(a) Compute and interpret the sample correlation coefficient.

(b) State necessary assumptions on random variables.

(c) Test the hypothesis that $\rho = 0.5$ against the alternative that $\rho > 0.5$. Use a P-value in the conclusion.

11.54 The business section of the *Washington Times* in March of 1997 listed 21 different used computers and printers and their sale prices. Also listed was the average hover bid. Partial results from regression analysis using *SAS* software are shown in Figure 11.29 on page 439.

(a) Explain the difference between the confidence interval on the mean and the prediction interval.

(b) Explain why the standard errors of prediction vary from observation to observation.

(c) Which observation has the lowest standard error of prediction? Why?

11.55 Consider the vehicle data from *Consumer Reports* in Figure 11.30 on page 440. Weight is in tons, mileage in miles per gallon, and drive ratio is also indicated. A regression model was fitted relating weight x to mileage y. A partial *SAS* printout in Figure 11.30 on page 440 shows some of the results of that regression analysis, and Figure 11.31 on page 441 gives a plot of the residuals and weight for each vehicle.

(a) From the analysis and the residual plot, does it appear that an improved model might be found by using a transformation? Explain.

(b) Fit the model by replacing weight with log weight. Comment on the results.

(c) Fit a model by replacing mpg with gallons per 100 miles traveled, as mileage is often reported in other countries. Which of the three models is preferable? Explain.

11.56 Observations on the yield of a chemical reaction taken at various temperatures were recorded as follows:

x (°C)	y (%)	x (°C)	y (%)
150	75.4	150	77.7
150	81.2	200	84.4
200	85.5	200	85.7
250	89.0	250	89.4
250	90.5	300	94.8
300	96.7	300	95.3

(a) Plot the data.

(b) Does it appear from the plot as if the relationship is linear?

(c) Fit a simple linear regression and test for lack of fit.

(d) Draw conclusions based on your result in (c).

11.57 Physical fitness testing is an important aspect of athletic training. A common measure of the magnitude of cardiovascular fitness is the maximum volume of oxygen uptake during strenuous exercise. A study was conducted on 24 middle-aged men to determine the influence on oxygen uptake of the time required to complete a two-mile run. Oxygen uptake

was measured with standard laboratory methods as the subjects performed on a treadmill. The work was published in "Maximal Oxygen Intake Prediction in Young and Middle Aged Males," *Journal of Sports Medicine* **9**, 1969, 17–22. The data are as follows:

Subject	y, Maximum Volume of O_2	x, Time in Seconds
1	42.33	918
2	53.10	805
3	42.08	892
4	50.06	962
5	42.45	968
6	42.46	907
7	47.82	770
8	49.92	743
9	36.23	1045
10	49.66	810
11	41.49	927
12	46.17	813
13	46.18	858
14	43.21	860
15	51.81	760
16	53.28	747
17	53.29	743
18	47.18	803
19	56.91	683
20	47.80	844
21	48.65	755
22	53.67	700
23	60.62	748
24	56.73	775

(a) Estimate the parameters in a simple linear regression model.

(b) Does the time it takes to run two miles have a significant influence on maximum oxygen uptake? Use H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$.

(c) Plot the residuals on a graph against x and comment on the appropriateness of the simple linear model.

11.58 Suppose a scientist postulates a model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \ i = 1, 2, \ldots, n,$$

and β_0 is a **known value**, not necessarily zero.

(a) What is the appropriate least squares estimator of β_1? Justify your answer.

(b) What is the variance of the slope estimator?

11.59 For the simple linear regression model, prove that $E(s^2) = \sigma^2$.

11.60 Assuming that the ϵ_i are independent and normally distributed with zero means and common variance σ^2, show that B_0, the least squares estimator of β_0 in $\mu_{Y|x} = \beta_0 + \beta_1 x$, is normally distributed with mean β_0 and variance

$$\sigma^2_{B_0} = \frac{\sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2} \sigma^2.$$

11.61 For a simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

where the ϵ_i are independent and normally distributed with zero means and equal variances σ^2, show that $\bar{Y}$ and

$$B_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) Y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

have zero covariance.

11.62 Show, in the case of a least squares fit to the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

that $\sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} e_i = 0$.

11.63 Consider the situation of Review Exercise 11.62 but suppose $n = 2$ (i.e., only two data points are available). Give an argument that the least squares regression line will result in $(y_1 - \hat{y}_1) = (y_2 - \hat{y}_2) = 0$. Also show that for this case $R^2 = 1.0$.

11.64 In Review Exercise 11.62, the student was required to show that $\sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$ for a standard simple linear regression model. Does the same hold for a model with zero intercept? Show why or why not.

11.65 Suppose that an experimenter postulates a model of the type

$$Y_i = \beta_0 + \beta_1 x_{1i} + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

when in fact an additional variable, say x_2, also contributes linearly to the response. The true model is then given by

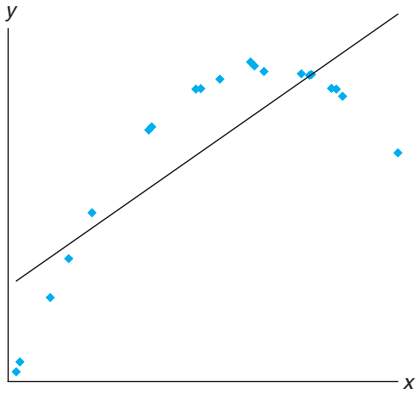
$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \quad i = 1, 2, \ldots, n.$$

Compute the expected value of the estimator

$$B_1 = \frac{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1) Y_i}{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2}.$$

11.66 Show the necessary steps in converting the equation $r = \frac{b_1}{s/\sqrt{S_{xx}}}$ to the equivalent form $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$.

11.67 Consider the fictitious set of data shown below, where the line through the data is the fitted simple linear regression line. Sketch a residual plot.



11.68 Project: This project can be done in groups or as individuals. Each group or person must find a set of data, preferably but not restricted to their field of study. The data need to fit the regression framework with a regression variable x and a response variable y. Carefully make the assignment as to which variable is x and which y. It may be necessary to consult a journal or periodical from your field if you do not have other research data available.

(a) Plot y versus x. Comment on the relationship as seen from the plot.

(b) Fit an appropriate regression model from the data. Use simple linear regression or fit a polynomial model to the data. Comment on measures of quality.

(c) Plot residuals as illustrated in the text. Check possible violation of assumptions. Show graphically a plot of confidence intervals on a mean response plotted against x. Comment.

R-Square	Coeff Var	Root MSE	Price Mean
0.967472	7.923338	70.83841	894.0476

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	59.93749137	38.34195754	1.56	0.1345
Buyer	1.04731316	0.04405635	23.77	<.0001

product	Buyer Price	Predict Value	Std Err Predict	Lower 95% Mean	Upper 95% Mean	Lower 95% Predict	Upper 95% Predict
IBM PS/1 486/66 420MB	325	375	400.31	25.8906	346.12	454.50	242.46	558.17
IBM ThinkPad 500	450	625	531.23	21.7232	485.76	576.70	376.15	686.31
IBM Think-Dad 755CX	1700	1850	1840.37	42.7041	1750.99	1929.75	1667.25	2013.49
AST Pentium 90 540MB	800	875	897.79	15.4590	865.43	930.14	746.03	1049.54
Dell Pentium 75 1GB	650	700	740.69	16.7503	705.63	775.75	588.34	893.05
Gateway 486/75 320MB	700	750	793.06	16.0314	759.50	826.61	641.04	945.07
Clone 586/133 1GB	500	600	583.59	20.2363	541.24	625.95	429.40	737.79
Compaq Contura 4/25 120MB	450	600	531.23	21.7232	485.76	576.70	376.15	686.31
Compaq Deskpro P90 1.2GB	800	850	897.79	15.4590	865.43	930.14	746.03	1049.54
Micron P75 810MB	800	675	897.79	15.4590	865.43	930.14	746.03	1049.54
Micron P100 1.2GB	900	975	1002.52	16.1176	968.78	1036.25	850.46	1154.58
Mac Quadra 840AV 500MB	450	575	531.23	21.7232	485.76	576.70	376.15	686.31
Mac Performer 6116 700MB	700	775	793.06	16.0314	759.50	826.61	641.04	945.07
PowerBook 540c 320MB	1400	1500	1526.18	30.7579	1461.80	1590.55	1364.54	1687.82
PowerBook 5300 500MB	1350	1575	1473.81	28.8747	1413.37	1534.25	1313.70	1633.92
Power Mac 7500/100 1GB	1150	1325	1264.35	21.9454	1218.42	1310.28	1109.13	1419.57
NEC Versa 486 340MB	800	900	897.79	15.4590	865.43	930.14	746.03	1049.54
Toshiba 1960CS 320MB	700	825	793.06	16.0314	759.50	826.61	641.04	945.07
Toshiba 4800VCT 500MB	1000	1150	1107.25	17.8715	1069.85	1144.66	954.34	1260.16
HP Laser jet III	350	475	426.50	25.0157	374.14	478.86	269.26	583.74
Apple Laser Writer Pro 63	750	800	845.42	15.5930	812.79	878.06	693.61	997.24

Figure 11.29: *SAS* printout, showing partial analysis of data of Review Exercise 11.54.

Obs	Model	WT	MPG	DR_RATIO
1	Buick Estate Wagon	4.360	16.9	2.73
2	Ford Country Squire Wagon	4.054	15.5	2.26
3	Chevy Ma libu Wagon	3.605	19.2	2.56
4	Chrysler LeBaron Wagon	3.940	18.5	2.45
5	Chevette	2.155	30.0	3.70
6	Toyota Corona	2.560	27.5	3.05
7	Datsun 510	2.300	27.2	3.54
8	Dodge Omni	2.230	30.9	3.37
9	Audi 5000	2.830	20.3	3.90
10	Volvo 240 CL	3.140	17.0	3.50
11	Saab 99 GLE	2.795	21.6	3.77
12	Peugeot 694 SL	3.410	16.2	3.58
13	Buick Century Special	3.380	20.6	2.73
14	Mercury Zephyr	3.070	20.8	3.08
15	Dodge Aspen	3.620	18.6	2.71
16	AMC Concord D/L	3.410	18.1	2.73
17	Chevy Caprice Classic	3.840	17.0	2.41
18	Ford LTP	3.725	17.6	2.26
19	Mercury Grand Marquis	3.955	16.5	2.26
20	Dodge St Regis	3.830	18.2	2.45
21	Ford Mustang 4	2.585	26.5	3.08
22	Ford Mustang Ghia	2.910	21.9	3.08
23	Macda GLC	1.975	34.1	3.73
24	Dodge Colt	1.915	35.1	2.97
25	AMC Spirit	2.670	27.4	3.08
26	VW Scirocco	1.990	31.5	3.78
27	Honda Accord LX	2.135	29.5	3.05
28	Buick Skylark	2.570	28.4	2.53
29	Chevy Citation	2.595	28.8	2.69
30	Olds Omega	2.700	26.8	2.84
31	Pontiac Phoenix	2.556	33.5	2.69
32	Plymouth Horizon	2.200	34.2	3.37
33	Datsun 210	2.020	31.8	3.70
34	Fiat Strada	2.130	37.3	3.10
35	VW Dasher	2.190	30.5	3.70
36	Datsun 810	2.815	22.0	3.70
37	BMW 320i	2.600	21.5	3.64
38	VW Rabbit	1.925	31.9	3.78

R-Square	Coeff Var	Root MSE	MPG Mean
0.817244	11.46010	2.837580	24.76053

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	48.67928080	1.94053995	25.09	<.0001
WT	-8.36243141	0.65908398	-12.69	<.0001

Figure 11.30: *SAS* printout, showing partial analysis of data of Review Exercise 11.55.

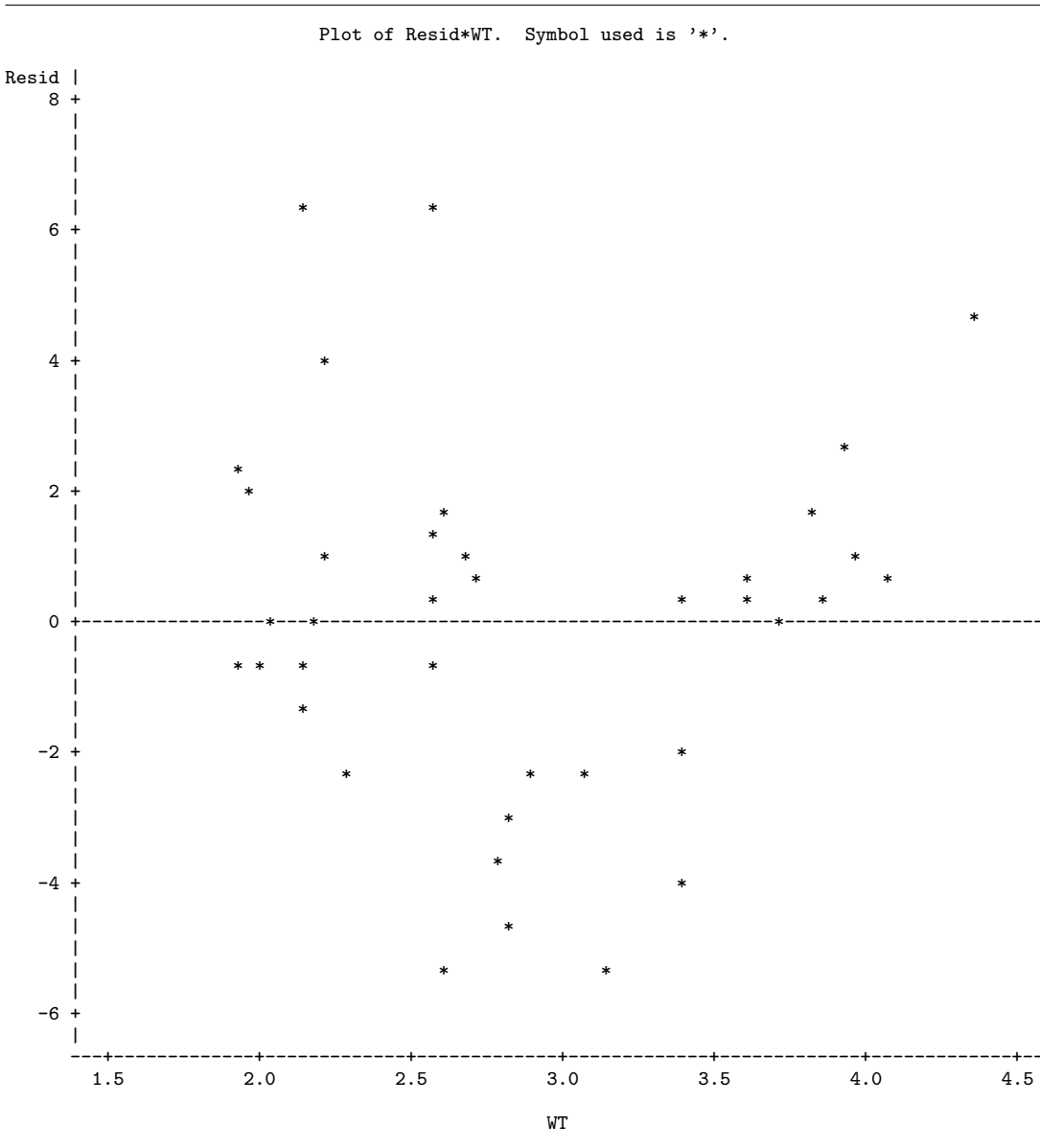


Figure 11.31: *SAS* printout, showing residual plot of Review Exercise 11.55.

11.13 Potential Misconceptions and Hazards; Relationship to Material in Other Chapters

Anytime one is considering the use of simple linear regression, a plot of the data is not only recommended but essential. A plot of the ordinary residuals and a normal probability plot of these residuals are always edifying. In addition, we introduce and illustrate an additional type of residual in Chapter 12 that is in a standardized form. All of these plots are designed to detect violation of assumptions.

The use of t-statistics for tests on regression coefficients is reasonably robust to the normality assumption. The homogeneous variance assumption is crucial, and residual plots are designed to detect a violation.

The material in this chapter is used heavily in Chapters 12 and 15. All of the information involving the method of least squares in the development of regression models carries over into Chapter 12. The difference is that Chapter 12 deals with the scientific conditions in which there is more than a single x variable, i.e., more than one regression variable. However, material in the current chapter that deals with regression diagnostics, types of residual plots, measures of model quality, and so on, applies and will carry over. The student will realize that more complications occur in Chapter 12 because the problems in multiple regression models often involve the backdrop of questions regarding how the various regression variables enter the model and even issues of which variables should remain in the model. Certainly Chapter 15 heavily involves the use of regression modeling, but we will preview the connection in the summary at the end of Chapter 12.

Chapter 12

Multiple Linear Regression and Certain Nonlinear Regression Models

12.1 Introduction

In most research problems where regression analysis is applied, more than one independent variable is needed in the regression model. The complexity of most scientific mechanisms is such that in order to be able to predict an important response, a **multiple regression model** is needed. When this model is linear in the coefficients, it is called a **multiple linear regression model**. For the case of k independent variables $x_1, x_2, \ldots, x_k$, the mean of $Y|x_1, x_2, \ldots, x_k$ is given by the multiple linear regression model

$$\mu_{Y|x_1,x_2,\ldots,x_k} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k,$$

and the estimated response is obtained from the sample regression equation

$$\hat{y} = b_0 + b_1 x_1 + \cdots + b_k x_k,$$

where each regression coefficient β_i is estimated by b_i from the sample data using the method of least squares. As in the case of a single independent variable, the multiple linear regression model can often be an adequate representation of a more complicated structure within certain ranges of the independent variables.

Similar least squares techniques can also be applied for estimating the coefficients when the linear model involves, say, powers and products of the independent variables. For example, when $k = 1$, the experimenter may believe that the means $\mu_{Y|x}$ do not fall on a straight line but are more appropriately described by the **polynomial regression model**

$$\mu_{Y|x} = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_r x^r,$$

and the estimated response is obtained from the polynomial regression equation

$$\hat{y} = b_0 + b_1 x + b_2 x^2 + \cdots + b_r x^r.$$

Confusion arises occasionally when we speak of a polynomial model as a linear model. However, statisticians normally refer to a linear model as one in which the parameters occur linearly, regardless of how the independent variables enter the model. An example of a nonlinear model is the **exponential relationship**

$$\mu_{Y|x} = \alpha\beta^x,$$

whose response is estimated by the regression equation

$$\hat{y} = ab^x.$$

There are many phenomena in science and engineering that are inherently nonlinear in nature, and when the true structure is known, an attempt should certainly be made to fit the actual model. The literature on estimation by least squares of nonlinear models is voluminous. The nonlinear models discussed in this chapter deal with nonideal conditions in which the analyst is certain that the response and hence the response model error are not normally distributed but, rather, have a binomial or Poisson distribution. These situations do occur extensively in practice.

A student who wants a more general account of nonlinear regression should consult *Classical and Modern Regression with Applications* by Myers (1990; see the Bibliography).

12.2 Estimating the Coefficients

In this section, we obtain the least squares estimators of the parameters $\beta_0, \beta_1, \ldots, \beta_k$ by fitting the multiple linear regression model

$$\mu_{Y|x_1,x_2,\ldots,x_k} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

to the data points

$$\{(x_{1i}, x_{2i}, \ldots, x_{ki}, y_i); \quad i = 1, 2, \ldots, n \text{ and } n > k\},$$

where y_i is the observed response to the values $x_{1i}, x_{2i}, \ldots, x_{ki}$ of the k independent variables $x_1, x_2, \ldots, x_k$. Each observation $(x_{1i}, x_{2i}, \ldots, x_{ki}, y_i)$ is assumed to satisfy the following equation.

Multiple Linear Regression Model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i$$

or

$$y_i = \hat{y}_i + e_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki} + e_i,$$

where ϵ_i and e_i are the random error and residual, respectively, associated with the response y_i and fitted value $\hat{y}_i$.

As in the case of simple linear regression, it is assumed that the ϵ_i are independent and identically distributed with mean 0 and common variance σ^2.

In using the concept of least squares to arrive at estimates $b_0, b_1, \ldots, b_k$, we minimize the expression

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} - \cdots - b_k x_{ki})^2.$$

Differentiating SSE in turn with respect to $b_0, b_1, \ldots, b_k$ and equating to zero, we generate the set of $k + 1$ **normal equations for multiple linear regression**.

Normal Estimation Equations for Multiple Linear Regression

$$\begin{aligned}
nb_0 + b_1\sum_{i=1}^{n} x_{1i} \quad &+ b_2\sum_{i=1}^{n} x_{2i} \quad + \cdots + b_k\sum_{i=1}^{n} x_{ki} \quad = \sum_{i=1}^{n} y_i \\
b_0\sum_{i=1}^{n} x_{1i} + b_1\sum_{i=1}^{n} x_{1i}^2 \quad &+ b_2\sum_{i=1}^{n} x_{1i}x_{2i} + \cdots + b_k\sum_{i=1}^{n} x_{1i}x_{ki} = \sum_{i=1}^{n} x_{1i}y_i \\
&\vdots \qquad\qquad \vdots \qquad\qquad\qquad \vdots \qquad\qquad\qquad \vdots \qquad\qquad \vdots \\
b_0\sum_{i=1}^{n} x_{ki} + b_1\sum_{i=1}^{n} x_{ki}x_{1i} &+ b_2\sum_{i=1}^{n} x_{ki}x_{2i} + \cdots + b_k\sum_{i=1}^{n} x_{ki}^2 \quad = \sum_{i=1}^{n} x_{ki}y_i
\end{aligned}$$

These equations can be solved for $b_0, b_1, b_2, \ldots, b_k$ by any appropriate method for solving systems of linear equations. Most statistical software can be used to obtain numerical solutions of the above equations.

Example 12.1: A study was done on a diesel-powered light-duty pickup truck to see if humidity, air temperature, and barometric pressure influence emission of nitrous oxide (in ppm). Emission measurements were taken at different times, with varying experimental conditions. The data are given in Table 12.2. The model is

$$\mu_{Y|x_1,x_2,x_3} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,$$

or, equivalently,

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i, \quad i = 1, 2, \ldots, 20.$$

Fit this multiple linear regression model to the given data and then estimate the amount of nitrous oxide emitted for the conditions where humidity is 50%, temperature is 76°F, and barometric pressure is 29.30.

Table 12.1: Data for Example 12.1

Nitrous Oxide, y	Humidity, x_1	Temp., x_2	Pressure, x_3	Nitrous Oxide, y	Humidity, x_1	Temp., x_2	Pressure, x_3
0.90	72.4	76.3	29.18	1.07	23.2	76.8	29.38
0.91	41.6	70.3	29.35	0.94	47.4	86.6	29.35
0.96	34.3	77.1	29.24	1.10	31.5	76.9	29.63
0.89	35.1	68.0	29.27	1.10	10.6	86.3	29.56
1.00	10.7	79.0	29.78	1.10	11.2	86.0	29.48
1.10	12.9	67.4	29.39	0.91	73.3	76.3	29.40
1.15	8.3	66.8	29.69	0.87	75.4	77.9	29.28
1.03	20.1	76.9	29.48	0.78	96.6	78.7	29.29
0.77	72.2	77.7	29.09	0.82	107.4	86.8	29.03
1.07	24.0	67.7	29.60	0.95	54.9	70.9	29.37

Source: Charles T. Hare, "Light-Duty Diesel Emission Correction Factors for Ambient Conditions," EPA-600/2-77-116. U.S. Environmental Protection Agency.

***Solution*:** The solution of the set of estimating equations yields the unique estimates

$$b_0 = -3.507778,\ b_1 = -0.002625,\ b_2 = 0.000799,\ b_3 = 0.154155.$$

Therefore, the regression equation is

$$\hat{y} = -3.507778 - 0.002625\,x_1 + 0.000799\,x_2 + 0.154155\,x_3.$$

For 50% humidity, a temperature of 76°F, and a barometric pressure of 29.30, the estimated amount of nitrous oxide emitted is

$$\begin{aligned}\hat{y} &= -3.507778 - 0.002625(50.0) + 0.000799(76.0) + 0.1541553(29.30)\\ &= 0.9384 \text{ ppm.}\end{aligned}$$

Polynomial Regression

Now suppose that we wish to fit the polynomial equation

$$\mu_{Y|x} = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_r x^r$$

to the n pairs of observations $\{(x_i, y_i);\ i = 1, 2, \ldots, n\}$. Each observation, y_i, satisfies the equation

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_r x_i^r + \epsilon_i$$

or

$$y_i = \hat{y}_i + e_i = b_0 + b_1 x_i + b_2 x_i^2 + \cdots + b_r x_i^r + e_i,$$

where r is the degree of the polynomial and ϵ_i and e_i are again the random error and residual associated with the response y_i and fitted value $\hat{y}_i$, respectively. Here, the number of pairs, n, must be at least as large as $r+1$, the number of parameters to be estimated.

Notice that the polynomial model can be considered a special case of the more general multiple linear regression model, where we set $x_1 = x, x_2 = x^2, \ldots, x_r = x^r$. The normal equations assume the same form as those given on page 445. They are then solved for $b_0, b_1, b_2, \ldots, b_r$.

Example 12.2: Given the data

x	0	1	2	3	4	5	6	7	8	9
y	9.1	7.3	3.2	4.6	4.8	2.9	5.7	7.1	8.8	10.2

fit a regression curve of the form $\mu_{Y|x} = \beta_0 + \beta_1 x + \beta_2 x^2$ and then estimate $\mu_{Y|2}$.

***Solution*:** From the data given, we find that

$$\begin{aligned}10b_0 + \quad 45\,b_1 + \quad 285\,b_2 &= 63.7,\\ 45b_0 + \quad 285b_1 + \quad 2025\,b_2 &= 307.3,\\ 285b_0 + 2025\,b_1 + 15{,}333b_2 &= 2153.3.\end{aligned}$$

Solving these normal equations, we obtain

$$b_0 = 8.698, \quad b_1 = -2.341, \quad b_2 = 0.288.$$

Therefore,

$$\hat{y} = 8.698 - 2.341x + 0.288x^2.$$

When $x = 2$, our estimate of $\mu_{Y|2}$ is

$$\hat{y} = 8.698 - (2.341)(2) + (0.288)(2^2) = 5.168.$$

Example 12.3: The data in Table 12.2 represent the percent of impurities that resulted for various temperatures and sterilizing times during a reaction associated with the manufacturing of a certain beverage. Estimate the regression coefficients in the polynomial model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{11} x_{1i}^2 + \beta_{22} x_{2i}^2 + \beta_{12} x_{1i} x_{2i} + \epsilon_i,$$

for $i = 1, 2, \ldots, 18$.

Table 12.2: Data for Example 12.3

Sterilizing Time, x_2 (min)	**Temperature, x_1 (°C)**		
	75	**100**	**125**
15	14.05	10.55	7.55
	14.93	9.48	6.59
20	16.56	13.63	9.23
	15.85	11.75	8.78
25	22.41	18.55	15.93
	21.66	17.98	16.44

***Solution*:** Using the normal equations, we obtain

$$b_0 = 56.4411, \quad b_1 = -0.36190, \quad b_2 = -2.75299,$$
$$b_{11} = 0.00081, \quad b_{22} = 0.08173, \quad b_{12} = 0.00314,$$

and our estimated regression equation is

$$\hat{y} = 56.4411 - 0.36190x_1 - 2.75299x_2 + 0.00081x_1^2 + 0.08173x_2^2 + 0.00314x_1x_2.$$

Many of the principles and procedures associated with the estimation of polynomial regression functions fall into the category of **response surface methodology**, a collection of techniques that have been used quite successfully by scientists and engineers in many fields. The x_i^2 are called **pure quadratic terms**, and the $x_i x_j$ $(i \neq j)$ are called **interaction terms**. Such problems as selecting a proper experimental design, particularly in cases where a large number of variables are in the model, and choosing optimum operating conditions for $x_1, x_2, \ldots, x_k$ are often approached through the use of these methods. For an extensive exposure, the reader is referred to *Response Surface Methodology: Process and Product Optimization Using Designed Experiments* by Myers, Montgomery, and Anderson-Cook (2009; see the Bibliography).

12.3 Linear Regression Model Using Matrices

In fitting a multiple linear regression model, particularly when the number of variables exceeds two, a knowledge of matrix theory can facilitate the mathematical manipulations considerably. Suppose that the experimenter has k independent

variables $x_1, x_2, \ldots, x_k$ and n observations $y_1, y_2, \ldots, y_n$, each of which can be expressed by the equation

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i.$$

This model essentially represents n equations describing how the response values are generated in the scientific process. Using matrix notation, we can write the following equation:

General Linear Model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Then the least squares method for estimation of β, illustrated in Section 12.2, involves finding $\mathbf{b}$ for which

$$SSE = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})$$

is minimized. This minimization process involves solving for $\mathbf{b}$ in the equation

$$\frac{\partial}{\partial \mathbf{b}}(SSE) = \mathbf{0}.$$

We will not present the details regarding solution of the equations above. The result reduces to the solution of $\mathbf{b}$ in

$$(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{y}.$$

Notice the nature of the $\mathbf{X}$ matrix. Apart from the initial element, the ith row represents the x-values that give rise to the response y_i. Writing

$$\mathbf{A} = \mathbf{X}'\mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{2i} & \cdots & \sum_{i=1}^{n} x_{ki} \\ \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{1i}^2 & \sum_{i=1}^{n} x_{1i}x_{2i} & \cdots & \sum_{i=1}^{n} x_{1i}x_{ki} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum_{i=1}^{n} x_{ki} & \sum_{i=1}^{n} x_{ki}x_{1i} & \sum_{i=1}^{n} x_{ki}x_{2i} & \cdots & \sum_{i=1}^{n} x_{ki}^2 \end{bmatrix}$$

and

$$\mathbf{g} = \mathbf{X}'\mathbf{y} = \begin{bmatrix} g_0 = \sum_{i=1}^{n} y_i \\ g_1 = \sum_{i=1}^{n} x_{1i}y_i \\ \vdots \\ g_k = \sum_{i=1}^{n} x_{ki}y_i \end{bmatrix}$$

allows the normal equations to be put in the matrix form

$$\mathbf{A}\mathbf{b} = \mathbf{g}.$$

If the matrix $\mathbf{A}$ is nonsingular, we can write the solution for the regression coefficients as

$$\mathbf{b} = \mathbf{A}^{-1}\mathbf{g} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

Thus, we can obtain the prediction equation or regression equation by solving a set of $k+1$ equations in a like number of unknowns. This involves the inversion of the $k+1$ by $k+1$ matrix $\mathbf{X}'\mathbf{X}$. Techniques for inverting this matrix are explained in most textbooks on elementary determinants and matrices. Of course, there are many high-speed computer packages available for multiple regression problems, packages that not only print out estimates of the regression coefficients but also provide other information relevant to making inferences concerning the regression equation.

Example 12.4: The percent survival rate of sperm in a certain type of animal semen, after storage, was measured at various combinations of concentrations of three materials used to increase chance of survival. The data are given in Table 12.3. Estimate the multiple linear regression model for the given data.

Table 12.3: Data for Example 12.4

y (% survival)	x_1 (weight %)	x_2 (weight %)	x_3 (weight %)
25.5	1.74	5.30	10.80
31.2	6.32	5.42	9.40
25.9	6.22	8.41	7.20
38.4	10.52	4.63	8.50
18.4	1.19	11.60	9.40
26.7	1.22	5.85	9.90
26.4	4.10	6.62	8.00
25.9	6.32	8.72	9.10
32.0	4.08	4.42	8.70
25.2	4.15	7.60	9.20
39.7	10.15	4.83	9.40
35.7	1.72	3.12	7.60
26.5	1.70	5.30	8.20

***Solution*:** The least squares estimating equations, $(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{y}$, are

$$\begin{bmatrix} 13.0 & 59.43 & 81.82 & 115.40 \\ 59.43 & 394.7255 & 360.6621 & 522.0780 \\ 81.82 & 360.6621 & 576.7264 & 728.3100 \\ 115.40 & 522.0780 & 728.3100 & 1035.9600 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 377.5 \\ 1877.567 \\ 2246.661 \\ 3337.780 \end{bmatrix}.$$

From a computer readout we obtain the elements of the inverse matrix

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 8.0648 & -0.0826 & -0.0942 & -0.7905 \\ -0.0826 & 0.0085 & 0.0017 & 0.0037 \\ -0.0942 & 0.0017 & 0.0166 & -0.0021 \\ -0.7905 & 0.0037 & -0.0021 & 0.0886 \end{bmatrix},$$

and then, using the relation $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, the estimated regression coefficients are obtained as

$$b_0 = 39.1574,\ b_1 = 1.0161,\ b_2 = -1.8616,\ b_3 = -0.3433.$$

Hence, our estimated regression equation is

$$\hat{y} = 39.1574 + 1.0161x_1 - 1.8616x_2 - 0.3433x_3.$$

Exercises

12.1 A set of experimental runs was made to determine a way of predicting cooking time y at various values of oven width x_1 and flue temperature x_2. The coded data were recorded as follows:

y	x_1	x_2
6.40	1.32	1.15
15.05	2.69	3.40
18.75	3.56	4.10
30.25	4.41	8.75
44.85	5.35	14.82
48.94	6.20	15.15
51.55	7.12	15.32
61.50	8.87	18.18
100.44	9.80	35.19
111.42	10.65	40.40

Estimate the multiple linear regression equation

$$\mu_{Y|x_1,x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

12.2 In *Applied Spectroscopy*, the infrared reflectance spectra properties of a viscous liquid used in the electronics industry as a lubricant were studied. The designed experiment consisted of the effect of band frequency x_1 and film thickness x_2 on optical density y using a Perkin-Elmer Model 621 infrared spectrometer. (*Source*: Pacansky, J., England, C. D., and Wattman, R., 1986.)

y	x_1	x_2
0.231	740	1.10
0.107	740	0.62
0.053	740	0.31
0.129	805	1.10
0.069	805	0.62
0.030	805	0.31
1.005	980	1.10
0.559	980	0.62
0.321	980	0.31
2.948	1235	1.10
1.633	1235	0.62
0.934	1235	0.31

Estimate the multiple linear regression equation

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2.$$

12.3 Suppose in Review Exercise 11.53 on page 437 that we were also given the number of class periods missed by the 12 students taking the chemistry course. The complete data are shown.

Student	Chemistry Grade, y	Test Score, x_1	Classes Missed, x_2
1	85	65	1
2	74	50	7
3	76	55	5
4	90	65	2
5	85	55	6
6	87	70	3
7	94	65	2
8	98	70	5
9	81	55	4
10	91	70	3
11	76	50	1
12	74	55	4

(a) Fit a multiple linear regression equation of the form $\hat{y} = b_0 + b_1 x_1 + b_2 x_2$.

(b) Estimate the chemistry grade for a student who has an intelligence test score of 60 and missed 4 classes.

12.4 An experiment was conducted to determine if the weight of an animal can be predicted after a given period of time on the basis of the initial weight of the animal and the amount of feed that was eaten. The following data, measured in kilograms, were recorded:

Final Weight, y	Initial Weight, x_1	Feed Weight, x_2
95	42	272
77	33	226
80	33	259
100	45	292
97	39	311
70	36	183
50	32	173
80	41	236
92	40	230
84	38	235

(a) Fit a multiple regression equation of the form

$$\mu_{Y|x_1,x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

(b) Predict the final weight of an animal having an initial weight of 35 kilograms that is given 250 kilograms of feed.

12.5 The electric power consumed each month by a chemical plant is thought to be related to the average ambient temperature x_1, the number of days in the month x_2, the average product purity x_3, and the tons of product produced x_4. The past year's historical data are available and are presented in the following table.

y	x_1	x_2	x_3	x_4
240	25	24	91	100
236	31	21	90	95
290	45	24	88	110
274	60	25	87	88
301	65	25	91	94
316	72	26	94	99
300	80	25	87	97
296	84	25	86	96
267	75	24	88	110
276	60	25	91	105
288	50	25	90	100
261	38	23	89	98

(a) Fit a multiple linear regression model using the above data set.

(b) Predict power consumption for a month in which $x_1 = 75°$F, $x_2 = 24$ days, $x_3 = 90\%$, and $x_4 = 98$ tons.

12.6 An experiment was conducted on a new model of a particular make of automobile to determine the stopping distance at various speeds. The following data were recorded.

Speed, v (km/hr)	35	50	65	80	95	110
Stopping Distance, d (m)	16	26	41	62	88	119

(a) Fit a multiple regression curve of the form $\mu_{D|v} = \beta_0 + \beta_1 v + \beta_2 v^2$.

(b) Estimate the stopping distance when the car is traveling at 70 kilometers per hour.

12.7 An experiment was conducted in order to determine if cerebral blood flow in human beings can be predicted from arterial oxygen tension (millimeters of mercury). Fifteen patients participated in the study, and the following data were collected:

Blood Flow, y	**Arterial Oxygen Tension, x**
84.33	603.40
87.80	582.50
82.20	556.20
78.21	594.60
78.44	558.90
80.01	575.20
83.53	580.10
79.46	451.20
75.22	404.00
76.58	484.00
77.90	452.40
78.80	448.40
80.67	334.80
86.60	320.30
78.20	350.30

Estimate the quadratic regression equation

$$\mu_{Y|x} = \beta_0 + \beta_1 x + \beta_2 x^2.$$

12.8 The following is a set of coded experimental data on the compressive strength of a particular alloy at various values of the concentration of some additive:

Concentration, x	**Compressive Strength, y**		
10.0	25.2	27.3	28.7
15.0	29.8	31.1	27.8
20.0	31.2	32.6	29.7
25.0	31.7	30.1	32.3
30.0	29.4	30.8	32.8

(a) Estimate the quadratic regression equation $\mu_{Y|x} = \beta_0 + \beta_1 x + \beta_2 x^2$.

(b) Test for lack of fit of the model.

12.9 (a) Fit a multiple regression equation of the form $\mu_{Y|x} = \beta_0 + \beta_1 x_1 + \beta_2 x^2$ to the data of Example 11.8 on page 420.

(b) Estimate the yield of the chemical reaction for a temperature of 225°C.

12.10 The following data are given:

x	0	1	2	3	4	5	6
y	1	4	5	3	2	3	4

(a) Fit the cubic model $\mu_{Y|x} = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$.

(b) Predict Y when $x = 2$.

12.11 An experiment was conducted to study the size of squid eaten by sharks and tuna. The regressor variables are characteristics of the beaks of the squid. The data are given as follows:

x_1	x_2	x_3	x_4	x_5	y
1.31	1.07	0.44	0.75	0.35	1.95
1.55	1.49	0.53	0.90	0.47	2.90
0.99	0.84	0.34	0.57	0.32	0.72
0.99	0.83	0.34	0.54	0.27	0.81
1.01	0.90	0.36	0.64	0.30	1.09
1.09	0.93	0.42	0.61	0.31	1.22
1.08	0.90	0.40	0.51	0.31	1.02
1.27	1.08	0.44	0.77	0.34	1.93
0.99	0.85	0.36	0.56	0.29	0.64
1.34	1.13	0.45	0.77	0.37	2.08
1.30	1.10	0.45	0.76	0.38	1.98
1.33	1.10	0.48	0.77	0.38	1.90
1.86	1.47	0.60	1.01	0.65	8.56
1.58	1.34	0.52	0.95	0.50	4.49
1.97	1.59	0.67	1.20	0.59	8.49
1.80	1.56	0.66	1.02	0.59	6.17
1.75	1.58	0.63	1.09	0.59	7.54
1.72	1.43	0.64	1.02	0.63	6.36
1.68	1.57	0.72	0.96	0.68	7.63
1.75	1.59	0.68	1.08	0.62	7.78
2.19	1.86	0.75	1.24	0.72	10.15
1.73	1.67	0.64	1.14	0.55	6.88

In the study, the regressor variables and response considered are

x_1 = rostral length, in inches,
x_2 = wing length, in inches,
x_3 = rostral to notch length, in inches,
x_4 = notch to wing length, in inches,
x_5 = width, in inches,
y = weight, in pounds.

Estimate the multiple linear regression equation

$$\mu_{Y|x_1,x_2,x_3,x_4,x_5} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5.$$

12.12 The following data reflect information from 17 U.S. Naval hospitals at various sites around the world. The regressors are workload variables, that is, items that result in the need for personnel in a hospital. A brief description of the variables is as follows:

y = monthly labor-hours,
x_1 = average daily patient load,
x_2 = monthly X-ray exposures,
x_3 = monthly occupied bed-days,
x_4 = eligible population in the area/1000,
x_5 = average length of patient's stay, in days.

Site	x_1	x_2	x_3	x_4	x_5	y
1	15.57	2463	472.92	18.0	4.45	566.52
2	44.02	2048	1339.75	9.5	6.92	696.82
3	20.42	3940	620.25	12.8	4.28	1033.15
4	18.74	6505	568.33	36.7	3.90	1003.62
5	49.20	5723	1497.60	35.7	5.50	1611.37
6	44.92	11,520	1365.83	24.0	4.60	1613.27
7	55.48	5779	1687.00	43.3	5.62	1854.17
8	59.28	5969	1639.92	46.7	5.15	2160.55
9	94.39	8461	2872.33	78.7	6.18	2305.58
10	128.02	20,106	3655.08	180.5	6.15	3503.93
11	96.00	13,313	2912.00	60.9	5.88	3571.59
12	131.42	10,771	3921.00	103.7	4.88	3741.40
13	127.21	15,543	3865.67	126.8	5.50	4026.52
14	252.90	36,194	7684.10	157.7	7.00	10,343.81
15	409.20	34,703	12,446.33	169.4	10.75	11,732.17
16	463.70	39,204	14,098.40	331.4	7.05	15,414.94
17	510.22	86,533	15,524.00	371.6	6.35	18,854.45

The goal here is to produce an empirical equation that will estimate (or predict) personnel needs for Naval hospitals. Estimate the multiple linear regression equation

$$\mu_{Y|x_1,x_2,x_3,x_4,x_5} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5.$$

12.13 A study was performed on a type of bearing to find the relationship of amount of wear y to x_1 = oil viscosity and x_2 = load. The following data were obtained. (From *Response Surface Methodology*, Myers, Montgomery, and Anderson-Cook, 2009.)

y	x_1	x_2	y	x_1	x_2
193	1.6	851	230	15.5	816
172	22.0	1058	91	43.0	1201
113	33.0	1357	125	40.0	1115

(a) Estimate the unknown parameters of the multiple linear regression equation

$$\mu_{Y|x_1,x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

(b) Predict wear when oil viscosity is 20 and load is 1200.

12.14 Eleven student teachers took part in an evaluation program designed to measure teacher effectiveness and determine what factors are important. The response measure was a quantitative evaluation of the teacher. The regressor variables were scores on four standardized tests given to each teacher. The data are as follows:

y	x_1	x_2	x_3	x_4
410	69	125	59.00	55.66
569	57	131	31.75	63.97
425	77	141	80.50	45.32
344	81	122	75.00	46.67
324	0	141	49.00	41.21
505	53	152	49.35	43.83
235	77	141	60.75	41.61
501	76	132	41.25	64.57
400	65	157	50.75	42.41
584	97	166	32.25	57.95
434	76	141	54.50	57.90

Estimate the multiple linear regression equation

$$\mu_{Y|x_1,x_2,x_3,x_4} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4.$$

12.15 The personnel department of a certain industrial firm used 12 subjects in a study to determine the relationship between job performance rating (y) and scores on four tests. The data are as follows:

y	x_1	x_2	x_3	x_4
11.2	56.5	71.0	38.5	43.0
14.5	59.5	72.5	38.2	44.8
17.2	69.2	76.0	42.5	49.0
17.8	74.5	79.5	43.4	56.3
19.3	81.2	84.0	47.5	60.2
24.5	88.0	86.2	47.4	62.0
21.2	78.2	80.5	44.5	58.1
16.9	69.0	72.0	41.8	48.1
14.8	58.1	68.0	42.1	46.0
20.0	80.5	85.0	48.1	60.3
13.2	58.3	71.0	37.5	47.1
22.5	84.0	87.2	51.0	65.2

Estimate the regression coefficients in the model

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4.$$

12.16 An engineer at a semiconductor company wants to model the relationship between the gain or hFE of a device (y) and three parameters: emitter-RS (x_1), base-RS (x_2), and emitter-to-base-RS (x_3). The data are shown below:

x_1, Emitter-RS	x_2, Base-RS	x_3, E-B-RS	y, hFE
14.62	226.0	7.000	128.40
15.63	220.0	3.375	52.62
14.62	217.4	6.375	113.90
15.00	220.0	6.000	98.01
14.50	226.5	7.625	139.90
15.25	224.1	6.000	102.60
16.12	220.5	3.375	48.14
15.13	223.5	6.125	109.60
15.50	217.6	5.000	82.68
15.13	228.5	6.625	112.60
15.50	230.2	5.750	97.52
16.12	226.5	3.750	59.06
15.13	226.6	6.125	111.80
15.63	225.6	5.375	89.09
15.38	234.0	8.875	171.90
15.50	230.0	4.000	66.80
14.25	224.3	8.000	157.10
14.50	240.5	10.870	208.40
14.62	223.7	7.375	133.40

(cont.)

(Data from Myers, Montgomery, and Anderson-Cook, 2009.)

(a) Fit a multiple linear regression to the data.

(b) Predict hFE when $x_1 = 14$, $x_2 = 220$, and $x_3 = 5$.

12.4 Properties of the Least Squares Estimators

The means and variances of the estimators $b_0, b_1, \ldots, b_k$ are readily obtained under certain assumptions on the random errors $\epsilon_1, \epsilon_2, \ldots, \epsilon_k$ that are identical to those made in the case of simple linear regression. When we assume these errors to be independent, each with mean 0 and variance σ^2, it can be shown that $b_0, b_1, \ldots, b_k$ are, respectively, unbiased estimators of the regression coefficients $\beta_0, \beta_1, \ldots, \beta_k$. In addition, the variances of the b's are obtained through the elements of the inverse of the $\mathbf{A}$ matrix. Note that the off-diagonal elements of $\mathbf{A} = \mathbf{X}'\mathbf{X}$ represent sums of products of elements in the columns of $\mathbf{X}$, while the diagonal elements of $\mathbf{A}$ represent sums of squares of elements in the columns of $\mathbf{X}$. The inverse matrix, $\mathbf{A}^{-1}$, apart from the multiplier σ^2, represents the **variance-covariance matrix** of the estimated regression coefficients. That is, the elements of the matrix $\mathbf{A}^{-1}\sigma^2$ display the variances of $b_0, b_1, \ldots, b_k$ on the main diagonal and covariances on the off-diagonal. For example, in a $k = 2$ multiple linear regression problem, we might write

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} c_{00} & c_{01} & c_{02} \\ c_{10} & c_{11} & c_{12} \\ c_{20} & c_{21} & c_{22} \end{bmatrix}$$

with the elements below the main diagonal determined through the symmetry of the matrix. Then we can write

$$\sigma^2_{b_i} = c_{ii}\sigma^2, \qquad i = 0, 1, 2,$$
$$\sigma_{b_ib_j} = \text{Cov}(b_i, b_j) = c_{ij}\sigma^2, \quad i \neq j.$$

Of course, the estimates of the variances and hence the standard errors of these estimators are obtained by replacing σ^2 with the appropriate estimate obtained through experimental data. An unbiased estimate of σ^2 is once again defined in

terms of the error sum of squares, which is computed using the formula established in Theorem 12.1. In the theorem, we are making the assumptions on the ϵ_i described above.

Theorem 12.1: For the linear regression equation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

an unbiased estimate of σ^2 is given by the error or residual mean square

$$s^2 = \frac{SSE}{n-k-1}, \quad \text{where} \quad SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

We can see that Theorem 12.1 represents a generalization of Theorem 11.1 for the simple linear regression case. The proof is left for the reader. As in the simpler linear regression case, the estimate s^2 is a measure of the variation in the prediction errors or residuals. Other important inferences regarding the fitted regression equation, based on the values of the individual residuals $e_i = y_i - \hat{y}_i$, $i = 1, 2, \ldots, n$, are discussed in Sections 12.10 and 12.11.

The error and regression sums of squares take on the same form and play the same role as in the simple linear regression case. In fact, the sum-of-squares identity

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

continues to hold, and we retain our previous notation, namely

$$SST = SSR + SSE,$$

with

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \text{total sum of squares}$$

and

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \text{regression sum of squares.}$$

There are k degrees of freedom associated with SSR, and, as always, SST has $n-1$ degrees of freedom. Therefore, after subtraction, SSE has $n-k-1$ degrees of freedom. Thus, our estimate of σ^2 is again given by the error sum of squares divided by its degrees of freedom. All three of these sums of squares will appear on the printouts of most multiple regression computer packages. Note that the condition $n > k$ in Section 12.2 guarantees that the degrees of freedom of SSE cannot be negative.

Analysis of Variance in Multiple Regression

The partition of the total sum of squares into its components, the regression and error sums of squares, plays an important role. An **analysis of variance** can be conducted to shed light on the quality of the regression equation. A useful hypothesis that determines if a significant amount of variation is explained by the model is

$$H_0: \ \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0.$$

The analysis of variance involves an F-test via a table given as follows:

Source	Sum of Squares	Degrees of Freedom	Mean Squares	F
Regression	SSR	k	$MSR = \frac{SSR}{k}$	$f = \frac{MSR}{MSE}$
Error	SSE	$n-(k+1)$	$MSE = \frac{SSE}{n-(k+1)}$	
Total	SST	$n-1$		

This test is an **upper-tailed test**. Rejection of H_0 implies that the **regression equation differs from a constant**. That is, at least one regressor variable is important. More discussion of the use of analysis of variance appears in subsequent sections.

Further utility of the mean square error (or residual mean square) lies in its use in hypothesis testing and confidence interval estimation, which is discussed in Section 12.5. In addition, the mean square error plays an important role in situations where the scientist is searching for the best from a set of competing models. Many model-building criteria involve the statistic s^2. Criteria for comparing competing models are discussed in Section 12.11.

12.5 Inferences in Multiple Linear Regression

A knowledge of the distributions of the individual coefficient estimators enables the experimenter to construct confidence intervals for the coefficients and to test hypotheses about them. Recall from Section 12.4 that the b_j $(j = 0, 1, 2, \ldots, k)$ are normally distributed with mean β_j and variance $c_{jj}\sigma^2$. Thus, we can use the statistic

$$t = \frac{b_j - \beta_{j0}}{s\sqrt{c_{jj}}}$$

with $n - k - 1$ degrees of freedom to test hypotheses and construct confidence intervals on β_j. For example, if we wish to test

$$H_0: \ \beta_j = \beta_{j0},$$
$$H_1: \ \beta_j \neq \beta_{j0},$$

we compute the above t-statistic and do not reject H_0 if $-t_{\alpha/2} < t < t_{\alpha/2}$, where $t_{\alpha/2}$ has $n - k - 1$ degrees of freedom.

Example 12.5: For the model of Example 12.4, test the hypothesis that $\beta_2 = -2.5$ at the 0.05 level of significance against the alternative that $\beta_2 > -2.5$.

***Solution*:**

$$H_0\colon\ \beta_2 = -2.5,$$
$$H_1\colon\ \beta_2 > -2.5.$$

Computations:

$$t = \frac{b_2 - \beta_{20}}{s\sqrt{c_{22}}} = \frac{-1.8616 + 2.5}{2.073\sqrt{0.0166}} = 2.390,$$
$$P = P(T > 2.390) = 0.04.$$

Decision: Reject H_0 and conclude that $\beta_2 > -2.5$.

Individual *t*-Tests for Variable Screening

The *t*-test most often used in multiple regression is the one that tests the importance of individual coefficients (i.e., H_0: $\beta_j = 0$ against the alternative H_1: $\beta_j \neq 0$). These tests often contribute to what is termed **variable screening**, where the analyst attempts to arrive at the most useful model (i.e., the choice of which regressors to use). It should be emphasized here that if a coefficient is found insignificant (i.e., the hypothesis H_0: $\beta_j = 0$ **is not rejected**), the conclusion drawn is that the **variable** is insignificant (i.e., explains an insignificant amount of variation in y), **in the presence of the other regressors in the model**. This point will be reaffirmed in a future discussion.

Inferences on Mean Response and Prediction

One of the most useful inferences that can be made regarding the quality of the predicted response y_0 corresponding to the values $x_{10}, x_{20}, \ldots, x_{k0}$ is the confidence interval on the mean response $\mu_{Y|x_{10},x_{20},\ldots,x_{k0}}$. We are interested in constructing a confidence interval on the mean response for the set of conditions given by

$$\mathbf{x}_0' = [1, x_{10}, x_{20}, \ldots, x_{k0}].$$

We augment the conditions on the x's by the number 1 in order to facilitate the matrix notation. Normality in the ϵ_i produces normality in the b_j and the mean and variance are still the same as indicated in Section 12.4. So is the covariance between b_i and b_j, for $i \neq j$. Hence,

$$\hat{y} = b_0 + \sum_{j=1}^{k} b_j x_{j0}$$

is likewise normally distributed and is, in fact, an unbiased estimator for the **mean response** on which we are attempting to attach a confidence interval. The variance of $\hat{y}_0$, written in matrix notation simply as a function of σ^2, $(\mathbf{X}'\mathbf{X})^{-1}$, and the condition vector $\mathbf{x}_0'$, is

$$\sigma^2_{\hat{y}_0} = \sigma^2 \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0.$$

If this expression is expanded for a given case, say $k = 2$, it is readily seen that it appropriately accounts for the variance of the b_j and the covariance of b_i and b_j, for $i \neq j$. After σ^2 is replaced by s^2 as given by Theorem 12.1, the $100(1-\alpha)\%$ confidence interval on $\mu_{Y|x_{10},x_{20},\ldots,x_{k0}}$ can be constructed from the statistic

$$T = \frac{\hat{y}_0 - \mu_{Y|x_{10},x_{20},\ldots,x_{k0}}}{s\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}},$$

which has a t-distribution with $n-k-1$ degrees of freedom.

Confidence Interval for $\mu_{Y|x_{10},x_{20},\ldots,x_{k0}}$

A $100(1-\alpha)\%$ confidence interval for the **mean response** $\mu_{Y|x_{10},x_{20},\ldots,x_{k0}}$ is

$$\hat{y}_0 - t_{\alpha/2}s\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} < \mu_{Y|x_{10},x_{20},\ldots,x_{k0}} < \hat{y}_0 + t_{\alpha/2}s\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0},$$

where $t_{\alpha/2}$ is a value of the t-distribution with $n-k-1$ degrees of freedom.

The quantity $s\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}$ is often called the **standard error of prediction** and appears on the printout of many regression computer packages.

Example 12.6: Using the data of Example 12.4, construct a 95% confidence interval for the mean response when $x_1 = 3\%$, $x_2 = 8\%$, and $x_3 = 9\%$.

***Solution*:** From the regression equation of Example 12.4, the estimated percent survival when $x_1 = 3\%$, $x_2 = 8\%$, and $x_3 = 9\%$ is

$$\hat{y} = 39.1574 + (1.0161)(3) - (1.8616)(8) - (0.3433)(9) = 24.2232.$$

Next, we find that

$$\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 = [1,3,8,9]\begin{bmatrix} 8.0648 & -0.0826 & -0.0942 & -0.7905 \\ -0.0826 & 0.0085 & 0.0017 & 0.0037 \\ -0.0942 & 0.0017 & 0.0166 & -0.0021 \\ -0.7905 & 0.0037 & -0.0021 & 0.0886 \end{bmatrix}\begin{bmatrix} 1 \\ 3 \\ 8 \\ 9 \end{bmatrix}$$
$$= 0.1267.$$

Using the mean square error, $s^2 = 4.298$ or $s = 2.073$, and Table A.4, we see that $t_{0.025} = 2.262$ for 9 degrees of freedom. Therefore, a 95% confidence interval for the mean percent survival for $x_1 = 3\%$, $x_2 = 8\%$, and $x_3 = 9\%$ is given by

$$24.2232 - (2.262)(2.073)\sqrt{0.1267} < \mu_{Y|3,8,9} < 24.2232 + (2.262)(2.073)\sqrt{0.1267},$$

or simply $22.5541 < \mu_{Y|3,8,9} < 25.8923$. ■

As in the case of simple linear regression, we need to make a clear distinction between the confidence interval on a mean response and the prediction interval on an *observed response*. The latter provides a bound within which we can say with a preselected degree of certainty that a new observed response will fall.

A prediction interval for a single predicted response y_0 is once again established by considering the difference $\hat{y}_0 - y_0$. The sampling distribution can be shown to be normal with mean

$$\mu_{\hat{y}_0 - y_0} = 0$$

and variance

$$\sigma^2_{\hat{y}_0 - y_0} = \sigma^2[1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0].$$

Thus, a $100(1-\alpha)\%$ prediction interval for a single prediction value y_0 can be constructed from the statistic

$$T = \frac{\hat{y}_0 - y_0}{s\sqrt{1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}},$$

which has a t-distribution with $n-k-1$ degrees of freedom.

Prediction Interval for y_0

A $100(1-\alpha)\%$ prediction interval for a **single response** y_0 is given by

$$\hat{y}_0 - t_{\alpha/2}s\sqrt{1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} < y_0 < \hat{y}_0 + t_{\alpha/2}s\sqrt{1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0},$$

where $t_{\alpha/2}$ is a value of the t-distribution with $n-k-1$ degrees of freedom.

Example 12.7: Using the data of Example 12.4, construct a 95% prediction interval for an individual percent survival response when $x_1 = 3\%$, $x_2 = 8\%$, and $x_3 = 9\%$.

***Solution*:** Referring to the results of Example 12.6, we find that the 95% prediction interval for the response y_0, when $x_1 = 3\%$, $x_2 = 8\%$, and $x_3 = 9\%$, is

$$24.2232 - (2.262)(2.073)\sqrt{1.1267} < y_0 < 24.2232 + (2.262)(2.073)\sqrt{1.1267},$$

which reduces to $19.2459 < y_0 < 29.2005$. Notice, as expected, that the prediction interval is considerably wider than the confidence interval for mean percent survival found in Example 12.6.

Annotated Printout for Data of Example 12.4

Figure 12.1 shows an annotated computer printout for a multiple linear regression fit to the data of Example 12.4. The package used is *SAS*.

Note the model parameter estimates, the standard errors, and the t-statistics shown in the output. The standard errors are computed from square roots of diagonal elements of $(\mathbf{X}'\mathbf{X})^{-1}s^2$. In this illustration, the variable x_3 is insignificant in the presence of x_1 and x_2 based on the t-test and the corresponding P-value of 0.5916. The terms CLM and CLI are confidence intervals on mean response and prediction limits on an individual observation, respectively. The f-test in the analysis of variance indicates that a significant amount of variability is explained. As an example of the interpretation of CLM and CLI, consider observation 10. With an observation of 25.2000 and a predicted value of 26.0676, we are 95% confident that the mean response is between 24.5024 and 27.6329, and a new observation will fall between 21.1238 and 31.0114 with probability 0.95. The R^2 value of 0.9117 implies that the model explains 91.17% of the variability in the response. More discussion about R^2 appears in Section 12.6.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	399.45437	133.15146	30.98	<.0001
Error	9	38.67640	4.29738		
Corrected Total	12	438.13077			

Root MSE	2.07301	R-Square	0.9117
Dependent Mean	29.03846	Adj R-Sq	0.8823
Coeff Var	7.13885		

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	39.15735	5.88706	6.65	<.0001
x1	1	1.01610	0.19090	5.32	0.0005
x2	1	-1.86165	0.26733	-6.96	<.0001
x3	1	-0.34326	0.61705	-0.56	0.5916

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	25.5000	27.3514	1.4152	24.1500	30.5528	21.6734	33.0294	-1.8514
2	31.2000	32.2623	0.7846	30.4875	34.0371	27.2482	37.2764	-1.0623
3	25.9000	27.3495	1.3588	24.2757	30.4234	21.7425	32.9566	-1.4495
4	38.4000	38.3096	1.2818	35.4099	41.2093	32.7960	43.8232	0.0904
5	18.4000	15.5447	1.5789	11.9730	19.1165	9.6499	21.4395	2.8553
6	26.7000	26.1081	1.0358	23.7649	28.4512	20.8658	31.3503	0.5919
7	26.4000	28.2532	0.8094	26.4222	30.0841	23.2189	33.2874	-1.8532
8	25.9000	26.2219	0.9732	24.0204	28.4233	21.0414	31.4023	-0.3219
9	32.0000	32.0882	0.7828	30.3175	33.8589	27.0755	37.1008	-0.0882
10	25.2000	26.0676	0.6919	24.5024	27.6329	21.1238	31.0114	-0.8676
11	39.7000	37.2524	1.3070	34.2957	40.2090	31.7086	42.7961	2.4476
12	35.7000	32.4879	1.4648	29.1743	35.8015	26.7459	38.2300	3.2121
13	26.5000	28.2032	0.9841	25.9771	30.4294	23.0122	33.3943	-1.7032

Figure 12.1: *SAS* printout for data in Example 12.4.

More on Analysis of Variance in Multiple Regression (Optional)

In Section 12.4, we discussed briefly the partition of the total sum of squares $\sum_{i=1}^{n}(y_i-\bar{y})^2$ into its two components, the regression model and error sums of squares (illustrated in Figure 12.1). The analysis of variance leads to a test of

$$H_0:\ \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0.$$

Rejection of the null hypothesis has an important interpretation for the scientist or engineer. (For those who are interested in more extensive treatment of the subject using matrices, it is useful to discuss the development of these sums of squares used in ANOVA.)

First, recall in Section 12.3, $\mathbf{b}$, the vector of least squares estimators, is given by

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

A partition of the **uncorrected sum of squares**

$$\mathbf{y}'\mathbf{y} = \sum_{i=1}^{n} \mathbf{y}_i^2$$

into two components is given by

$$\begin{aligned}\mathbf{y}'\mathbf{y} &= \mathbf{b}'\mathbf{X}'\mathbf{y} + (\mathbf{y}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{y}) \\ &= \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} + [\mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}].\end{aligned}$$

The second term (in brackets) on the right-hand side is simply the error sum of squares $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. The reader should see that an alternative expression for the error sum of squares is

$$SSE = \mathbf{y}'[\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y}.$$

The term $\mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is called the **regression sum of squares**. However, it is not the expression $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ used for testing the "importance" of the terms $b_1, b_2, \ldots, b_k$ but, rather,

$$\mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \sum_{i=1}^{n} \hat{y}_i^2,$$

which is a regression sum of squares uncorrected for the mean. As such, it would only be used in testing if the regression equation differs significantly from zero, that is,

$$H_0\text{: } \beta_0 = \beta_1 = \beta_2 = \cdots = \beta_k = 0.$$

In general, this is not as important as testing

$$H_0\text{: } \beta_1 = \beta_2 = \cdots = \beta_k = 0,$$

since the latter states that the mean response is a constant, not necessarily zero.

Degrees of Freedom

Thus, the partition of sums of squares and degrees of freedom reduces to

Source	Sum of Squares	d.f.
Regression	$\sum_{i=1}^{n} \hat{y}_i^2 = \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$	$k+1$
Error	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \mathbf{y}'[\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y}$	$n-(k+1)$
Total	$\sum_{i=1}^{n} y_i^2 = \mathbf{y}'\mathbf{y}$	n

Hypothesis of Interest

Now, of course, the hypotheses of interest for an ANOVA must eliminate the role of the intercept described previously. Strictly speaking, if H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$, then the estimated regression line is merely $\hat{y}_i = \bar{y}$. As a result, we are actually seeking evidence that the regression equation "varies from a constant." Thus, the total and regression sums of squares must be corrected for the mean. As a result, we have

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

In matrix notation this is simply

$$\mathbf{y}'[\mathbf{I}_n - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}']\mathbf{y} = \mathbf{y}'[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}']\mathbf{y} + \mathbf{y}'[\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y}.$$

In this expression, $\mathbf{1}$ is a vector of n ones. As a result, we are merely subtracting

$$\mathbf{y}'\mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'\mathbf{y} = \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2$$

from $\mathbf{y}'\mathbf{y}$ and from $\mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ (i.e., correcting the total and regression sums of squares for the mean).

Finally, the appropriate partitioning of sums of squares with degrees of freedom is as follows:

Source	Sum of Squares	d.f.
Regression	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \mathbf{y}'[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}]\mathbf{y}$	k
Error	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \mathbf{y}'[\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y}$	$n - (k+1)$
Total	$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \mathbf{y}'[\mathbf{I}_n - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}']\mathbf{y}$	$n - 1$

This is the ANOVA table that appears in the computer printout of Figure 12.1. The expression $\mathbf{y}'[\mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}']\mathbf{y}$ is often called the **regression sum of squares associated with the mean**, and 1 degree of freedom is allocated to it.

Exercises

12.17 For the data of Exercise 12.2 on page 450, estimate σ^2.

12.18 For the data of Exercise 12.1 on page 450, estimate σ^2.

12.19 For the data of Exercise 12.5 on page 450, estimate σ^2.

12.20 Obtain estimates of the variances and the covariance of the estimators b_1 and b_2 of Exercise 12.2 on page 450.

12.21 Referring to Exercise 12.5 on page 450, find the estimate of
(a) $\sigma^2_{b_2}$;
(b) $\text{Cov}(b_1, b_4)$.

12.22 For the model of Exercise 12.7 on page 451,

test the hypothesis that $\beta_2 = 0$ at the 0.05 level of significance against the alternative that $\beta_2 \neq 0$.

12.23 For the model of Exercise 12.2 on page 450, test the hypothesis that $\beta_1 = 0$ at the 0.05 level of significance against the alternative that $\beta_1 \neq 0$.

12.24 For the model of Exercise 12.1 on page 450, test the hypotheses that $\beta_1 = 2$ against the alternative that $\beta_1 \neq 2$. Use a P-value in your conclusion.

12.25 Using the data of Exercise 12.2 on page 450 and the estimate of σ^2 from Exercise 12.17, compute 95% confidence intervals for the predicted response and the mean response when $x_1 = 900$ and $x_2 = 1.00$.

12.26 For Exercise 12.8 on page 451, construct a 90% confidence interval for the mean compressive strength when the concentration is $x = 19.5$ and a quadratic model is used.

12.27 Using the data of Exercise 12.5 on page 450 and the estimate of σ^2 from Exercise 12.19, compute 95% confidence intervals for the predicted response and the mean response when $x_1 = 75$, $x_2 = 24$, $x_3 = 90$, and $x_4 = 98$.

12.28 Consider the following data from Exercise 12.13 on page 452.

y (wear)	x_1 (oil viscosity)	x_2 (load)
193	1.6	851
230	15.5	816
172	22.0	1058
91	43.0	1201
113	33.0	1357
125	40.0	1115

(a) Estimate σ^2 using multiple regression of y on x_1 and x_2.

(b) Compute predicted values, a 95% confidence interval for mean wear, and a 95% prediction interval for observed wear if $x_1 = 20$ and $x_2 = 1000$.

12.29 Using the data from Exercise 12.28, test the following at the 0.05 level.

(a) H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$;

(b) H_0: $\beta_2 = 0$ versus H_1: $\beta_2 \neq 0$.

(c) Do you have any reason to believe that the model in Exercise 12.28 should be changed? Why or why not?

12.30 Use the data from Exercise 12.16 on page 453.

(a) Estimate σ^2 using the multiple regression of y on x_1, x_2, and x_3,

(b) Compute a 95% prediction interval for the observed gain with the three regressors at $x_1 = 15.0$, $x_2 = 220.0$, and $x_3 = 6.0$.

12.6 Choice of a Fitted Model through Hypothesis Testing

In many regression situations, individual coefficients are of importance to the experimenter. For example, in an economics application, $\beta_1, \beta_2, \ldots$ might have some particular significance, and thus confidence intervals and tests of hypotheses on these parameters would be of interest to the economist. However, consider an industrial chemical situation in which the postulated model assumes that reaction yield is linearly dependent on reaction temperature and concentration of a certain catalyst. It is probably known that this is not the true model but an adequate approximation, so interest is likely to be not in the individual parameters but rather in the ability of the entire function to predict the true response in the range of the variables considered. Therefore, in this situation, one would put more emphasis on $\sigma^2_{\hat{Y}}$, confidence intervals on the mean response, and so forth, and likely deemphasize inferences on individual parameters.

The experimenter using regression analysis is also interested in deletion of variables when the situation dictates that, in addition to arriving at a workable prediction equation, he or she must find the "best regression" involving only variables that are useful predictors. There are a number of computer programs that sequentially arrive at the so-called best regression equation depending on certain criteria. We discuss this further in Section 12.9.

One criterion that is commonly used to illustrate the adequacy of a fitted regression model is the **coefficient of determination**, or $\boldsymbol{R^2}$.

Coefficient of Determination, or R^2

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{SSE}{SST}.$$

Note that this parallels the description of R^2 in Chapter 11. At this point the explanation might be clearer since we now focus on SSR as the **variability explained**. The quantity R^2 merely indicates what proportion of the total variation in the response Y is explained by the fitted model. Often an experimenter will report $R^2 \times 100\%$ and interpret the result as percent variation explained by the postulated model. The square root of R^2 is called the **multiple correlation coefficient** between Y and the set $x_1, x_2, \ldots, x_k$. The value of R^2 for the case in Example 12.4, indicating the proportion of variation explained by the three independent variables x_1, x_2, and x_3, is

$$R^2 = \frac{SSR}{SST} = \frac{399.45}{438.13} = 0.9117,$$

which means that 91.17% of the variation in percent survival has been explained by the linear regression model.

The regression sum of squares can be used to give some indication concerning whether or not the model is an adequate explanation of the true situation. We can test the hypothesis H_0 that the **regression is not significant** by merely forming the ratio

$$f = \frac{SSR/k}{SSE/(n-k-1)} = \frac{SSR/k}{s^2}$$

and rejecting H_0 at the α-level of significance when $f > f_\alpha(k, n-k-1)$. For the data of Example 12.4, we obtain

$$f = \frac{399.45/3}{4.298} = 30.98.$$

From the printout in Figure 12.1, the P-value is less than 0.0001. This should not be misinterpreted. Although it does indicate that the regression explained by the model is significant, this does not rule out the following possibilities:

1. The linear regression model for this set of x's is not the only model that can be used to explain the data; indeed, there may be other models with transformations on the x's that give a larger value of the F-statistic.
2. The model might have been more effective with the inclusion of other variables in addition to x_1, x_2, and x_3 or perhaps with the deletion of one or more of the variables in the model, say x_3, which has a $P = 0.5916$.

The reader should recall the discussion in Section 11.5 regarding the pitfalls in the use of R^2 as a criterion for comparing competing models. These pitfalls are certainly relevant in multiple linear regression. In fact, in its employment in multiple regression, the dangers are even more pronounced since the temptation

to overfit is so great. One should always keep in mind that $R^2 \approx 1.0$ can always be achieved at the expense of error degrees of freedom when an excess of model terms is employed. However, $R^2 = 1$, describing a model with a near perfect fit, does not always result in a model that predicts well.

The Adjusted Coefficient of Determination (R^2_{adj})

In Chapter 11, several figures displaying computer printout from both *SAS* and *MINITAB* featured a statistic called *adjusted* R^2 or adjusted coefficient of determination. Adjusted R^2 is a variation on R^2 that provides an **adjustment for degrees of freedom**. The coefficient of determination as defined on page 407 cannot decrease as terms are added to the model. In other words, R^2 does not decrease as the error degrees of freedom $n - k - 1$ are reduced, the latter result being produced by an increase in k, the number of model terms. Adjusted R^2 is computed by dividing SSE and SST by their respective degrees of freedom as follows.

Adjusted R^2

$$R^2_{adj} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}.$$

To illustrate the use of R^2_{adj}, Example 12.4 will be revisited.

How Are R^2 and R^2_{adj} Affected by Removal of x_3?

The t-test (or corresponding F-test) for x_3 suggests that a simpler model involving only x_1 and x_2 may well be an improvement. In other words, the complete model with all the regressors may be an overfitted model. It is certainly of interest to investigate R^2 and R^2_{adj} for both the full (x_1, x_2, x_3) and the reduced (x_1, x_2) models. We already know that $R^2_{full} = 0.9117$ from Figure 12.1. The SSE for the reduced model is 40.01, and thus $R^2_{reduced} = 1 - \frac{40.01}{438.13} = 0.9087$. Thus, more variability is explained with x_3 in the model. However, as we have indicated, this will occur even if the model is an overfitted model. Now, of course, R^2_{adj} is designed to provide a statistic that punishes an overfitted model, so we might expect it to favor the reduced model. Indeed, for the full model

$$R^2_{adj} = 1 - \frac{38.6764/9}{438.1308/12} = 1 - \frac{4.2974}{36.5109} = 0.8823,$$

whereas for the reduced model (deletion of x_3)

$$R^2_{adj} = 1 - \frac{40.01/10}{438.1308/12} = 1 - \frac{4.001}{36.5109} = 0.8904.$$

Thus, R^2_{adj} does indeed favor the reduced model and confirms the evidence produced by the t- and F-tests, suggesting that the reduced model is preferable to the model containing all three regressors. The reader may expect that other statistics would suggest rejection of the overfitted model. See Exercise 12.40 on page 471.

Test on an Individual Coefficient

The addition of any single variable to a regression system *will increase the regression sum of squares* and thus *reduce the error sum of squares.* Consequently, we must decide whether the increase in regression is sufficient to warrant using the variable in the model. As we might expect, the use of unimportant variables can reduce the effectiveness of the prediction equation by increasing the variance of the estimated response. We shall pursue this point further by considering the importance of x_3 in Example 12.4. Initially, we can test

$$H_0\colon\ \beta_3 = 0,$$
$$H_1\colon\ \beta_3 \neq 0$$

by using the t-distribution with 9 degrees of freedom. We have

$$t = \frac{b_3 - 0}{s\sqrt{c_{33}}} = \frac{-0.3433}{2.073\sqrt{0.0886}} = -0.556,$$

which indicates that β_3 does not differ significantly from zero, and hence we may very well feel justified in removing x_3 from the model. Suppose that we consider the regression of Y on the set (x_1, x_2), the least squares normal equations now reducing to

$$\begin{bmatrix} 13.0 & 59.43 & 81.82 \\ 59.43 & 394.7255 & 360.6621 \\ 81.82 & 360.6621 & 576.7264 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 377.50 \\ 1877.5670 \\ 2246.6610 \end{bmatrix}.$$

The estimated regression coefficients for this reduced model are

$$b_0 = 36.094, \quad b_1 = 1.031, \quad b_2 = -1.870,$$

and the resulting regression sum of squares with 2 degrees of freedom is

$$R(\beta_1, \beta_2) = 398.12.$$

Here we use the notation $R(\beta_1, \beta_2)$ to indicate the regression sum of squares of the restricted model; it should not be confused with SSR, the regression sum of squares of the original model with 3 degrees of freedom. The new error sum of squares is then

$$SST - R(\beta_1, \beta_2) = 438.13 - 398.12 = 40.01,$$

and the resulting mean square error with 10 degrees of freedom becomes

$$s^2 = \frac{40.01}{10} = 4.001.$$

Does a Single Variable *t*-Test Have an *F* Counterpart?

From Example 12.4, the amount of variation in the percent survival that is attributed to x_3, in the presence of the variables x_1 and x_2, is

$$R(\beta_3 \mid \beta_1, \beta_2) = SSR - R(\beta_1, \beta_2) = 399.45 - 398.12 = 1.33,$$

which represents a small proportion of the entire regression variation. This amount of added regression is statistically insignificant, as indicated by our previous test on β_3. An equivalent test involves the formation of the ratio

$$f = \frac{R(\beta_3 \mid \beta_1, \beta_2)}{s^2} = \frac{1.33}{4.298} = 0.309,$$

which is a value of the F-distribution with 1 and 9 degrees of freedom. Recall that the basic relationship between the t-distribution with v degrees of freedom and the F-distribution with 1 and v degrees of freedom is

$$t^2 = f(1, v),$$

and note that the f-value of 0.309 is indeed the square of the t-value of -0.56.

To generalize the concepts above, we can assess the work of an independent variable x_i in the general multiple linear regression model

$$\mu_{Y|x_1,x_2,\ldots,x_k} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

by observing the amount of regression attributed to x_i **over and above that attributed to the other variables**, that is, the regression on x_i *adjusted for the other variables.* For example, we say that x_1 is assessed by calculating

$$R(\beta_1 \mid \beta_2, \beta_3, \ldots, \beta_k) = SSR - R(\beta_2, \beta_3, \ldots, \beta_k),$$

where $R(\beta_2, \beta_3, \ldots, \beta_k)$ is the regression sum of squares with $\beta_1 x_1$ removed from the model. To test the hypothesis

$$\begin{aligned} H_0&: \ \beta_1 = 0, \\ H_1&: \ \beta_1 \neq 0, \end{aligned}$$

we compute

$$f = \frac{R(\beta_1 \mid \beta_2, \beta_3, \ldots, \beta_k)}{s^2},$$

and compare it with $f_\alpha(1, n - k - 1)$.

Partial F-Tests on Subsets of Coefficients

In a similar manner, we can test for the significance of a *set* of the variables. For example, to investigate simultaneously the importance of including x_1 and x_2 in the model, we test the hypothesis

$$\begin{aligned} H_0&: \ \beta_1 = \beta_2 = 0, \\ H_1&: \ \beta_1 \text{ and } \beta_2 \text{ are not both zero}, \end{aligned}$$

by computing

$$f = \frac{[R(\beta_1, \beta_2 \mid \beta_3, \beta_4, \ldots, \beta_k)]/2}{s^2} = \frac{[SSR - R(\beta_3, \beta_4, \ldots, \beta_k)]/2}{s^2}$$

and comparing it with $f_\alpha(2, n-k-1)$. The number of degrees of freedom associated with the numerator, in this case 2, equals the number of variables in the set being investigated.

Suppose we wish to test the hypothesis

$$H_0: \beta_2 = \beta_3 = 0,$$
$$H_1: \beta_2 \text{ and } \beta_3 \text{ are not both zero}$$

for Example 12.4. If we develop the regression model

$$y = \beta_0 + \beta_1 x_1 + \epsilon,$$

we can obtain $R(\beta_1) = SSR_{\text{reduced}} = 187.31179$. From Figure 12.1 on page 459, we have $s^2 = 4.29738$ for the full model. Hence, the f-value for testing the hypothesis is

$$f = \frac{R(\beta_2, \beta_3 \mid \beta_1)/2}{s^2} = \frac{[R(\beta_1, \beta_2, \beta_3) - R(\beta_1)]/2}{s^2} = \frac{[SSR_{\text{full}} - SSR_{\text{reduced}}]/2}{s^2}$$
$$= \frac{(399.45437 - 187.31179)/2}{4.29738} = 24.68278.$$

This implies that β_2 and β_3 are not simultaneously zero. Using statistical software such as *SAS* one can directly obtain the above result with a P-value of 0.0002. Readers should note that in statistical software package output there are P-values associated with each individual model coefficient. The null hypothesis for each is that the coefficient is zero. However, it should be noted that the insignificance of any coefficient does not necessarily imply that it does not belong in the final model. It merely suggests that it is insignificant in the presence of all other variables in the problem. The case study at the end of this chapter illustrates this further.

12.7 Special Case of Orthogonality (Optional)

Prior to our original development of the general linear regression problem, the assumption was made that the independent variables are measured without error and are often controlled by the experimenter. Quite often they occur as a result of an *elaborately designed experiment*. In fact, we can increase the effectiveness of the resulting prediction equation with the use of a suitable experimental plan.

Suppose that we once again consider the $\mathbf{X}$ matrix as defined in Section 12.3. We can rewrite it as

$$\mathbf{X} = [\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k],$$

where $\mathbf{1}$ represents a column of ones and $\mathbf{x}_j$ is a column vector representing the levels of x_j. If

$$\mathbf{x}_p'\mathbf{x}_q = \mathbf{0}, \quad \text{for } p \neq q,$$

the variables x_p and x_q are said to be *orthogonal* to each other. There are certain obvious advantages to having a completely orthogonal situation where $\mathbf{x}_p'\mathbf{x}_q = \mathbf{0}$

Chapter 3

Logit Models for Binary Data

We now turn our attention to regression models for dichotomous data, including logistic regression and probit analysis. These models are appropriate when the response takes one of only two possible values representing success and failure, or more generally the presence or absence of an attribute of interest.

3.1 Introduction to Logistic Regression

We start by introducing an example that will be used to illustrate the analysis of binary data. We then discuss the stochastic structure of the data in terms of the Bernoulli and binomial distributions, and the systematic structure in terms of the logit transformation. The result is a generalized linear model with binomial response and link logit.

3.1.1 The Contraceptive Use Data

Table 3.1, adapted from Little (1978), shows the distribution of 1607 currently married and fecund women interviewed in the Fiji Fertility Survey of 1975, classified by current age, level of education, desire for more children, and contraceptive use.

In our analysis of these data we will view current use of contraception as the response or dependent variable of interest and age, education and desire for more children as predictors. Note that the response has two categories: use and non-use. In this example all predictors are treated as categorical

TABLE 3.1: Current Use of Contraception Among Married Women by Age, Education and Desire for More Children Fiji Fertility Survey, 1975

Age	Education	Desires More Children?	Contraceptive Use		Total
			No	Yes	
<25	Lower	Yes	53	6	59
		No	10	4	14
	Upper	Yes	212	52	264
		No	50	10	60
25–29	Lower	Yes	60	14	74
		No	19	10	29
	Upper	Yes	155	54	209
		No	65	27	92
30–39	Lower	Yes	112	33	145
		No	77	80	157
	Upper	Yes	118	46	164
		No	68	78	146
40–49	Lower	Yes	35	6	41
		No	46	48	94
	Upper	Yes	8	8	16
		No	12	31	43
Total			1100	507	1607

variables, but the techniques to be studied can be applied more generally to both discrete factors and continuous variates.

The original dataset includes the date of birth of the respondent and the date of interview in month/year form, so it is possible to calculate age in single years, but we will use ten-year age groups for convenience. Similarly, the survey included information on the highest level of education attained and the number of years completed at that level, so one could calculate completed years of education, but we will work here with a simple distinction between lower primary or less and upper primary or more. Finally, desire for more children is measured as a simple dichotomy coded yes or no, and therefore is naturally a categorical variate.

The fact that we treat all predictors as discrete factors allows us to summarize the data in terms of the numbers using and not using contraception in each of sixteen different groups defined by combinations of values of the pre-

dictors. For models involving discrete factors we can obtain exactly the same results working with grouped data or with individual data, but grouping is convenient because it leads to smaller datasets. If we were to incorporate continuous predictors into the model we would need to work with the original 1607 observations. Alternatively, it might be possible to group cases with identical covariate patterns, but the resulting dataset may not be much smaller than the original one.

The basic aim of our analysis will be to describe the way in which contraceptive use varies by age, education and desire for more children. An example of the type of research question that we will consider is the extent to which the association between education and contraceptive use is affected by the fact that women with upper primary or higher education are younger and tend to prefer smaller families than women with lower primary education or less.

3.1.2 The Binomial Distribution

We consider first the case where the response y_i is binary, assuming only two values that for convenience we code as one or zero. For example, we could define

$$y_i = \begin{cases} 1 & \text{if the } i\text{-th woman is using contraception} \\ 0 & \text{otherwise.} \end{cases}$$

We view y_i as a realization of a random variable Y_i that can take the values one and zero with probabilities π_i and $1-\pi_i$, respectively. The distribution of Y_i is called a *Bernoulli* distribution with parameter π_i, and can be written in compact form as

$$\Pr\{Y_i = y_i\} = \pi_i^{y_i}(1-\pi_i)^{1-y_i}, \tag{3.1}$$

for $y_i = 0, 1$. Note that if $y_i = 1$ we obtain π_i, and if $y_i = 0$ we obtain $1-\pi_i$.

It is fairly easy to verify by direct calculation that the expected value and variance of Y_i are

$$\begin{aligned} \mathrm{E}(Y_i) &= \mu_i = \pi_i, \text{ and} \\ \mathrm{var}(Y_i) &= \sigma_i^2 = \pi_i(1-\pi_i). \end{aligned} \tag{3.2}$$

Note that the mean and variance depend on the underlying probability π_i. Any factor that affects the probability will alter not just the mean but also the variance of the observations. This suggest that a linear model that allows

the predictors to affect the mean but assumes that the variance is constant will not be adequate for the analysis of binary data.

Suppose now that the units under study can be classified according to the factors of interest into k groups in such a way that all individuals in a group have identical values of all covariates. In our example, women may be classified into 16 different groups in terms of their age, education and desire for more children. Let n_i denote the number of observations in group i, and let y_i denote the number of units who have the attribute of interest in group i. For example, let

$$y_i = \text{number of women using contraception in group } i.$$

We view y_i as a realization of a random variable Y_i that takes the values $0, 1, \ldots, n_i$. If the n_i observations in each group are *independent*, and they all have the same probability π_i of having the attribute of interest, then the distribution of Y_i is *binomial* with parameters π_i and n_i, which we write

$$Y_i \sim B(n_i, \pi_i).$$

The probability distribution function of Y_i is given by

$$\Pr\{Y_i = y_i\} = \left(\begin{array}{c} n_i \\ y_i \end{array} \right) \pi_i^{y_i} (1-\pi_i)^{n_i - y_i} \tag{3.3}$$

for $y_i = 0, 1, \ldots, n_i$. Here $\pi_i^{y_i}(1-\pi_i)^{n_i-y_i}$ is the probability of obtaining y_i successes and $n_i - y_i$ failures in some specific order, and the combinatorial coefficient is the number of ways of obtaining y_i successes in n_i trials.

The mean and variance of Y_i can be shown to be

$$\begin{aligned} \mathrm{E}(Y_i) &= \mu_i = n_i \pi_i, \text{ and} \\ \mathrm{var}(Y_i) &= \sigma_i^2 = n_i \pi_i (1-\pi_i). \end{aligned} \tag{3.4}$$

The easiest way to obtain this result is as follows. Let Y_{ij} be an indicator variable that takes the values one or zero if the j-th unit in group i is a success or a failure, respectively. Note that Y_{ij} is a Bernoulli random variable with mean and variance as given in Equation 3.2. We can write the number of successes Y_i in group i as a sum of the individual indicator variables, so $Y_i = \sum_j Y_{ij}$. The mean of Y_i is then the sum of the individual means, and by independence, its variance is the sum of the individual variances, leading to the result in Equation 3.4. Note again that the mean and variance depend

on the underlying probability π_i. Any factor that affects this probability will affect both the mean and the variance of the observations.

From a mathematical point of view the grouped data formulation given here is the most general one; it includes individual data as the special case where we have n groups of size one, so $k = n$ and $n_i = 1$ for all i. It also includes as a special case the other extreme where the underlying probability is the same for all individuals and we have a single group, with $k = 1$ and $n_1 = n$. Thus, all we need to consider in terms of estimation and testing is the binomial distribution.

From a practical point of view it is important to note that if the predictors are discrete factors and the outcomes are independent, we can use the Bernoulli distribution for the individual zero-one data or the binomial distribution for grouped data consisting of counts of successes in each group. The two approaches are equivalent, in the sense that they lead to exactly the same likelihood function and therefore the same estimates and standard errors. Working with grouped data when it is possible has the additional advantage that, depending on the size of the groups, it becomes possible to test the goodness of fit of the model. In terms of our example we can work with 16 groups of women (or fewer when we ignore some of the predictors) and obtain exactly the same estimates as we would if we worked with the 1607 individuals.

In Appendix B we show that the binomial distribution belongs to Nelder and Wedderburn's (1972) exponential family, so it fits in our general theoretical framework.

3.1.3 The Logit Transformation

The next step in defining a model for our data concerns the systematic structure. We would like to have the probabilities π_i depend on a vector of observed covariates $\mathbf{x}_i$. The simplest idea would be to let π_i be a linear function of the covariates, say

$$\pi_i = \mathbf{x}_i'\boldsymbol{\beta}, \tag{3.5}$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients. Model 3.5 is sometimes called the *linear probability model*. This model is often estimated from individual data using ordinary least squares (OLS).

One problem with this model is that the probability π_i on the left-hand-side has to be between zero and one, but the linear predictor $\mathbf{x}_i'\boldsymbol{\beta}$ on the right-hand-side can take any real value, so there is no guarantee that the

predicted values will be in the correct range unless complex restrictions are imposed on the coefficients.

A simple solution to this problem is to *transform* the probability to remove the range restrictions, and model the transformation as a linear function of the covariates. We do this in two steps.

First, we move from the probability π_i to the *odds*

$$\text{odds}_i = \frac{\pi_i}{1-\pi_i},$$

defined as the ratio of the probability to its complement, or the ratio of favorable to unfavorable cases. If the probability of an event is a half, the odds are one-to-one or even. If the probability is 1/3, the odds are one-to-two. If the probability is very small, the odds are said to be long. In some contexts the language of odds is more natural than the language of probabilities. In gambling, for example, odds of $1:k$ indicate that the fair payoff for a stake of one is k. The key from our point of view is that the languages are equivalent, i.e. one can easily be translated into the other, but odds can take any positive value and therefore have no ceiling restriction.

Second, we take logarithms, calculating the *logit* or log-odds

$$\eta_i = \text{logit}(\pi_i) = \log \frac{\pi_i}{1-\pi_i}, \tag{3.6}$$

which has the effect of removing the floor restriction. To see this point note that as the probability goes down to zero the odds approach zero and the logit approaches $-\infty$. At the other extreme, as the probability approaches one the odds approach $+\infty$ and so does the logit. Thus, logits map probabilities from the range $(0, 1)$ to the entire real line. Note that if the probability is 1/2 the odds are even and the logit is zero. Negative logits represent probabilities below one half and positive logits correspond to probabilities above one half. Figure 3.1 illustrates the logit transformation.

Logits may also be defined in terms of the binomial mean $\mu_i = n_i\pi_i$ as the log of the ratio of expected successes μ_i to expected failures $n_i - \mu_i$. The result is exactly the same because the binomial denominator n_i cancels out when calculating the odds.

In the contraceptive use data there are 507 users of contraception among 1607 women, so we estimate the probability as $507/1607 = 0.316$. The odds are 507/1100 or 0.461 to one, so non-users outnumber users roughly two to one. The logit is $\log(0.461) = -0.775$.

The logit transformation is one-to-one. The inverse transformation is sometimes called the *antilogit*, and allows us to go back from logits to prob-

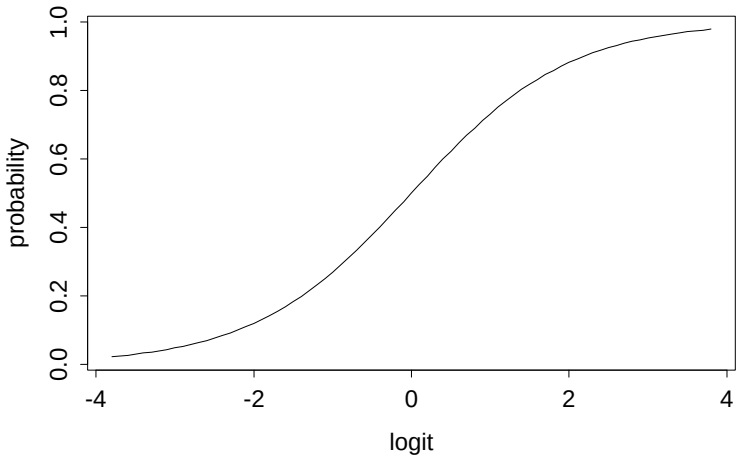


FIGURE 3.1: The Logit Transformation

abilities. Solving for π_i in Equation 3.6 gives

$$\pi_i = \text{logit}^{-1}(\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}}. \tag{3.7}$$

In the contraceptive use data the estimated logit was -0.775. Exponentiating this value we obtain odds of $\exp(-0.775) = 0.461$ and from this we obtain a probability of $0.461/(1 + 0.461) = 0.316$.

We are now in a position to define the logistic regression model, by assuming that the *logit* of the probability π_i, rather than the probability itself, follows a linear model.

3.1.4 The Logistic Regression Model

Suppose that we have k independent observations $y_1, \ldots, y_k$, and that the i-th observation can be treated as a realization of a random variable Y_i. We assume that Y_i has a binomial distribution

$$Y_i \sim B(n_i, \pi_i) \tag{3.8}$$

with binomial denominator n_i and probability π_i. With individual data $n_i = 1$ for all i. This defines the stochastic structure of the model.

Suppose further that the *logit* of the underlying probability π_i is a linear function of the predictors

$$\text{logit}(\pi_i) = \mathbf{x}_i'\boldsymbol{\beta}, \tag{3.9}$$

where $\mathbf{x}_i$ is a vector of covariates and $\boldsymbol{\beta}$ is a vector of regression coefficients. This defines the systematic structure of the model.

The model defined in Equations 3.8 and 3.9 is a generalized linear model with binomial response and link logit. Note, incidentally, that it is more natural to consider the distribution of the response Y_i than the distribution of the implied error term $Y_i - \mu_i$.

The regression coefficients $\boldsymbol{\beta}$ can be interpreted along the same lines as in linear models, bearing in mind that the left-hand-side is a logit rather than a mean. Thus, β_j represents the change in the *logit* of the probability associated with a unit change in the j-th predictor holding all other predictors constant. While expressing results in the logit scale will be unfamiliar at first, it has the advantage that the model is rather simple in this particular scale.

Exponentiating Equation 3.9 we find that the odds for the i-th unit are given by

$$\frac{\pi_i}{1-\pi_i} = \exp\{\mathbf{x}_i'\boldsymbol{\beta}\}. \tag{3.10}$$

This expression defines a multiplicative model for the odds. For example if we were to change the j-th predictor by one unit while holding all other variables constant, we would multiply the odds by $\exp\{\beta_j\}$. To see this point suppose the linear predictor is $\mathbf{x}_i'\boldsymbol{\beta}$ and we increase x_j by one, to obtain $\mathbf{x}_i'\boldsymbol{\beta} + \beta_j$. Exponentiating we get $\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}$ times $\exp\{\beta_j\}$. Thus, the exponentiated coefficient $\exp\{\beta_j\}$ represents an *odds ratio*. Translating the results into multiplicative effects on the odds, or odds ratios, is often helpful, because we can deal with a more familiar scale while retaining a relatively simple model.

Solving for the probability π_i in the logit model in Equation 3.9 gives the more complicated model

$$\pi_i = \frac{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}{1+\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}. \tag{3.11}$$

While the left-hand-side is in the familiar probability scale, the right-hand-side is a non-linear function of the predictors, and there is no simple way to express the effect on the probability of increasing a predictor by one unit while holding the other variables constant. We can obtain an approximate answer by taking derivatives with respect to x_j, which of course makes sense only for continuous predictors. Using the quotient rule we get

$$\frac{\mathrm{d}\pi_i}{\mathrm{d}x_{ij}} = \beta_j \pi_i (1-\pi_i).$$

Thus, the effect of the j-th predictor on the probability π_i depends on the coefficient β_j *and* the value of the probability. Analysts sometimes evaluate this product setting π_i to the sample mean (the proportion of cases with the attribute of interest in the sample). The result approximates the effect of the covariate near the mean of the response.

In the examples that follow we will emphasize working directly in the logit scale, but we will often translate effects into odds ratios to help in interpretation.

Before we leave this topic it may be worth considering the linear probability model of Equation 3.5 one more time. In addition to the fact that the linear predictor $\mathbf{x}_i'\boldsymbol{\beta}$ may yield values outside the $(0, 1)$ range, one should consider whether it is reasonable to assume linear effects on a probability scale that is subject to floor and ceiling effects. An incentive, for example, may increase the probability of taking an action by ten percentage points when the probability is a half, but couldn't possibly have that effect if the baseline probability was 0.95. This suggests that the assumption of a linear effect across the board may not be reasonable.

In contrast, suppose the effect of the incentive is 0.4 in the logit scale, which is equivalent to approximately a 50% increase in the odds of taking the action. If the original probability is a half the logit is zero, and adding 0.4 to the logit gives a probability of 0.6, so the effect is ten percentage points, just as before. If the original probability is 0.95, however, the logit is almost three, and adding 0.4 in the logit scale gives a probability of 0.97, an effect of just two percentage points. An effect that is constant in the logit scale translates into varying effects on the probability scale, adjusting automatically as one approaches the floor of zero or the ceiling of one. This feature of the transformation is clearly seen from Figure 3.1.

3.2 Estimation and Hypothesis Testing

The logistic regression model just developed is a generalized linear model with binomial errors and link logit. We can therefore rely on the general theory developed in Appendix B to obtain estimates of the parameters and to test hypotheses. In this section we summarize the most important results needed in the applications.

3.2.1 Maximum Likelihood Estimation

Although you will probably use a statistical package to compute the estimates, here is a brief description of the underlying procedure. The likelihood

function for n independent binomial observations is a product of densities given by Equation 3.3. Taking logs we find that, except for a constant involving the combinatorial terms, the log-likelihood function is

$$\log L(\boldsymbol{\beta}) = \sum \{y_i \log(\pi_i) + (n_i - y_i) \log(1 - \pi_i)\},$$

where π_i depends on the covariates $\mathbf{x}_i$ and a vector of p parameters $\boldsymbol{\beta}$ through the logit transformation of Equation 3.9.

At this point we could take first and expected second derivatives to obtain the score and information matrix and develop a Fisher scoring procedure for maximizing the log-likelihood. As shown in Appendix B, the procedure is equivalent to iteratively re-weighted least squares (IRLS). Given a current estimate $\hat{\boldsymbol{\beta}}$ of the parameters, we calculate the linear predictor $\hat{\boldsymbol{\eta}} = \mathbf{x}_i'\hat{\boldsymbol{\beta}}$ and the fitted values $\hat{\boldsymbol{\mu}} = \text{logit}^{-1}(\boldsymbol{\eta})$. With these values we calculate the working dependent variable $\mathbf{z}$, which has elements

$$z_i = \hat{\eta}_i + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i(n_i - \hat{\mu}_i)} n_i,$$

where n_i are the binomial denominators. We then regress $\mathbf{z}$ on the covariates calculating the weighted least squares estimate

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{z},$$

where $\mathbf{W}$ is a diagonal matrix of weights with entries

$$w_{ii} = \hat{\mu}_i(n_i - \hat{\mu}_i)/n_i.$$

(You may be interested to know that the weight is inversely proportional to the estimated variance of the working dependent variable.) The resulting estimate of $\boldsymbol{\beta}$ is used to obtain improved fitted values and the procedure is iterated to convergence.

Suitable initial values can be obtained by applying the link to the data. To avoid problems with counts of 0 or n_i (which is always the case with individual zero-one data), we calculate empirical logits adding $1/2$ to both the numerator and denominator, i.e. we calculate

$$z_i = \log \frac{y_i + 1/2}{n_i - y_i + 1/2},$$

and then regress this quantity on $\mathbf{x}_i$ to obtain an initial estimate of $\boldsymbol{\beta}$.

The resulting estimate is consistent and its large-sample variance is given by

$$\text{var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} \tag{3.12}$$

where $\mathbf{W}$ is the matrix of weights evaluated in the last iteration.

Alternatives to maximum likelihood estimation include weighted least squares, which can be used with grouped data, and a method that minimizes Pearson's chi-squared statistic, which can be used with both grouped and individual data. We will not consider these alternatives further.

3.2.2 Goodness of Fit Statistics

Suppose we have just fitted a model and want to assess how well it fits the data. A measure of discrepancy between observed and fitted values is the *deviance* statistic, which is given by

$$D = 2\sum\{y_i \log(\frac{y_i}{\hat{\mu}_i}) + (n_i - y_i)\log(\frac{n_i - y_i}{n_i - \hat{\mu}_i})\}, \tag{3.13}$$

where y_i is the observed and $\hat{\mu}_i$ is the fitted value for the i-th observation. Note that this statistic is twice a sum of 'observed times log of observed over expected', where the sum is over both successes and failures (i.e. we compare both y_i and $n_i - y_i$ with their expected values). In a perfect fit the ratio observed over expected is one and its logarithm is zero, so the deviance is zero.

In Appendix B we show that this statistic may be constructed as a likelihood ratio test that compares the model of interest with a saturated model that has one parameter for each observation.

With grouped data, the distribution of the deviance statistic as the group sizes $n_i \to \infty$ for all i, converges to a chi-squared distribution with $n - p$ d.f., where n is the number of *groups* and p is the number of parameters in the model, including the constant. Thus, for reasonably large groups, the deviance provides a goodness of fit test for the model. With individual data the distribution of the deviance does not converge to a chi-squared (or any other known) distribution, and cannot be used as a goodness of fit test. We will, however, consider other diagnostic tools that can be used with individual data.

An alternative measure of goodness of fit is *Pearson's chi-squared statistic*, which for binomial data can be written as

$$\chi^2_P = \sum_i \frac{n_i(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i(n_i - \hat{\mu}_i)}. \tag{3.14}$$

Note that each term in the sum is the squared difference between observed and fitted values y_i and $\hat{\mu}_i$, divided by the variance of y_i, which is $\mu_i(n_i -$

$\mu_i)/n_i$, estimated using $\hat{\mu}_i$ for μ_i. This statistic can also be derived as a sum of 'observed minus expected squared over expected', where the sum is over both successes and failures.

With grouped data Pearson's statistic has approximately in large samples a chi-squared distribution with $n-p$ d.f., and is asymptotically equivalent to the deviance or likelihood-ratio chi-squared statistic. The statistic can not be used as a goodness of fit test with individual data, but provides a basis for calculating residuals, as we shall see when we discuss logistic regression diagnostics.

3.2.3 Tests of Hypotheses

Let us consider the problem of testing hypotheses in logit models. As usual, we can calculate Wald tests based on the large-sample distribution of the m.l.e., which is approximately normal with mean $\boldsymbol{\beta}$ and variance-covariance matrix as given in Equation 3.12.

In particular, we can test the hypothesis

$$H_0 : \beta_j = 0$$

concerning the significance of a single coefficient by calculating the ratio of the estimate to its standard error

$$z = \frac{\hat{\beta}_j}{\sqrt{\hat{\text{var}}(\hat{\beta}_j)}}.$$

This statistic has approximately a standard normal distribution in large samples. Alternatively, we can treat the square of this statistic as approximately a chi-squared with one d.f.

The Wald test can be use to calculate a confidence interval for β_j. We can assert with $100(1-\alpha)\%$ confidence that the true parameter lies in the interval with boundaries

$$\hat{\beta}_j \pm z_{1-\alpha/2}\sqrt{\hat{\text{var}}(\hat{\beta}_j)},$$

where $z_{1-\alpha/2}$ is the normal critical value for a two-sided test of size α. Confidence intervals for effects in the logit scale can be translated into confidence intervals for odds ratios by exponentiating the boundaries.

The Wald test can be applied to tests hypotheses concerning several coefficients by calculating the usual quadratic form. This test can also be inverted to obtain confidence regions for vector-value parameters, but we will not consider this extension.

For more general problems we consider the likelihood ratio test. A key to construct these tests is the deviance statistic introduced in the previous subsection. In a nutshell, the likelihood ratio test to compare two nested models is based on the *difference* between their deviances.

To fix ideas, consider partitioning the model matrix and the vector of coefficients into two components

$$\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2) \qquad \text{and} \qquad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}$$

with p_1 and p_2 elements, respectively. Consider testing the hypothesis

$$H_0 : \boldsymbol{\beta}_2 = \mathbf{0},$$

that the variables in $\mathbf{X}_2$ have no effect on the response, i.e. the joint significance of the coefficients in $\boldsymbol{\beta}_2$.

Let $D(\mathbf{X}_1)$ denote the deviance of a model that includes only the variables in $\mathbf{X}_1$ and let $D(\mathbf{X}_1 + \mathbf{X}_2)$ denote the deviance of a model that includes all variables in $\mathbf{X}$. Then the difference

$$\chi^2 = D(\mathbf{X}_1) - D(\mathbf{X}_1 + \mathbf{X}_2)$$

has approximately in large samples a chi-squared distribution with p_2 d.f. Note that p_2 is the difference in the number of parameters between the two models being compared.

The deviance plays a role similar to the residual sum of squares. In fact, in Appendix B we show that in models for normally distributed data the deviance *is* the residual sum of squares. Likelihood ratio tests in generalized linear models are based on scaled deviances, obtained by dividing the deviance by a scale factor. In linear models the scale factor was σ^2, and we had to divide the RSS's (or their difference) by an estimate of σ^2 in order to calculate the test criterion. With binomial data the scale factor is one, and there is no need to scale the deviances.

The Pearson chi-squared statistic in the previous subsection, while providing an alternative goodness of fit test for grouped data, cannot be used in general to compare nested models. In particular, differences in deviances have chi-squared distributions but differences in Pearson chi-squared statistics do not. This is the main reason why in statistical modelling we use the deviance or likelihood ratio chi-squared statistic rather than the more traditional Pearson chi-squared of elementary statistics.

3.3 The Comparison of Two Groups

We start our applications of logit regression with the simplest possible example: a two by two table. We study a binary outcome in two groups, and introduce the odds ratio and the logit analogue of the two-sample t test.

3.3.1 A 2-by-2 Table

We will use the contraceptive use data classified by desire for more children, as summarized in Table 3.2

TABLE 3.2: Contraceptive Use by Desire for More Children

Desires	Using	Not Using	All
i	y_i	$n_i - y_i$	n_i
Yes	219	753	972
No	288	347	635
All	507	1100	1607

We treat the counts of users y_i as realizations of independent random variables Y_i having binomial distributions $B(n_i, \pi_i)$ for $i = 1, 2$, and consider models for the logits of the probabilities.

3.3.2 Testing Homogeneity

There are only two possible models we can entertain for these data. The first one is the *null* model. This model assumes homogeneity, so the two groups have the same probability and therefore the same logit

$$\text{logit}(\pi_i) = \eta.$$

The m.l.e. of the common logit is -0.775, which happens to be the logit of the sample proportion $507/1607 = 0.316$. The standard error of the estimate is 0.054. This value can be used to obtain an approximate 95% confidence limit for the logit with boundaries $(-0.880, -0.669)$. Calculating the antilogit of these values, we obtain a 95% confidence interval for the overall probability of using contraception of $(0.293, 0.339)$.

The deviance for the null model happens to be 91.7 on one d.f. (two groups minus one parameter). This value is highly significant, indicating that this model does not fit the data, i.e. the two groups classified by desire for more children do not have the same probability of using contraception.

The value of the deviance is easily verified by hand. The estimated probability of 0.316, applied to the sample sizes in Table 3.2, leads us to expect 306.7 and 200.3 users of contraception in the two groups, and therefore 665.3 and 434.7 non-users . Comparing the observed and expected numbers of users and non-users in the two groups using Equation 3.13 gives 91.7.

You can also compare the observed and expected frequencies using Pearson's chi-squared statistic from Equation 3.14. The result is 92.6 on one d.f., and provides an alternative test of the goodness of fit of the null model.

3.3.3 The Odds Ratio

The other model that we can entertain for the two-by-two table is the *one-factor* model, where we write

$$\text{logit}(\pi_i) = \eta + \alpha_i,$$

where η is an overall logit and α_i is the effect of group i on the logit. Just as in the one-way anova model, we need to introduce a restriction to identify this model. We use the reference cell method, and set $\alpha_1 = 0$. The model can then be written

$$\text{logit}(\pi_i) = \begin{cases} \eta & i = 1 \\ \eta + \alpha_2 & i = 2 \end{cases}$$

so that η becomes the logit of the reference cell, and α_2 is the effect of level two of the factor compared to level one, or more simply the difference in logits between level two and the reference cell. Table 3.3 shows parameter estimates and standard errors for this model.

TABLE 3.3: Parameter Estimates for Logit Model of Contraceptive Use by Desire for More Children

Parameter	Symbol	Estimate	Std. Error	z-ratio
Constant	η	-1.235	0.077	-16.09
Desire	α_2	1.049	0.111	9.48

The estimate of η is, as you might expect, the logit of the observed proportion using contraception among women who desire more children, $\text{logit}(219/972) = -1.235$. The estimate of α_2 is the difference between the logits of the two groups, $\text{logit}(288/635) - \text{logit}(219/972) = 1.049$.

Exponentiating the additive logit model we obtain a multiplicative model for the odds:

$$\frac{\pi_i}{1-\pi_i} = \begin{cases} e^{\eta} & i = 1 \\ e^{\eta}e^{\alpha_2} & i = 2 \end{cases}$$

so that e^{η} becomes the odds for the reference cell and e^{α_2} is the ratio of the odds for level 2 of the factor to the odds for the reference cell. Not surprisingly, e^{α_2} is called the *odds ratio*.

In our example, the effect of 1.049 in the logit scale translates into an odds ratio of 2.85. Thus, the odds of using contraception among women who want no more children are nearly three times as high as the odds for women who desire more children.

From the estimated logit effect of 1.049 and its standard error we can calculate a 95% confidence interval with boundaries $(0.831, 1.267)$. Exponentiating these boundaries we obtain a 95% confidence interval for the odds ratio of $(2.30, 3.55)$. Thus, we conclude with 95% confidence that the odds of using contraception among women who want no more children are between two and three-and-a-half times the corresponding odds for women who want more children.

The estimate of the odds ratio can be calculated directly as the cross-product of the frequencies in the two-by-two table. If we let f_{ij} denote the frequency in cell i, j then the estimated odds ratio is

$$\frac{f_{11}f_{22}}{f_{12}f_{21}}.$$

The deviance of this model is zero, because the model is saturated: it has two parameters to represent two groups, so it has to do a perfect job. The reduction in deviance of 91.7 from the null model down to zero can be interpreted as a test of

$$H_0 : \alpha_2 = 0,$$

the significance of the effect of desire for more children.

An alternative test of this effect is obtained from the m.l.e of 1.049 and its standard error of 0.111, and gives a z-ratio of 9.47. Squaring this value we obtain a chi-squared of 89.8 on one d.f. Note that the Wald test is similar, but not identical, to the likelihood ratio test. Recall that in linear models the two tests were identical. In logit models they are only asymptotically equivalent.

The logit of the observed proportion $p_i = y_i/n_i$ has large-sample variance

$$\text{var}(\text{logit}(p_i)) = \frac{1}{\mu_i} + \frac{1}{n_i - \mu_i},$$

which can be estimated using y_i to estimate μ_i for $i = 1, 2$. Since the two groups are independent samples, the variance of the difference in logits is the sum of the individual variances. You may use these results to verify the Wald test given above.

3.3.4 The Conventional Analysis

It might be instructive to compare the results obtained here with the conventional analysis of this type of data, which focuses on the sample proportions and their difference. In our example, the proportions using contraception are 0.225 among women who want another child and 0.453 among those who do not. The difference of 0.228 has a standard error of 0.024 (calculated using the pooled estimate of the proportion). The corresponding z-ratio is 9.62 and is equivalent to a chi-squared of 92.6 on one d.f.

Note that the result coincides with the Pearson chi-squared statistic testing the goodness of fit of the null model. In fact, Pearson's chi-squared and the conventional test for equality of two proportions are one and the same.

In the case of two samples it is debatable whether the group effect is best measured in terms of a difference in probabilities, the odds-ratio, or even some other measures such as the relative difference proposed by Sheps (1961). For arguments on all sides of this issue see Fleiss (1973).

3.4 The Comparison of Several Groups

Let us take a more general look at logistic regression models with a single predictor by considering the comparison of k groups. This will help us illustrate the logit analogues of one-way analysis of variance and simple linear regression models.

3.4.1 A k-by-Two Table

Consider a cross-tabulation of contraceptive use by age, as summarized in Table 3.4. The structure of the data is the same as in the previous section, except that we now have four groups rather than two.

The analysis of this table proceeds along the same lines as in the two-by-two case. The null model yields exactly the same estimate of the overall logit and its standard error as before. The deviance, however, is now 79.2 on three d.f. This value is highly significant, indicating that the assumption of a common probability of using contraception for the four age groups is not tenable.

TABLE 3.4: Contraceptive Use by Age

Age i	Using y_i	Not Using $n_i - y_i$	Total n_i
<25	72	325	397
25–29	105	299	404
30–39	237	375	612
40–49	93	101	194
Total	507	1100	1607

3.4.2 The One-Factor Model

Consider now a one-factor model, where we allow each group or level of the discrete factor to have its own logit. We write the model as

$$\text{logit}(\pi_i) = \eta + \alpha_i.$$

To avoid redundancy we adopt the reference cell method and set $\alpha_1 = 0$, as before. Then η is the logit of the reference group, and α_i measures the difference in logits between level i of the factor and the reference level. This model is exactly analogous to an analysis of variance model. The model matrix $\mathbf{X}$ consists of a column of ones representing the constant and $k-1$ columns of dummy variables representing levels two to k of the factor.

Fitting this model to Table 3.4 leads to the parameter estimates and standard errors in Table 3.5. The deviance for this model is of course zero because the model is saturated: it uses four parameters to model four groups.

TABLE 3.5: Estimates and Standard Errors for Logit Model of Contraceptive Use by Age in Groups

Parameter		Symbol	Estimate	Std. Error	z-ratio
Constant		η	−1.507	0.130	−11.57
Age	25–29	α_2	0.461	0.173	2.67
	30–39	α_3	1.048	0.154	6.79
	40–49	α_4	1.425	0.194	7.35

The baseline logit of −1.51 for women under age 25 corresponds to odds of 0.22. Exponentiating the age coefficients we obtain odds ratios of 1.59, 2.85 and 4.16. Thus, the odds of using contraception increase by 59% and

185% as we move to ages 25–29 and 30–39, and are quadrupled for ages 40–49, all compared to women under age 25.

All of these estimates can be obtained directly from the frequencies in Table 3.4 in terms of the logits of the observed proportions. For example the constant is $\text{logit}(72/397) = -1.507$, and the effect for women 25–29 is logit(105/404) minus the constant.

To test the hypothesis of no age effects we can compare this model with the null model. Since the present model is saturated, the difference in deviances is exactly the same as the deviance of the null model, which was 79.2 on three d.f. and is highly significant. An alternative test of

$$H_0 : \alpha_2 = \alpha_3 = \alpha_4 = 0$$

is based on the estimates and their variance-covariance matrix. Let $\boldsymbol{\alpha} = (\alpha_2, \alpha_3, \alpha_4)'$. Then

$$\hat{\boldsymbol{\alpha}} = \begin{pmatrix} 0.461 \\ 1.048 \\ 1.425 \end{pmatrix} \qquad \text{and} \qquad \hat{\text{var}}(\hat{\boldsymbol{\alpha}}) = \begin{pmatrix} 0.030 & 0.017 & 0.017 \\ 0.017 & 0.024 & 0.017 \\ 0.017 & 0.017 & 0.038 \end{pmatrix},$$

and the Wald statistic is

$$W = \hat{\boldsymbol{\alpha}}' \, \hat{\text{var}}^{-1}(\hat{\boldsymbol{\alpha}}) \, \hat{\boldsymbol{\alpha}} = 74.4$$

on three d.f. Again, the Wald test gives results similar to the likelihood ratio test.

3.4.3 A One-Variate Model

Note that the estimated logits in Table 3.5 (and therefore the odds and probabilities) increase monotonically with age. In fact, the logits seem to increase by approximately the same amount as we move from one age group to the next. This suggests that the effect of age may actually be linear in the logit scale.

To explore this idea we treat age as a variate rather than a factor. A thorough exploration would use the individual data with age in single years (or equivalently, a 35 by two table of contraceptive use by age in single years from 15 to 49). However, we can obtain a quick idea of whether the model would be adequate by keeping age grouped into four categories but representing these by the *mid-points* of the age groups. We therefore consider a model analogous to simple linear regression, where

$$\text{logit}(\pi_i) = \alpha + \beta x_i,$$

where x_i takes the values 20, 27.5, 35 and 45, respectively, for the four age groups. This model fits into our general framework, and corresponds to the special case where the model matrix $\mathbf{X}$ has two columns, a column of ones representing the constant and a column with the mid-points of the age groups, representing the linear effect of age.

Fitting this model gives a deviance of 2.40 on two d.f. , which indicates a very good fit. The parameter estimates and standard errors are shown in Table 3.6. Incidentally, there is no explicit formula for the estimates of the constant and slope in this model, so we must rely on iterative procedures to obtain the estimates.

TABLE 3.6: Estimates and Standard Errors for Logit Model of Contraceptive Use with a Linear Effect of Age

Parameter	Symbol	Estimate	Std. Error	z-ratio
Constant	α	-2.673	0.233	-11.46
Age (linear)	β	0.061	0.007	8.54

The slope indicates that the logit of the probability of using contraception increases 0.061 for every year of age. Exponentiating this value we note that the odds of using contraception are multiplied by 1.063—that is, increase 6.3%—for every year of age. Note, by the way, that $e^{\beta} \approx 1 + \beta$ for small $|\beta|$. Thus, when the logit coefficient is small in magnitude, 100β provides a quick approximation to the percent change in the odds associated with a unit change in the predictor. In this example the effect is 6.3% and the approximation is 6.1%.

To test the significance of the slope we can use the Wald test, which gives a z statistic of 8.54 or equivalently a chi-squared of 73.9 on one d.f. Alternatively, we can construct a likelihood ratio test by comparing this model with the null model. The difference in deviances is 76.8 on one d.f. Comparing these results with those in the previous subsection shows that we have captured most of the age effect using a single degree of freedom.

Adding the estimated constant to the product of the slope by the mid-points of the age groups gives estimated logits at each age, and these may be compared with the logits of the observed proportions using contraception. The results of this exercise appear in Figure 3.2. The visual impression of the graph confirms that the fit is quite good. In this example the assumption of linear effects on the logit scale leads to a simple and parsimonious model. It would probably be worthwhile to re-estimate this model using the individual

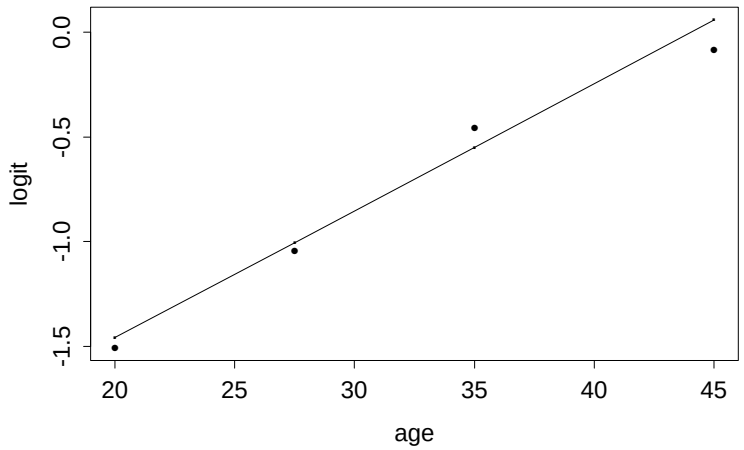


FIGURE 3.2: Observed and Fitted Logits for Model of Contraceptive Use with a Linear Effect of Age

ages.

3.5 Models With Two Predictors

We now consider models involving two predictors, and discuss the binary data analogues of two-way analysis of variance, multiple regression with dummy variables, and analysis of covariance models. An important element of the discussion concerns the key concepts of main effects and interactions.

3.5.1 Age and Preferences

Consider the distribution of contraceptive use by age and desire for more children, as summarized in Table 3.7. We have a total of eight groups, which will be indexed by a pair of subscripts i, j, with $i = 1, 2, 3, 4$ referring to the four age groups and $j = 1, 2$ denoting the two categories of desire for more children. We let y_{ij} denote the number of women using contraception and n_{ij} the total number of women in age group i and category j of desire for more children.

We now analyze these data under the usual assumption of a binomial error structure, so the y_{ij} are viewed as realizations of independent random variables $Y_{ij} \sim B(n_{ij}, \pi_{ij})$.

TABLE 3.7: Contraceptive Use by Age and Desire for More Children

Age i	Desires j	Using y_{ij}	Not Using $n_{ij} - y_{ij}$	All n_{ij}
<25	Yes	58	265	323
	No	14	60	74
25–29	Yes	68	215	283
	No	37	84	121
30–39	Yes	79	230	309
	No	158	145	303
40–49	Yes	14	43	57
	No	79	58	137
Total		507	1100	1607

3.5.2 The Deviance Table

There are five basic models of interest for the systematic structure of these data, ranging from the null to the saturated model. These models are listed in Table 3.8, which includes the name of the model, a descriptive notation, the formula for the linear predictor, the deviance or goodness of fit likelihood ratio chi-squared statistic, and the degrees of freedom.

Note first that the null model does not fit the data: the deviance of 145.7 on 7 d.f. is much greater than 14.1, the 95-th percentile of the chi-squared distribution with 7 d.f. This result is not surprising, since we already knew that contraceptive use depends on desire for more children and varies by age.

TABLE 3.8: Deviance Table for Models of Contraceptive Use by Age (Grouped) and Desire for More Children

Model	Notation	$\text{logit}(\pi_{ij})$	Deviance	d.f.
Null	ϕ	η	145.7	7
Age	A	$\eta + \alpha_i$	66.5	4
Desire	D	$\eta + \beta_j$	54.0	6
Additive	$A + D$	$\eta + \alpha_i + \beta_j$	16.8	3
Saturated	AD	$\eta + \alpha_i + \beta_j + (\alpha\beta)_{ij}$	0	0

Introducing age in the model reduces the deviance to 66.5 on four d.f. The difference in deviances between the null model and the age model provides a test for the *gross* effect of age. The difference is 79.2 on three d.f.,

and is highly significant. This value is exactly the same that we obtained in the previous section, when we tested for an age effect using the data classified by age only. Moreover, the estimated age effects based on fitting the age model to the three-way classification in Table 3.7 would be exactly the same as those estimated in the previous section, and have the property of reproducing exactly the proportions using contraception in each age group.

This equivalence illustrate an important property of binomial models. All information concerning the gross effect of age on contraceptive use is contained in the marginal distribution of contraceptive use by age. We can work with the data classified by age only, by age and desire for more children, by age, education and desire for more children, or even with the individual data. In all cases the estimated effects, standard errors, and likelihood ratio tests based on differences between deviances will be the same.

The deviances themselves will vary, however, because they depend on the context. In the previous section the deviance of the age model was zero, because treating age as a factor reproduces exactly the proportions using contraception by age. In this section the deviance of the age model is 66.5 on four d.f. and is highly significant, because the age model does not reproduce well the table of contraceptive use by both age and preferences. In both cases, however, the difference in deviances between the age model and the null model is 79.2 on three d.f.

The next model in Table 3.8 is the model with a main effect of desire for more children, and has a deviance of 54.0 on six d.f. Comparison of this value with the deviance of the null model shows a gain of 97.1 at the expense of one d.f., indicating a highly significant *gross* effect of desire for more children. This is, of course, the same result that we obtained in Section 3.3, when we first looked at contraceptive use by desire for more children. Note also that this model does not fit the data, as it own deviance is highly significant.

The fact that the effect of desire for more children has a chi-squared statistic of 91.7 with only one d.f., whereas age gives 79.2 on three d.f., suggests that desire for more children has a stronger effect on contraceptive use than age does. Note, however, that the comparison is informal; the models are not nested, and therefore we cannot construct a significance test from their deviances.

3.5.3 The Additive Model

Consider now the two-factor additive model, denoted $A + D$ in Table 3.8. In this model the logit of the probability of using contraception in age group i

and in category j of desire for more children is

$$\text{logit}(\pi_{ij}) = \eta + \alpha_i + \beta_j,$$

where η is a constant, the α_i are age effects and the β_j are effects of desire for more children. To avoid redundant parameters we adopt the reference cell method and set $\alpha_1 = \beta_1 = 0$. The parameters may then be interpreted as follows:

- η is the logit of the probability of using contraception for women under 25 who want more children, who serve as the reference cell,
- α_i for $i = 2, 3, 4$ represents the *net* effect of ages 25–29, 30–39 and 40–49, compared to women under age 25 in the same category of desire for more children,
- β_2 represents the *net* effect of wanting no more children, compared to women who want more children in the same age group.

The model is additive in the logit scale, in the usual sense that the effect of one variable does not depend on the value of the other. For example, the effect of desiring no more children is β_2 in all four age groups. (This assumption must obviously be tested, and we shall see that it is not consistent with the data.)

The deviance of the additive model is 16.8 on three d.f. With this value we can calculate three different tests of interest, all of which involve comparisons between nested models.

- As we move from model D to $A + D$ the deviance decreases by 37.2 while we lose three d.f. This statistic tests the hypothesis $H_0 : \alpha_i = 0$ for all i, concerning the *net* effect of age after adjusting for desire for more children, and is highly significant.
- As we move from model A to $A + D$ we reduce the deviance by 49.7 at the expense of one d.f. This chi-squared statistic tests the hypothesis $H_0 : \beta_2 = 0$ concerning the *net* effect of desire for more children after adjusting for age. This value is highly significant, so we reject the hypothesis of no net effects.
- Finally, the deviance of 16.8 on three d.f. is a measure of goodness of fit of the additive model: it compares this model with the saturated model, which adds an interaction between the two factors. Since the deviance exceeds 11.3, the one-percent critical value in the chi-squared

distribution for three d.f., we conclude that the additive model fails to fit the data.

Table 3.9 shows parameter estimates for the additive model. We show briefly how they would be interpreted, although we have evidence that the additive model does not fit the data.

TABLE 3.9: Parameter Estimates for Additive Logit Model of Contraceptive Use by Age (Grouped) and Desire for Children

Parameter		Symbol	Estimate	Std. Error	z-ratio
Constant		η	−1.694	0.135	−12.53
Age	25–29	α_2	0.368	0.175	2.10
	30–39	α_3	0.808	0.160	5.06
	40–49	α_4	1.023	0.204	5.01
Desire	No	β_2	0.824	0.117	7.04

The estimates of the α_j's show a monotonic effect of age on contraceptive use. Although there is evidence that this effect may vary depending on whether women desire more children, on average the odds of using contraception among women age 40 or higher are nearly three times the corresponding odds among women under age 25 in the same category of desire for another child.

Similarly, the estimate of β_2 shows a strong effect of wanting no more children. Although there is evidence that this effect may depend on the woman's age, on average the odds of using contraception among women who desire no more children are more than double the corresponding odds among women in the same age group who desire another child.

3.5.4 A Model With Interactions

We now consider a model which includes an interaction of age and desire for more children, denoted AD in Table 3.8. The model is

$$\text{logit}(\pi_{ij}) = \eta + \alpha_i + \beta_j + (\alpha\beta)_{ij},$$

where η is a constant, the α_i and β_j are the main effects of age and desire, and $(\alpha\beta)_{ij}$ is the interaction effect. To avoid redundancies we follow the reference cell method and set to zero all parameters involving the first cell, so that $\alpha_1 = \beta_1 = 0$, $(\alpha\beta)_{1j} = 0$ for all j and $(\alpha\beta)_{i1} = 0$ for all i. The remaining parameters may be interpreted as follows:

η is the logit of the reference group: women under age 25 who desire more children.

α_i for $i = 2, 3, 4$ are the effects of the age groups 25–29, 30–39 and 40–49, compared to ages under 25, for women who want another child.

β_2 is the effect of desiring no more children, compared to wanting another child, for women under age 25.

$(\alpha\beta)_{i2}$ for $i = 2, 3, 4$ is the *additional* effect of desiring no more children, compared to wanting another child, for women in age group i rather than under age 25. (This parameter is also the *additional* effect of age group i, compared to ages under 25, for women who desire no more children rather than those who want more.)

One way to simplify the presentation of results involving interactions is to combine the interaction terms with one of the main effects, and present them as effects of one factor within categories or levels of the other. In our example, we can combine the interactions $(\alpha\beta)_{i2}$ with the main effects of desire β_2, so that

$\beta_2 + (\alpha\beta)_{i2}$ is the effect of desiring no more children, compared to wanting another child, for women in age group i.

Of course, we could also combine the interactions with the main effects of age, and speak of age effects which are specific to women in each category of desire for more children. The two formulations are statistically equivalent, but the one chosen here seems demographically more sensible.

To obtain estimates based on this parameterization of the model we have to define the columns of the model matrix as follows. Let a_i be a dummy variable representing age group i, for $i = 2, 3, 4$, and let d take the value one for women who want no more children and zero otherwise. Then the model matrix $\mathbf{X}$ should have a column of ones to represent the constant or reference cell, the age dummies a_2, a_3 and a_4 to represent the age effects for women in the reference cell, and then the dummy d and the products a_2d, a_3d and a_4d, to represent the effect of wanting no more children at ages < 25, 25–29, 30–39 and 40–49, respectively. The resulting estimates and standard errors are shown in Table 3.10.

The results indicate that contraceptive use among women who desire more children varies little by age, increasing up to age 35–39 and then declining somewhat. On the other hand, the effect of wanting no more children

TABLE 3.10: Parameter Estimates for Model of Contraceptive Use With an Interaction Between Age (Grouped) and Desire for More Children

Parameter		Estimate	Std. Error	z-ratio
Constant		−1.519	0.145	−10.481
Age	25–29	0.368	0.201	1.832
	30–39	0.451	0.195	2.311
	40–49	0.397	0.340	1.168
Desires	<25	0.064	0.330	0.194
No More	25–29	0.331	0.241	1.372
at Age	30–39	1.154	0.174	6.640
	40–49	1.431	0.353	4.057

increases dramatically with age, from no effect among women below age 25 to an odds ratio of 4.18 at ages 40–49. Thus, in the older cohort the odds of using contraception among women who want no more children are four times the corresponding odds among women who desire more children. The results can also be summarized by noting that contraceptive use for spacing (i.e. among women who desire more children) does not vary much by age, but contraceptive use for limiting fertility (i.e among women who want no more children) increases sharply with age.

3.5.5 Analysis of Covariance Models

Since the model with an age by desire interaction is saturated, we have essentially reproduced the observed data. We now consider whether we could attain a more parsimonious fit by treating age as a variate and desire for more children as a factor, in the spirit of covariance analysis models.

Table 3.11 shows deviances for three models that include a linear effect of age using, as before, the midpoints of the age groups. To emphasize this point we use X rather than A to denote age.

The first model assumes that the logits are linear functions of age. This model fails to fit the data, which is not surprising because it ignores desire for more children, a factor that has a large effect on contraceptive use.

The next model, denoted $X + D$, is analogous to the two-factor additive model. It allows for an effect of desire for more children which is the same at all ages. This common effect is modelled by allowing each category of desire for more children to have its own constant, and results in two parallel lines. The common slope is the effect of age within categories of desire for

TABLE 3.11: Deviance Table for Models of Contraceptive Use by Age (Linear) and Desire for More Children

Model	Notation	$\text{logit}(\pi_{ij})$	Deviance	d.f.
One Line	X	$\alpha + \beta x_i$	68.88	6
Parallel Lines	$X + D$	$\alpha_j + \beta x_i$	18.99	5
Two Lines	XD	$\alpha_j + \beta_j x_i$	9.14	4

more children. The reduction in deviance of 39.9 on one d.f. indicates that desire for no more children has a strong effect on contraceptive use after controlling for a linear effect of age. However, the attained deviance of 19.0 on five d.f. is significant, indicating that the assumption of two parallel lines is not consistent with the data.

The last model in the table, denoted XD, includes an interaction between the linear effect of age and desire, and thus allows the effect of desire for more children to vary by age. This variation is modelled by allowing each category of desire for more children to have its own slope in addition to its own constant, and results in two regression lines. The reduction in deviance of 9.9 on one d.f. is a test of the hypothesis of parallelism or common slope $H_0 : \beta_1 = \beta_2$, which is rejected with a P-value of 0.002. The model deviance of 9.14 on four d.f. is just below the five percent critical value of the chi-squared distribution with four d.f., which is 9.49. Thus, we have no evidence against the assumption of two straight lines.

Before we present parameter estimates we need to discuss briefly the choice of parameterization. Direct application of the reference cell method leads us to use four variables: a dummy variable always equal to one, a variable x with the mid-points of the age groups, a dummy variable d which takes the value one for women who want no more children, and a variable dx equal to the product of this dummy by the mid-points of the age groups. This choice leads to parameters representing the constant and slope for women who want another child, and parameters representing the *difference* in constants and slopes for women who want no more children.

An alternative is to simply report the constants and slopes for the two groups defined by desire for more children. This parameterization can be easily obtained by omitting the constant and using the following four variables: d and $1 - d$ to represent the two constants and dx and $(1 - d)x$ to represent the two slopes. One could, of course, obtain the constant and slope for women who want no more children from the previous parameterization

simply by adding the main effect and the interaction. The simplest way to obtain the standard errors, however, is to change parameterization.

In both cases the constants represent effects at age zero and are not very meaningful. To obtain parameters that are more directly interpretable, we can center age around the sample mean, which is 30.6 years. Table 3.12 shows parameter estimates obtained under the two parameterizations discussed above, using the mid-points of the age groups minus the mean.

TABLE 3.12: Parameter Estimates for Model of Contraceptive Use With an Interaction Between Age (Linear) and Desire for More Children

Desire	Age	Symbol	Estimate	Std. Error	z-ratio
More	Constant	α_1	−1.1944	0.0786	−15.20
	Slope	β_1	0.0218	0.0104	2.11
No More	Constant	α_2	−0.4369	0.0931	−4.69
	Slope	β_2	0.0698	0.0114	6.10
Difference	Constant	$\alpha_2 - \alpha_1$	0.7575	0.1218	6.22
	Slope	$\beta_2 - \beta_1$	0.0480	0.0154	3.11

Thus, we find that contraceptive use increases with age, but at a faster rate among women who want no more children. The estimated slopes correspond to increases in the odds of two and seven percent per year of age for women who want and do not want more children, respectively. The difference of the slopes is significant by a likelihood ratio test or by Wald's test, with a z-ratio of 3.11.

Similarly, the effect of wanting no more children increases with age. The odds ratio around age 30.6—which we obtain by exponentiating the difference in constants–is 2.13, so not wanting more children at this age is associated with a doubling of the odds of using contraception. The difference in slopes of 0.048 indicates that this differential increases five percent per year of age.

The parameter estimates in Table 3.12 may be used to produce fitted logits for each age group and category of desire for more children. In turn, these can be compared with the empirical logits for the original eight groups, to obtain a visual impression of the nature of the relationships studied and the quality of the fit. The comparison appears in Figure 3.3, with the solid line representing the linear age effects (the dotted lines are discussed below). The graph shows clearly how the effect of wanting no more children increases with age (or, alternatively, how age has much stronger effects among limiters

than among spacers).

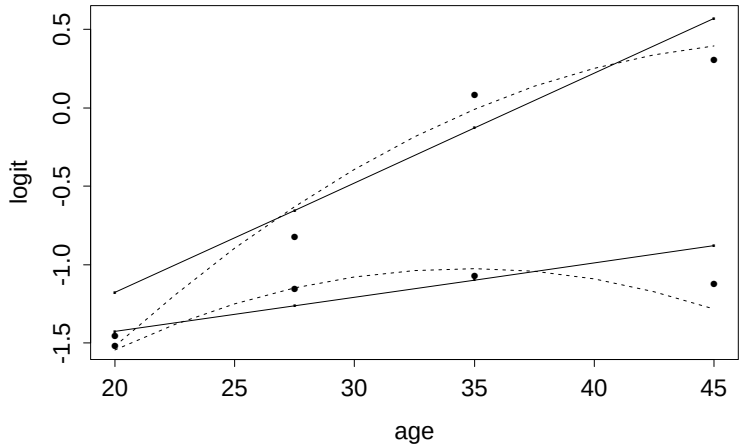


FIGURE 3.3: Observed and Fitted Logits for Models of Contraceptive Use With Effects of Age (Linear and Quadratic), Desire for More Children and a Linear Age by Desire Interaction.

The graph also shows that the assumption of linearity of age effects, while providing a reasonably parsimonious description of the data, is somewhat suspect, particularly at higher ages. We can improve the fit by adding higher-order terms on age. In particular

- Introducing a quadratic term on age yields an excellent fit, with a deviance of 2.34 on three d.f. This model consists of two parabolas, one for each category of desire for more children, but with the same curvature.

- Adding a quadratic age by desire interaction further reduces the deviance to 1.74 on two d.f. This model allows for two separate parabolas tracing contraceptive use by age, one for each category of desire.

Although the linear model passes the goodness of fit test, the fact that we can reduce the deviance by 6.79 at the expense of one d.f. indicates significant curvature. The dotted line in Figure 3.3 shows the intermediate model, where the curvature by age is the same for the two groups. While the fit is much better, the overall substantive conclusions do not change.

3.6 Multi-factor Models: Model Selection

Let us consider a full analysis of the contraceptive use data in Table 3.1, including all three predictors: age, education and desire for more children.

We use three subscripts to reflect the structure of the data, so π_{ijk} is the probability of using contraception in the (i, j, k)-th group, where $i = 1, 2, 3, 4$ indexes the age groups, $j = 1, 2$ the levels of education and $k = 1, 2$ the categories of desire for more children.

3.6.1 Deviances for One and Two-Factor Models

There are 19 basic models of interest for these data, which are listed for completeness in Table 3.13. Not all of these models would be of interest in any given analysis. The table shows the model in abbreviated notation, the formula for the linear predictor, the deviance and its degrees of freedom.

Note first that the null model does not fit the data. The assumption of a common probability of using contraception for all 16 groups of women is clearly untenable.

Next in the table we find the three possible one-factor models. Comparison of these models with the null model provides evidence of significant *gross* effects of age and desire for more children, but not of education. The likelihood ratio chi-squared tests are 91.7 on one d.f. for desire, 79.2 on three d.f. for age, and 0.7 on one d.f. for education.

Proceeding down the table we find the six possible two-factor models, starting with the additive ones. Here we find evidence of significant *net effects* of age and desire for more children after controlling for one other factor. For example the test for an effect of desire net of age is a chi-squared of 49.7 on one d.f., obtained by comparing the additive model $A + D$ on age and desire the one-factor model A with age alone. Education has a significant effect net of age, but not net of desire for more children. For example the test for the net effect of education controlling for age is 6.2 on one d.f., and follows from the comparison of the $A + E$ model with A. None of the additive models fits the data, but the closest one to a reasonable fit is $A + D$.

Next come the models involving *interactions* between two factors. We use the notation ED to denote the model with the main effects of E and D as well as the $E \times D$ interaction. Comparing each of these models with the corresponding additive model on the same two factors we obtain a test of the interaction effect. For example comparing the model ED with the additive model $E + D$ we can test whether the effect of desire for more children varies

TABLE 3.13: Deviance Table for Logit Models of Contraceptive Use by Age, Education and Desire for More Children

Model	$\text{logit}(\pi_{ijk})$	Dev.	d.f.
Null	η	165.77	15
One Factor			
*A*ge	$\eta+\alpha_i$	86.58	12
*E*ducation	$\eta+\beta_j$	165.07	14
*D*esire	$\eta+\gamma_k$	74.10	14
Two Factors			
$A+E$	$\eta+\alpha_i+\beta_j$	80.42	11
$A+D$	$\eta+\alpha_i+\gamma_k$	36.89	11
$E+D$	$\eta+\beta_j+\gamma_k$	73.87	13
AE	$\eta+\alpha_i+\beta_j+(\alpha\beta)_{ij}$	73.03	8
AD	$\eta+\alpha_i+\gamma_k+(\alpha\gamma)_{ik}$	20.10	8
ED	$\eta+\beta_j+\gamma_k+(\beta\gamma)_{jk}$	67.64	12
Three Factors			
$A+E+D$	$\eta+\alpha_i+\beta_j+\gamma_k$	29.92	10
$AE+D$	$\eta+\alpha_i+\beta_j+\gamma_k+(\alpha\beta)_{ij}$	23.15	7
$AD+E$	$\eta+\alpha_i+\beta_j+\gamma_k+(\alpha\gamma)_{ik}$	12.63	7
$A+ED$	$\eta+\alpha_i+\beta_j+\gamma_k+(\beta\gamma)_{jk}$	23.02	9
$AE+AD$	$\eta+\alpha_i+\beta_j+\gamma_k+(\alpha\beta)_{ij}+(\alpha\gamma)_{ik}$	5.80	4
$AE+ED$	$\eta+\alpha_i+\beta_j+\gamma_k+(\alpha\beta)_{ij}+(\beta\gamma)_{jk}$	13.76	6
$AD+ED$	$\eta+\alpha_i+\beta_j+\gamma_k+(\alpha\gamma)_{ik}+(\beta\gamma)_{jk}$	10.82	6
$AE+AD+ED$	$\eta+\alpha_i+\beta_j+\gamma_k+(\alpha\beta)_{ij}+(\alpha\gamma)_{ik}+(\beta\gamma)_{jk}$	2.44	3

with education. Making these comparisons we find evidence of interactions between age and desire for more children ($\chi^2 = 16.8$ on three d.f.), and between education and desire for more children ($\chi^2 = 6.23$ on one d.f.), but not between age and education ($\chi^2 = 7.39$ on three d.f.).

All of the results described so far could be obtained from two-dimensional tables of the type analyzed in the previous sections. The new results begin

to appear as we consider the nine possible three-factor models.

3.6.2 Deviances for Three-Factor Models

The first entry is the *additive* model $A+E+D$, with a deviance of 29.9 on ten d.f. This value represents a significant improvement over any of the additive models on two factors. Thus, we have evidence that there are significant net effects of age, education and desire for more children, considering each factor after controlling the other two. For example the test for a net effect of education controlling the other two variables compares the three-factor additive model $A+E+D$ with the model without education, namely $A+D$. The difference of 6.97 on one d.f. is significant, with a P-value of 0.008. However, the three-factor additive model does not fit the data.

The next step is to add *one interaction* between two of the factors. For example the model $AE + D$ includes the main effects of A, E and D and the $A \times E$ interaction. The interactions of desire for more children with age and with education produce significant gains over the additive model ($\chi^2 = 17.3$ on three d.f. and $\chi^2 = 6.90$ on one d.f., respectively), whereas the interaction between age and education is not significant ($\chi^2 = 6.77$ with three d.f.). These tests for interactions differ from those based on two-factor models in that they take into account the third factor. The best of these models is clearly the one with an interaction between age and desire for more children, $AD+E$. This is also the first model in our list that actually passes the goodness of fit test, with a deviance of 12.6 on seven d.f.

Does this mean that we can stop our search for an adequate model? Unfortunately, it does not. The goodness of fit test is a joint test for all terms omitted in the model. In this case we are testing for the AE, ED and AED interactions simultaneously, a total of seven parameters. This type of omnibus test lacks power against specific alternatives. It is possible that one of the omitted terms (or perhaps some particular contrast) would be significant by itself, but its effect may not stand out in the aggregate. At issue is whether the remaining deviance of 12.6 is spread out uniformly over the remaining d.f. or is concentrated in a few d.f. If you wanted to be absolutely sure of not missing anything you might want to aim for a deviance below 3.84, which is the five percent critical value for one d.f., but this strategy would lead to over-fitting if followed blindly.

Let us consider the models involving *two interactions* between two factors, of which there are three. Since the AD interaction seemed important we restrict attention to models that include this term, so we start from $AD + E$, the best model so far. Adding the age by education interaction

AE to this model reduces the deviance by 6.83 at the expense of three d.f. A formal test concludes that this interaction is not significant. If we add instead the education by desire interaction ED we reduce the deviance by only 1.81 at the expense of one d.f. This interaction is clearly not significant. A model-building strategy based on *forward selection* of variables would stop here and choose $AD + E$ as the best model on grounds of parsimony and goodness of fit.

An alternative approach is to start with the saturated model and impose progressive simplification. Deleting the *three-factor interaction* yields the model $AE + AD + ED$ with three two-factor interactions, which fits the data rather well, with a deviance of just 2.44 on three d.f. If we were to delete the AD interaction the deviance would rise by 11.32 on three d.f., a significant loss. Similarly, removing the AE interaction would incur a significant loss of 8.38 on 3 d.f. We can, however, drop the ED interaction with a non-significant increase in deviance of 3.36 on one d.f. At this point we can also eliminate the AE interaction, which is no longer significant, with a further loss of 6.83 on three d.f. Thus, a *backward elimination* strategy ends up choosing the same model as forward selection.

Although you may find these results reassuring, there is a fact that both approaches overlook: the AE and DE interactions are jointly significant! The change in deviance as we move from $AD+E$ to the model with three two-factor interactions is 10.2 on four d.f., and exceeds (although not by much) the five percent critical value of 9.5. This result indicates that we need to consider the more complicated model with all three two-factor interactions. Before we do that, however, we need to discuss parameter estimates for selected models.

3.6.3 The Additive Model: Gross and Net Effects

Consider first Table 3.14, where we adopt an approach similar to multiple classification analysis to compare the gross and net effects of all three factors. We use the reference cell method, and include the omitted category for each factor (with a dash where the estimated effect would be) to help the reader identify the baseline.

The gross or unadjusted effects are based on the single-factor models A, E and D. These effects represent overall differences between levels of each factor, and as such they have descriptive value even if the one-factor models do not tell the whole story. The results can easily be translated into odds ratios. For example not wanting another child is associated with an increase in the odds of using contraception of 185%. Having upper primary or higher

TABLE 3.14: Gross and Net Effects of Age, Education and Desire for More Children on Current Use of Contraception

Variable and category		Gross effect	Net effect
Constant		–	−1.966
Age	<25	–	–
	25–29	0.461	0.389
	30–39	1.048	0.909
	40–49	1.425	1.189
Education			
	Lower	–	–
	Upper	-0.093	0.325
Desires More			
	Yes	–	–
	No	1.049	0.833

education rather than lower primary or less appears to reduce the odds of using contraception by almost 10%.

The net or adjusted effects are based on the three-factor additive model $A+E+D$. This model assumes that the effect of each factor is the same for all categories of the others. We know, however, that this is not the case—particularly with desire for more children, which has an effect that varies by age—so we have to interpret the results carefully. The net effect of desire for more children shown in Table 3.14 represents an average effect across all age groups and may not be representative of the effect at any particular age. Having said that, we note that desire for no more children has an important effect net of age and education: on the average, it is associated with an increase in the odds of using contraception of 130%.

The result for education is particularly interesting. Having upper primary or higher education is associated with an increase in the odds or using contraception of 38%, compared to having lower primary or less, after we control for age and desire for more children. The gross effect was close to zero. To understand this result bear in mind that contraceptive use in Fiji occurs mostly among older women who want no more children. Education has no effect when considered by itself because in Fiji more educated women are likely to be younger than less educated women, and thus at a stage of their lives when they are less likely to have reached their desired family size,

even though they may want fewer children. Once we adjust for their age, calculating the net effect, we obtain the expected association. In this example age is said to act as a *suppressor* variable, masking the association between education and contraceptive use.

We could easily add columns to Table 3.14 to trace the effects of one factor after controlling for one or both of the other factors. We could, for example, examine the effect of education adjusted for age, the effect adjusted for desire for more children, and finally the effect adjusted for both factors. This type of analysis can yield useful insights into the confounding influences of other variables.

3.6.4 The Model with One Interaction Effect

Let us now examine parameter estimates for the model with an age by desire for more children interaction $AD + E$, where

$$\text{logit}(\pi_{ijk}) = \eta + \alpha_i + \beta_j + \gamma_j + (\alpha\gamma)_{ik}.$$

The parameter estimates depend on the restrictions used in estimation. We use the reference cell method, so that $\alpha_1 = \beta_1 = \gamma_1 = 0$, and $(\alpha\gamma)_{ik} = 0$ when either $i = 1$ or $k = 1$.

In this model η is the logit of the probability of using contraception in the reference cell, that is, for women under 25 with lower primary or less education who want another child. On the other hand β_2 is the effect of upper primary or higher education, compared to lower primary or less, for women in any age group or category of desire for another child. The presence of an interaction makes interpretation of the estimates for age and desire somewhat more involved:

α_i represents the effect of age group i, compared to age < 25, for women who want more children.

γ_2 represents the effect of wanting no more children, compared to desiring more, for women under age 25.

$(\alpha\gamma)_{i2}$, the interaction term, can be interpreted as the *additional* effect of wanting no more children among women in age group i, compared to women under age 25.

It is possible to simplify slightly the presentation of the results by combining the interactions with some of the main effects. In the present example, it is convenient to present the estimates of α_i as the age effects for women who

TABLE 3.15: The Estimates

Variable	Category	Symbol	Estimate	Std. Err	z-ratio
Constant		η	-1.803	0.180	-10.01
Age	25–29	α_2	0.395	0.201	1.96
	30–39	α_3	0.547	0.198	2.76
	40–49	α_4	0.580	0.347	1.67
Education	Upper	β_2	0.341	0.126	2.71
Desires	<25	γ_2	0.066	0.331	0.20
no more	25–29	$\gamma_2 + (\alpha\gamma)_{22}$	0.325	0.242	1.35
at age	30–39	$\gamma_2 + (\alpha\gamma)_{32}$	1.179	0.175	6.74
	40–49	$\gamma_2 + (\alpha\gamma)_{42}$	1.428	0.354	4.04

want another child, and to present $\gamma_2 + (\alpha\gamma)_{i2}$ as the effect of not wanting another child for women in age group i.

Calculation of the necessary dummy variables proceeds exactly as in Section 3.5. This strategy leads to the parameter estimates in Table 3.15.

To aid in interpretation as well as model criticism, Figure 3.4 plots observed logits based on the original data in Table 3.1, and fitted logits based on the model with an age by desire interaction.

The graph shows four curves tracing contraceptive use by age for groups defined by education and desire for more children. The curves are labelled using L and U for lower and upper education, and Y and N for desire for more children. The lowest curve labelled LY corresponds to women with lower primary education or less who want more children, and shows a slight increase in contraceptive use up to age 35–39 and then a small decline. The next curve labelled UY is for women with upper primary education or more who also want more children. This curve is parallel to the previous one because the effect of education is additive on age. The constant difference between these two curves corresponds to a 41% increase in the odds ratio as we move from lower to upper primary education. The third curve, labelled LN, is for women with lower primary education or less who want no more children. The distance between this curve and the first one represents the effect of wanting no more children at different ages. This effect increases sharply with age, reaching an odds ratio of four by age 40–49. The fourth curve, labelled UN, is for women with upper primary education or more who want no more children. The distance between this curve and the previous one is the effect of education, which is the same whether women want more children or not, and is also the same at every age.

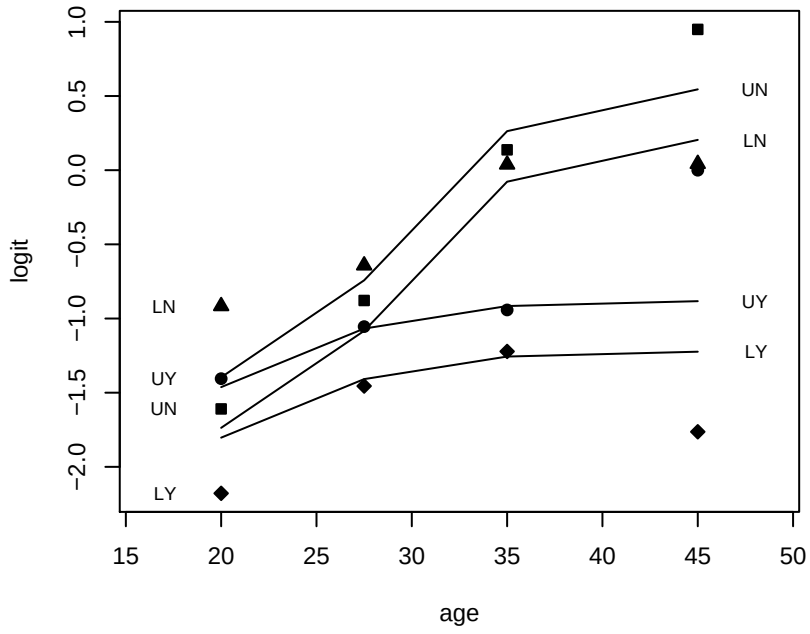


FIGURE 3.4: Logit Model of Contraceptive Use By Age, Education and Desire for Children, With Age by Desire Interaction

The graph also shows the observed logits, plotted using different symbols for each of the four groups defined by education and desire. Comparison of observed and fitted logits shows clearly the strengths and weaknesses of this model: it does a fairly reasonable job reproducing the logits of the proportions using contraception in each group *except* for ages 40–49 (and to a lesser extend the group < 25), where it seems to underestimate the educational differential. There is also some indication that this failure may be more pronounced for women who want more children.

3.6.5 Best Fitting and Parsimonious Models

How can we improve the model of the last section? The most obvious solution is to move to the model with all three two-factor interactions, $AE + AD + ED$, which has a deviance of 2.44 on three d.f. and therefore fits the data

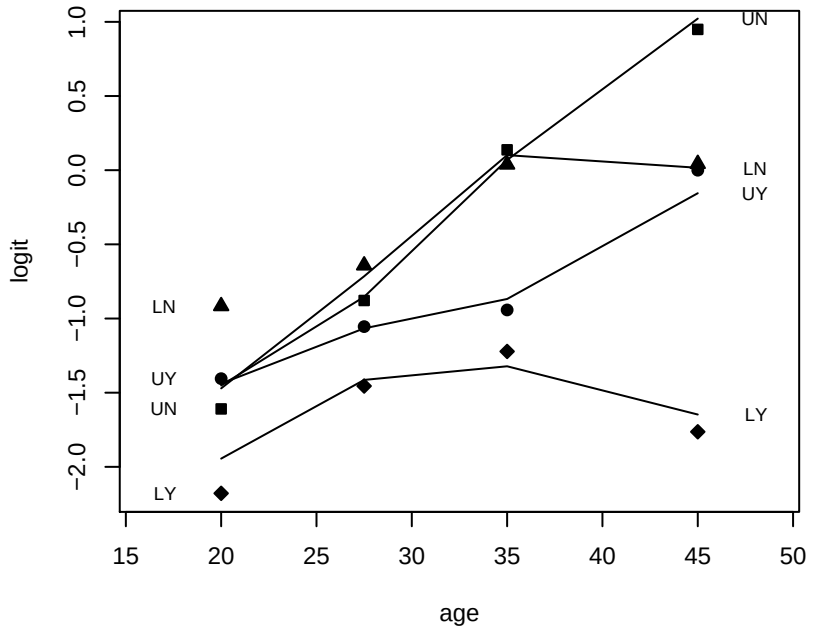


FIGURE 3.5: Observed and Fitted Logits of Contraceptive Use Based on Model with Three Two-Factor Interactions

extremely well. This model implies that the effect of each factor depends on the levels of the other two, but not on the combination of levels of the other two. Interpretation of the coefficients in this model is not as simple as it would be in an additive model, or in a model involving only one interaction. The best strategy in this case is to plot the fitted model and inspect the resulting curves.

Figure 3.5 shows fitted values based on the more complex model. The plot tells a simple story. Contraceptive use for spacing increases slightly up to age 35 and then declines for the less educated but continues to increase for the more educated. Contraceptive use for limiting increases sharply with age up to age 35 and then levels off for the less educated, but continues to increase for the more educated. The figure shows that the effect of wanting no more children increases with age, and appears to do so for both educational groups in the same way (look at the distance between the LY and LN curves, and

between the UY and UN curves). On the other hand, the effect of education is clearly more pronounced at ages 40–49 than at earlier ages, and also seems slightly larger for women who want more children than for those who do not (look at the distance between the LY and UY curves, and between the LN and UN curves).

One can use this knowledge to propose improved models that fit the data without having to use all three two-factor interactions. One approach would note that all interactions with age involve contrasts between ages 40–49 and the other age groups, so one could collapse age into only two categories for purposes of modelling the interactions. A simplified version of this approach is to start from the model $AD + E$ and add one d.f. to model the larger educational effect for ages 40–49. This can be done by adding a dummy variable that takes the value one for women aged 40–49 who have upper primary or more education. The resulting model has a deviance of 6.12 on six d.f., indicating a good fit. Comparing this value with the deviance of 12.6 on seven d.f. for the $AD + E$ model, we see that we reduced the deviance by 6.5 at the expense of a single d.f. The model $AD + AE$ includes all three d.f. for the age by education interaction, and has a deviance of 5.8 on four d.f. Thus, the total contribution of the AE interaction is 6.8 on three d.f. Our one-d.f. improvement has captured roughly 90% of this interaction.

An alternative approach is to model the effects of education and desire for no more children as smooth functions of age. The logit of the probability of using contraception is very close to a linear function of age for women with upper primary education who want no more children, who could serve as a new reference cell. The effect of wanting more children could be modelled as a linear function of age, and the effect of education could be modelled as a quadratic function of age. Let L_{ijk} take the value one for lower primary or less education and zero otherwise, and let M_{ijk} be a dummy variable that takes the value one for women who want more children and zero otherwise. Then the proposed model can be written as

$$\text{logit}(\pi_{ijk}) = \alpha + \beta x_{ijk} + (\alpha_E + \beta_E x_{ijk} + \gamma_E x_{ijk}^2)L_{ijk} + (\alpha_D + \beta_D x_{ijk})M_{ijk}.$$

Fitting this model, which requires only seven parameters, gives a deviance of 7.68 on nine d.f. The only weakness of the model is that it assumes equal effects of education on use for limiting and use for spacing, but these effects are not well-determined. Further exploration of these models is left as an exercise.

3.7 Other Choices of Link

All the models considered so far use the logit transformation of the probabilities, but other choices are possible. In fact, any transformation that maps probabilities into the real line could be used to produce a generalized linear model, as long as the transformation is one-to-one, continuous and differentiable.

In particular, suppose $F(.)$ is the cumulative distribution function (c.d.f.) of a random variable defined on the real line, and write

$$\pi_i = F(\eta_i),$$

for $-\infty < \eta_i < \infty$. Then we could use the inverse transformation

$$\eta_i = F^{-1}(\pi_i),$$

for $0 < \pi_i < 1$ as the link function.

Popular choices of c.d.f.'s in this context are the normal, logistic and extreme value distributions. In this section we motivate this general approach by introducing models for binary data in terms of latent variables.

3.7.1 A Latent Variable Formulation

Let Y_i denote a random variable representing a binary response coded zero and one, as usual. We will call Y_i the *manifest* response. Suppose that there is an unobservable continuous random variable Y_i^* which can take any value in the real line, and such that Y_i takes the value one if an only if Y_i^* exceeds a certain threshold θ. We will call Y_i^* the *latent* response. Figure 3.6 shows the relationship between the latent variable and the response when the threshold is zero.

The interpretation of Y_i and Y_i^* depends on the context. An economist, for example, may view Y_i as a binary choice, such as purchasing or renting a home, and Y_i^* as the difference in the utilities of purchasing and renting. A psychologist may view Y_i as a response to an item in an attitude scale, such as agreeing or disagreeing with school vouchers, and Y_i^* as the underlying attitude. Biometricians often view Y_i^* as a dose and Y_i as a response, hence the name dose-response models.

Since a positive outcome occurs only when the latent response exceeds the threshold, we can write the probability π_i of a positive outcome as

$$\pi_i = \Pr\{Y_i = 1\} = \Pr\{Y_i^* > \theta\}.$$

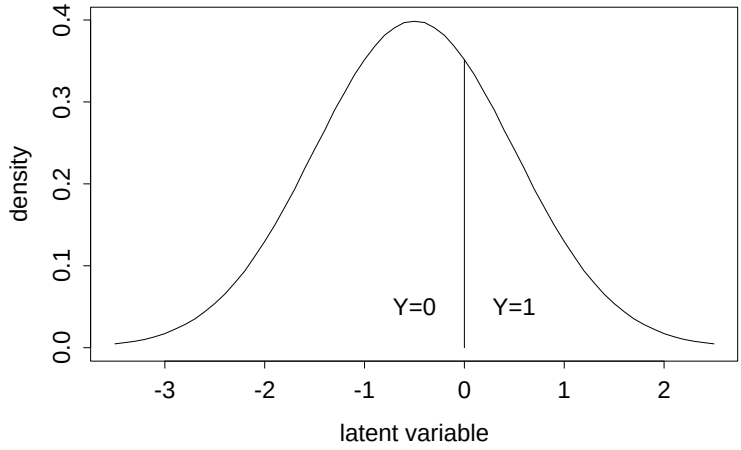


FIGURE 3.6: Latent Variable and Manifest Response

As often happens with latent variables, the location and scale of Y_i^* are arbitrary. We can add a constant a to both Y_i^* and the threshold θ, or multiply both by a constant c, without changing the probability of a positive outcome. To identify the model we take the threshold to be zero, and standardize Y_i^* to have standard deviation one (or any other fixed value).

Suppose now that the outcome depends on a vector of covariates $\mathbf{x}$. To model this dependence we use an ordinary linear model for the *latent* variable, writing

$$Y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + U_i, \tag{3.15}$$

where $\boldsymbol{\beta}$ is a vector of coefficients of the covariates $\mathbf{x}_i$ and U_i is the error term, assumed to have a distribution with c.d.f. $F(u)$, not necessarily the normal distribution.

Under this model, the probability π_i of observing a positive outcome is

$$\begin{aligned} \pi_i &= \Pr\{Y_i > 0\} \\ &= \Pr\{U_i > -\eta_i\} \\ &= 1 - F(-\eta_i), \end{aligned}$$

where $\eta_i = \mathbf{x}_i'\boldsymbol{\beta}$ is the linear predictor. If the distribution of the error term U_i is symmetric about zero, so $F(u) = 1 - F(-u)$, we can write

$$\pi_i = F(\eta_i)$$

This expression defines a generalized linear model with Bernoulli response and link

$$\eta_i = F^{-1}(\pi_i). \tag{3.16}$$

In the more general case where the distribution of the error term is not necessarily symmetric, we still have a generalized linear model with link

$$\eta_i = -F^{-1}(1 - \pi_i). \tag{3.17}$$

We now consider some specific distributions.

3.7.2 Probit Analysis

The obvious choice of an error distribution is the normal. Assuming that the error term has a standard normal distribution $U_i \sim N(0, 1)$, the results of the previous section lead to

$$\pi_i = \Phi(\eta_i),$$

where Φ is the standard normal c.d.f. The inverse transformation, which gives the linear predictor as a function of the probability

$$\eta_i = \Phi^{-1}(\pi_i),$$

is called the *probit*.

It is instructive to consider the more general case where the error term $U_i \sim N(0, \sigma^2)$ has a normal distribution with variance σ^2. Following the same steps as before we find that

$$\begin{aligned}\pi_i &= \Pr\{Y_i^* > 0\} \\ &= \Pr\{U_i > -\mathbf{x}_i'\boldsymbol{\beta}\} = \Pr\{U_i/\sigma > -\mathbf{x}_i'\boldsymbol{\beta}/\sigma\} \\ &= 1 - \Phi(-\mathbf{x}_i'\boldsymbol{\beta}/\sigma) = \Phi(\mathbf{x}_i'\boldsymbol{\beta}/\sigma),\end{aligned}$$

where we have divided by σ to obtain a standard normal variate, and used the symmetry of the normal distribution to obtain the last result.

This development shows that we cannot identify $\boldsymbol{\beta}$ and σ separately, because the probability depends on them only through their ratio $\boldsymbol{\beta}/\sigma$. This is another way of saying that the scale of the latent variable is not identified. We therefore take $\sigma = 1$, or equivalently interpret the β's in units of standard deviation of the latent variable.

As a simple example, consider fitting a probit model to the contraceptive use data by age and desire for more children. In view of the results in Section 3.5, we introduce a main effect of wanting no more children, a linear effect

TABLE 3.16: Estimates for Probit Model of Contraceptive Use With a Linear Age by Desire Interaction

Parameter	Symbol	Estimate	Std. Error	z-ratio
Constant	α_1	−0.7297	0.0460	−15.85
Age	β_1	0.0129	0.0061	2.13
Desire	$\alpha_2 - \alpha_1$	0.4572	0.0731	6.26
Age × Desire	$\beta_2 - \beta_1$	0.0305	0.0092	3.32

of age, and a linear age by desire interaction. Fitting this model gives a deviance of 8.91 on four d.f. Estimates of the parameters and standard errors appear in Table 3.16

To interpret these results we imagine a latent continuous variable representing the woman's motivation to use contraception (or the utility of using contraception, compared to not using). At the average age of 30.6, not wanting more children increases the motivation to use contraception by almost half a standard deviation. Each year of age is associated with an increase in motivation of 0.01 standard deviations if she wants more children and 0.03 standard deviations more (for a total of 0.04) if she does not. In the next section we compare these results with logit estimates.

A slight disadvantage of using the normal distribution as a link for binary response models is that the c.d.f. does not have a closed form, although excellent numerical approximations and computer algorithms are available for computing both the normal probability integral and its inverse, the probit.

3.7.3 Logistic Regression

An alternative to the normal distribution is the standard logistic distribution, whose shape is remarkably similar to the normal distribution but has the advantage of a closed form expression

$$\pi_i = F(\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}},$$

for $-\infty < \eta_i < \infty$. The standard logistic distribution is symmetric, has mean zero, and has variance $\pi^2/3$. The shape is very close to the normal, except that it has heavier tails. The inverse transformation, which can be obtained solving for η_i in the expression above is

$$\eta_i = F^{-1}(\pi_i) = \log \frac{\pi_i}{1 - \pi_i},$$

our good old friend, the *logit*.

Thus, coefficients in a logit regression model can be interpret not only in terms of log-odds, but also as effects of the covariates on a latent variable that follows a linear model with logistic errors.

The logit and probit transformations are almost linear functions of each other for values of π_i in the range from 0.1 to 0.9, and therefore tend to give very similar results. Comparison of probit and logit coefficients should take into account the fact that the standard normal and the standard logistic distributions have different variances. Recall that with binary data we can only estimate the ratio $\boldsymbol{\beta}/\sigma$. In probit analysis we have implicitly set $\sigma = 1$. In a logit model, by using a standard logistic error term, we have effectively set $\sigma = \pi/\sqrt{3}$. Thus, coefficients in a logit model should be standardized dividing by $\pi/\sqrt{3}$ before comparing them with probit coefficients.

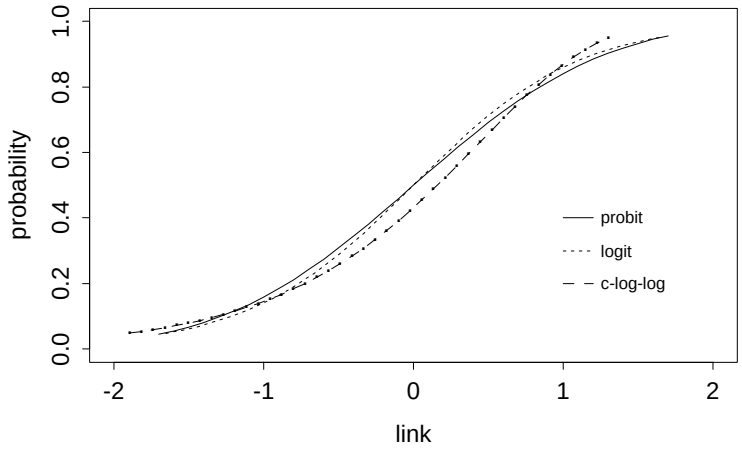


FIGURE 3.7: The Standardized Probit, Logit and C-Log-Log Links

Figure 3.7 compares the logit and probit links (and a third link discussed below) after standardizing the logits to unit variance. The solid line is the probit and the dotted line is the logit divided by $\pi/\sqrt{3}$. As you can see, they are barely distinguishable.

To illustrate the similarity of these links in practice, consider our models of contraceptive use by age and desire for more children in Tables 3.10 and 3.16. The deviance of 9.14 for the logit model is very similar to the deviance of 8.91 for the probit model, indicating an acceptable fit. The Wald tests of individual coefficients are also very similar, for example the test for the effect of wanting no more children at age 30.6 is 6.22 in the logit model and 6.26

in the probit model. The coefficients themselves look somewhat different, but of course they are not standardized. The effect of wanting no more children at the average age is 0.758 in the logit scale. Dividing by $\pi/\sqrt{3}$, the standard deviation of the underlying logistic distribution, we find this effect equivalent to an increase in the latent variable of 0.417 standard deviations. The probit analysis estimates the effect as 0.457 standard deviations.

3.7.4 The Complementary Log-Log Transformation

A third choice of link is the complementary log-log transformation

$$\eta_i = \log(-\log(1-\pi_i)),$$

which is the inverse of the c.d.f. of the extreme value (or log-Weibull) distribution, with c.d.f.

$$F(\eta_i) = 1 - e^{-e^{\eta_i}}.$$

For small values of π_i the complementary log-log transformation is close to the logit. As the probability increases, the transformation approaches infinity more slowly that either the probit or logit.

This particular choice of link function can also be obtained from our general latent variable formulation if we assume that $-U_i$ (note the minus sign) has a standard extreme value distribution, so the error term itself has a *reverse* extreme value distribution, with c.d.f.

$$F(U_i) = e^{-e^{-U_i}}.$$

The reverse extreme value distribution is asymmetric, with a long tail to the right. It has mean equal to Euler's constant 0.577 and variance $\pi^2/6 = 1.645$. The median is $-\log\log 2 = 0.367$ and the quartiles are -0.327 and 1.246.

Inverting the reverse extreme value c.d.f. and applying Equation 3.17, which is valid for both symmetric and asymmetric distributions, we find that the link corresponding to this error distribution is the complementary log-log.

Thus, coefficients in a generalized linear model with binary response and a complementary log-log link can be interpreted as effects of the covariates on a latent variable which follows a linear model with reverse extreme value errors.

To compare these coefficients with estimates based on a probit analysis we should standardize them, dividing by $\pi/\sqrt{6}$. To compare coefficients with logit analysis we should divide by $\sqrt{2}$, or standardize both c-log-log and logit coefficients.

Figure 3.7 compares the c-log-log link with the probit and logit after standardizing it to have mean zero and variance one. Although the c-log-log link differs from the other two, one would need extremely large sample sizes to be able to discriminate empirically between these links.

The complementary log-log transformation has a direct interpretation in terms of hazard ratios, and thus has practical applications in terms of hazard models, as we shall see later in the sequel.

3.8 Regression Diagnostics for Binary Data

Model checking is just as important in logistic regression and probit analysis as it is in classical linear models. The raw materials are again the residuals, or differences between observed and fitted values. Unlike the case of linear models, however, we now have to make allowance for the fact that the observations have different variances. There are two types of residuals in common use.

3.8.1 Pearson Residuals

A very simple approach to the calculation of residuals is to take the difference between observed and fitted values and divide by an estimate of the standard deviation of the observed value. The resulting residual has the form

$$p_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i(n_i - \hat{\mu}_i)/n_i}}, \tag{3.18}$$

where $\hat{\mu}_i$ is the fitted value and the denominator follows from the fact that $\text{var}(y_i) = n_i\pi_i(1-\pi_i)$.

The result is called the Pearson residual because the square of p_i is the contribution of the i-th observation to Pearson's chi-squared statistic, which was introduced in Section 3.2.2, Equation 3.14.

With grouped data the Pearson residuals are approximately normally distributed, but this is not the case with individual data. In both cases, however, observations with a Pearson residual exceeding two in absolute value may be worth a closer look.

3.8.2 Deviance Residuals

An alternative residual is based on the deviance or likelihood ratio chi-squared statistic. The deviance residual is defined as

$$d_i = \sqrt{2[y_i \log(\frac{y_i}{\hat{\mu}_i}) + (n_i - y_i)\log(\frac{n_i - y_i}{n_i - \hat{\mu}_i})]}, \tag{3.19}$$

with the same sign as the raw residual $y_i - \hat{y}_i$. Squaring these residuals and summing over all observations yields the deviance statistic. Observations with a deviance residual in excess of two may indicate lack of fit.

3.8.3 Studentized Residuals

The residuals defined so far are not fully standardized. They take into account the fact that different observations have different variances, but they make no allowance for additional variation arising from estimation of the parameters, in the way studentized residuals in classical linear models do.

Pregibon (1981) has extended to logit models some of the standard regression diagnostics. A key in this development is the weighted *hat* matrix

$$\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{1/2},$$

where $\mathbf{W}$ is the diagonal matrix of iteration weights from Section 3.2.1, with entries $w_{ii} = \mu_i(n_i - \mu_i)/n_i$, evaluated at the m.l.e.'s. Using this expression it can be shown that the variance of the raw residual is, to a first-order approximation,

$$\text{var}(y_i - \hat{\mu}_i) \approx (1 - h_{ii})\text{var}(y_i),$$

where $h_i i$ is the leverage or diagonal element of the weighted hat matrix. Thus, an internally studentized residual can be obtained dividing the Pearson residual by the square root of $1 - h_{ii}$, to obtain

$$s_i = \frac{p_i}{\sqrt{1 - h_{ii}}} = \frac{y_i - \hat{\mu}_i}{\sqrt{(1 - h_{ii})\hat{\mu}_i(n_i - \hat{\mu}_i)/n_i}}.$$

A similar standardization can be applied to deviance residuals. In both cases the standardized residuals have the same variance only approximately because the correction is first order, unlike the case of linear models where the correction was exact.

Consider now calculating jack-knifed residuals by omitting one observation. Since estimation relies on iterative procedures, this calculation would

be expensive. Suppose, however, that we start from the final estimates and do only one iteration of the IRLS procedure. Since this step is a standard weighted least squares calculation, we can apply the standard regression updating formulas to obtain the new coefficients and thus the predictive residuals. Thus, we can calculate a jack-knifed residual as a function of the standardized residual using the same formula as in linear models

$$t_i = s_i \sqrt{\frac{n-p-1}{n-p-s_i^2}}$$

and view the result as a one-step approximation to the true jack-knifed residual.

3.8.4 Leverage and Influence

The diagonal elements of the hat matrix can be interpreted as leverages just as in linear models. To measure actual rather than potential influence we could calculate Cook's distance, comparing $\hat{\boldsymbol{\beta}}$ with $\hat{\boldsymbol{\beta}}_{(i)}$, the m.l.e.'s of the coefficients with and without the i-th observation. Calculation of the later would be expensive if we iterated to convergence. Pregibon (1981), however, has shown that we can use the standard linear models formula

$$D_i = s_i^2 \frac{h_{ii}}{(1-h_{ii})p},$$

and view the result as a one-step approximation to Cook's distance, based on doing one iteration of the IRLS algorithm towards $\hat{\boldsymbol{\beta}}_{(i)}$ starting from the complete data estimate $\hat{\boldsymbol{\beta}}$.

3.8.5 Testing Goodness of Fit

With grouped data we can assess goodness of fit by looking directly at the deviance, which has approximately a chi-squared distribution for large n_i. A common rule of thumb is to require all expected frequencies (both expected successes $\hat{\mu}_i$ and failures $n_i - \hat{\mu}_i$) to exceed one, and 80% of them to exceed five.

With individual data this test is not available, but one can always group the data according to their covariate patterns. If the number of possible combinations of values of the covariates is not too large relative to the total sample size, it may be possible to group the data and conduct a formal goodness of fit test. Even when the number of covariate patterns is large, it is possible that a few patterns will account for most of the observations. In this

case one could compare observed and fitted counts at least for these common patterns, using either the deviance or Pearson's chi-squared statistic.

Hosmer and Lemeshow (1980, 1989) have proposed an alternative procedure that can be used with individual data even if there are no common covariate patterns. The basic idea is to use predicted probabilities to create groups. These authors recommend forming ten groups, with predicted probabilities of 0–0.1, 0.1–0.2, and so on, with the last group being 0.9–1. One can then compute expected counts of successes (and failures) for each group by summing the predicted values (and their complements), and compare these with observed values using Pearson's chi-squared statistic. Simulation studies show that the resulting statistic has approximately in large samples the usual chi-squared distribution, with degrees of freedom equal to $g - 2$, where g is the number of groups, usually ten. It seems reasonable to assume that this result would also apply if one used the deviance rather than Pearson's chi-squared.

Another measure that has been proposed in the literature is a pseudo-R^2, based on the proportion of deviance explained by a model. This is a direct extension of the calculations based on RSS's for linear models. These measures compare a given model with the null model, and as such do not necessarily measure goodness of fit. A more direct measure of goodness of fit would compare a given model with the saturated model, which brings us back again to the deviance.

Yet another approach to assessing goodness of fit is based on prediction errors. Suppose we were to use the fitted model to predict 'success' if the fitted probability exceeds 0.5 and 'failure' otherwise. We could then crosstabulate the observed and predicted responses, and calculate the proportion of cases predicted correctly. While intuitively appealing, one problem with this approach is that a model that fits the data may not necessarily predict well, since this depends on how predictable the outcome is. If prediction was the main objective of the analysis, however, the proportion classified correctly would be an ideal criterion for model comparison.

12.42 In Example 12.8, a case is made for eliminating x_1, powder temperature, from the model since the P-value based on the F-test is 0.2156 while P-values for x_2 and x_3 are near zero.

(a) Reduce the model by eliminating x_1, thereby producing a full and a restricted (or reduced) model, and compare them on the basis of R^2_{adj}.

(b) Compare the full and restricted models using the width of the 95% prediction intervals on a new observation. The better of the two models would be that with the tightened prediction intervals. Use the average of the width of the prediction intervals.

12.43 Consider the data of Exercise 12.13 on page 452. Can the response, wear, be explained adequately by a single variable (either viscosity or load) in an SLR rather than with the full two-variable regression? Justify your answer thoroughly through tests of hypotheses as well as comparison of the three competing models.

12.44 For the data set given in Exericise 12.16 on page 453, can the response be explained adequately by any two regressor variables? Discuss.

12.8 Categorical or Indicator Variables

An extremely important special-case application of multiple linear regression occurs when one or more of the regressor variables are **categorical**, **indicator**, or **dummy variables**. In a chemical process, the engineer may wish to model the process yield against regressors such as process temperature and reaction time. However, there is interest in using two different catalysts and somehow including "the catalyst" in the model. The catalyst effect cannot be measured on a continuum and is hence a categorical variable. An analyst may wish to model the price of homes against regressors that include square feet of living space x_1, the land acreage x_2, and age of the house x_3. These regressors are clearly continuous in nature. However, it is clear that cost of homes may vary substantially from one area of the country to another. If data are collected on homes in the east, midwest, south, and west, we have an indicator variable with **four categories**. In the chemical process example, if two catalysts are used, we have an indicator variable with two categories. In a biomedical example in which a drug is to be compared to a placebo, all subjects are evaluated on several continuous measurements such as age, blood pressure, and so on, as well as gender, which of course is categorical with two categories. So, included along with the continuous variables are two indicator variables: treatment with two categories (active drug and placebo) and gender with two categories (male and female).

Model with Categorical Variables

Let us use the chemical processing example to illustrate how indicator variables are involved in the model. Suppose y = yield and x_1 = temperature and x_2 = reaction time. Now let us denote the indicator variable by z. Let $z = 0$ for catalyst 1 and $z = 1$ for catalyst 2. The assignment of the $(0, 1)$ indicator to the catalyst is arbitrary. As a result, the model becomes

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 z_i + \epsilon_i, \quad i = 1, 2, \ldots, n.$$

Three Categories

The estimation of coefficients by the method of least squares continues to apply. In the case of three levels or categories of a single indicator variable, the model will

include **two** regressors, say z_1 and z_2, where the $(0, 1)$ assignment is as follows:

$$\begin{array}{cc} z_1 & z_2 \\ \begin{bmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \end{array},$$

where **0** and **1** are vectors of 0's and 1's, respectively. In other words, if there are ℓ categories, the model includes $\ell - 1$ actual model terms.

It may be instructive to look at a graphical representation of the model with three categories. For the sake of simplicity, let us assume a single continuous variable x. As a result, the model is given by

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_{1i} + \beta_3 z_{2i} + \epsilon_i.$$

Thus, Figure 12.2 reflects the nature of the model. The following are model expressions for the three categories.

$$\begin{aligned} E(Y) &= (\beta_0 + \beta_2) + \beta_1 x, && \text{category 1,} \\ E(Y) &= (\beta_0 + \beta_3) + \beta_1 x, && \text{category 2,} \\ E(Y) &= \beta_0 + \beta_1 x, && \text{category 3.} \end{aligned}$$

As a result, the model involving categorical variables essentially involves a **change in the intercept** as we change from one category to another. Here of course we are assuming that the **coefficients of continuous variables remain the same across the categories**.

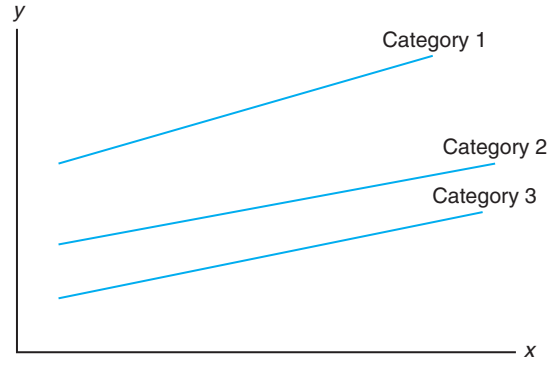


Figure 12.2: Case of three categories.

Example 12.9: Consider the data in Table 12.7. The response y is the amount of suspended solids in a coal cleansing system. The variable x is the pH of the system. Three different polymers are used in the system. Thus, "polymer" is categorical with three categories and hence produces two model terms. The model is given by

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_{1i} + \beta_3 z_{2i} + \epsilon_i, \qquad i = 1, 2, \ldots, 18.$$

Here we have

$$z_1 = \begin{cases} 1, & \text{for polymer 1,} \\ 0, & \text{otherwise,} \end{cases} \quad \text{and} \quad z_2 = \begin{cases} 1, & \text{for polymer 2,} \\ 0, & \text{otherwise.} \end{cases}$$

From the analysis in Figure 12.3, the following conclusions are drawn. The coefficient b_1 for pH is the estimate of the **common slope** that is assumed in the regression analysis. All model terms are statistically significant. Thus, pH and the nature of the polymer have an impact on the amount of cleansing. The signs and magnitudes of the coefficients of z_1 and z_2 indicate that polymer 1 is most effective (producing higher suspended solids) for cleansing, followed by polymer 2. Polymer 3 is least effective.

Table 12.7: Data for Example 12.9

x, (pH)	y, (amount of suspended solids)	Polymer
6.5	292	1
6.9	329	1
7.8	352	1
8.4	378	1
8.8	392	1
9.2	410	1
6.7	198	2
6.9	227	2
7.5	277	2
7.9	297	2
8.7	364	2
9.2	375	2
6.5	167	3
7.0	225	3
7.2	247	3
7.6	268	3
8.7	288	3
9.2	342	3

Slope May Vary with Indicator Categories

In the discussion given here, we have assumed that the indicator variable model terms enter the model in an additive fashion. This suggests that the slopes, as in Figure 12.2, are constant across categories. Obviously, this is not always going to be the case. We can account for the possibility of varying slopes and indeed test for this condition of **parallelism** by including product or **interaction** terms between indicator terms and continuous variables. For example, suppose a model with one continuous regressor and an indicator variable with two levels is chosen. The model is given by

$$y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz + \epsilon.$$

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	80181.73127	26727.24376	73.68	<.0001
Error	14	5078.71318	362.76523		
Corrected Total	17	85260.44444			

R-Square	Coeff Var	Root MSE	y Mean
0.940433	6.316049	19.04640	301.5556

Parameter	Estimate	Error	t Value	Standard Pr > |t|
Intercept	-161.8973333	37.43315576	-4.32	0.0007
x	54.2940260	4.75541126	11.42	<.0001
z1	89.9980606	11.05228237	8.14	<.0001
z2	27.1656970	11.01042883	2.47	0.0271

Figure 12.3: *SAS* printout for Example 12.9.

This model suggests that for category l ($z = 1$),

$$E(y) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x,$$

while for category 2 ($z = 0$),

$$E(y) = \beta_0 + \beta_1 x.$$

Thus, we allow for varying intercepts and slopes for the two categories. Figure 12.4 displays the regression lines with varying slopes for the two categories.

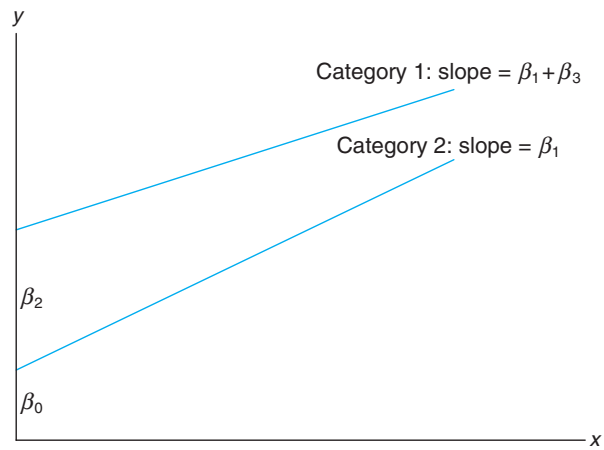


Figure 12.4: Nonparallelism in categorical variables.

In this case, β_0, β_1, and β_2 are positive while β_3 is negative with $|\beta_3| < \beta_1$. Obviously, if the interaction coefficient β_3 is insignificant, we are back to the common slope model.

Exercises

12.45 A study was done to assess the cost effectiveness of driving a four-door sedan instead of a van or an SUV (sports utility vehicle). The continuous variables are odometer reading and octane of the gasoline used. The response variable is miles per gallon. The data are presented here.

MPG	Car Type	Odometer	Octane
34.5	sedan	75,000	87.5
33.3	sedan	60,000	87.5
30.4	sedan	88,000	78.0
32.8	sedan	15,000	78.0
35.0	sedan	25,000	90.0
29.0	sedan	35,000	78.0
32.5	sedan	102,000	90.0
29.6	sedan	98,000	87.5
16.8	van	56,000	87.5
19.2	van	72,000	90.0
22.6	van	14,500	87.5
24.4	van	22,000	90.0
20.7	van	66,500	78.0
25.1	van	35,000	90.0
18.8	van	97,500	87.5
15.8	van	65,500	78.0
17.4	van	42,000	78.0
15.6	SUV	65,000	78.0
17.3	SUV	55,500	87.5
20.8	SUV	26,500	87.5
22.2	SUV	11,500	90.0
16.5	SUV	38,000	78.0
21.3	SUV	77,500	90.0
20.7	SUV	19,500	78.0
24.1	SUV	87,000	90.0

(a) Fit a linear regression model including two indicator variables. Use $(0, 0)$ to denote the four-door sedan.

(b) Which type of vehicle appears to get the best gas mileage?

(c) Discuss the difference between a van and an SUV in terms of gas mileage.

12.46 A study was done to determine whether the gender of the credit card holder was an important factor in generating profit for a certain credit card company. The variables considered were income, the number of family members, and the gender of the card holder. The data are as follows:

Profit	Income	Gender	Family Members
157	45,000	M	1
−181	55,000	M	2
−253	45,800	M	4
158	38,000	M	3
75	75,000	M	4
202	99,750	M	4
−451	28,000	M	1
146	39,000	M	2
89	54,350	M	1
−357	32,500	M	1
522	36,750	F	1
78	42,500	F	3
5	34,250	F	2
−177	36,750	F	3
123	24,500	F	2
251	27,500	F	1
−56	18,000	F	1
453	24,500	F	1
288	88,750	F	1
−104	19,750	F	2

(a) Fit a linear regression model using the variables available. Based on the fitted model, would the company prefer male or female customers?

(b) Would you say that income was an important factor in explaining the variability in profit?

12.9 Sequential Methods for Model Selection

At times, the significance tests outlined in Section 12.6 are quite adequate for determining which variables should be used in the final regression model. These tests are certainly effective if the experiment can be planned and the variables are orthogonal to each other. Even if the variables are not orthogonal, the individual t-tests can be of some use in many problems where the number of variables under investigation is small. However, there are many problems where it is necessary to use more elaborate techniques for screening variables, particularly when the experiment exhibits a substantial deviation from orthogonality. Useful measures of **multicollinearity** (linear dependency) among the independent variables are provided by the sample correlation coefficients $r_{x_i x_j}$. Since we are concerned only

with linear dependency among independent variables, no confusion will result if we drop the x's from our notation and simply write $r_{x_i x_j} = r_{ij}$, where

$$r_{ij} = \frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}}.$$

Note that the r_{ij} do not give true estimates of population correlation coefficients in the strict sense, since the x's are actually not random variables in the context discussed here. Thus, the term *correlation*, although standard, is perhaps a misnomer.

When one or more of these sample correlation coefficients deviate substantially from zero, it can be quite difficult to find the most effective subset of variables for inclusion in our prediction equation. In fact, for some problems the multicollinearity will be so extreme that a suitable predictor cannot be found unless all possible subsets of the variables are investigated. Informative discussions of model selection in regression by Hocking (1976) are cited in the Bibliography. Procedures for detection of multicollinearity are discussed in the textbook by Myers (1990), also cited.

The user of multiple linear regression attempts to accomplish one of three objectives:

1. Obtain estimates of individual coefficients in a complete model.
2. Screen variables to determine which have a significant effect on the response.
3. Arrive at the most effective prediction equation.

In (1) it is known a priori that all variables are to be included in the model. In (2) prediction is secondary, while in (3) individual regression coefficients are not as important as the quality of the estimated response $\hat{y}$. For each of the situations above, multicollinearity in the experiment can have a profound effect on the success of the regression.

In this section, some standard sequential procedures for selecting variables are discussed. They are based on the notion that a single variable or a collection of variables should not appear in the estimating equation unless the variables result in a significant increase in the regression sum of squares or, equivalently, a significant increase in R^2, the coefficient of multiple determination.

Illustration of Variable Screening in the Presence of Collinearity

Example 12.10: Consider the data of Table 12.8, where measurements were taken for nine infants. The purpose of the experiment was to arrive at a suitable estimating equation relating the length of an infant to all or a subset of the independent variables. The sample correlation coefficients, indicating the linear dependency among the independent variables, are displayed in the symmetric matrix

$$\begin{array}{c} \begin{array}{cccc} x_1 & x_2 & x_3 & x_4 \end{array} \\ \begin{bmatrix} 1.0000 & 0.9523 & 0.5340 & 0.3900 \\ 0.9523 & 1.0000 & 0.2626 & 0.1549 \\ 0.5340 & 0.2626 & 1.0000 & 0.7847 \\ 0.3900 & 0.1549 & 0.7847 & 1.0000 \end{bmatrix} \end{array}$$

Table 12.8: Data Relating to Infant Length*

Infant Length, y (cm)	**Age, x_1 (days)**	**Length at Birth, x_2 (cm)**	**Weight at Birth, x_3 (kg)**	**Chest Size at Birth, x_4 (cm)**
57.5	78	48.2	2.75	29.5
52.8	69	45.5	2.15	26.3
61.3	77	46.3	4.41	32.2
67.0	88	49.0	5.52	36.5
53.5	67	43.0	3.21	27.2
62.7	80	48.0	4.32	27.7
56.2	74	48.0	2.31	28.3
68.5	94	53.0	4.30	30.3
69.2	102	58.0	3.71	28.7

*Data analyzed by the Statistical Consulting Center, Virginia Tech, Blacksburg, Virginia.

Note that there appears to be an appreciable amount of multicollinearity. Using the least squares technique outlined in Section 12.2, the estimated regression equation was fitted using the complete model and is

$$\hat{y} = 7.1475 + 0.1000x_1 + 0.7264x_2 + 3.0758x_3 - 0.0300x_4.$$

The value of s^2 with 4 degrees of freedom is 0.7414, and the value for the coefficient of determination for this model is found to be 0.9908. Regression sums of squares, measuring the variation attributed to each individual variable in the presence of the others, and the corresponding t-values are given in Table 12.9.

Table 12.9: t-Values for the Regression Data of Table 12.8

Variable x_1	**Variable x_2**	**Variable x_3**	**Variable x_4**
$R(\beta_1 \mid \beta_2, \beta_3, \beta_4)$	$R(\beta_2 \mid \beta_1, \beta_3, \beta_4)$	$R(\beta_3 \mid \beta_1, \beta_2, \beta_4)$	$R(\beta_4 \mid \beta_1, \beta_2, \beta_3)$
$= 0.0644$	$= 0.6334$	$= 6.2523$	$= 0.0241$
$t = 0.2947$	$t = 0.9243$	$t = 2.9040$	$t = -0.1805$

A two-tailed critical region with 4 degrees of freedom at the 0.05 level of significance is given by $|t| > 2.776$. Of the four computed t-values, **only variable x_3 appears to be significant**. However, recall that although the t-statistic described in Section 12.6 measures the worth of a variable adjusted for all other variables, it does not detect the potential importance of a variable in combination with a subset of the variables. For example, consider the model with only the variables x_2 and x_3 in the equation. The data analysis gives the regression function

$$\hat{y} = 2.1833 + 0.9576x_2 + 3.3253x_3,$$

with $R^2 = 0.9905$, certainly not a substantial reduction from $R^2 = 0.9907$ for the complete model. However, unless the performance characteristics of this particular combination had been observed, one would not be aware of its predictive potential. This, of course, lends support for a methodology that observes *all possible regressions* or a systematic sequential procedure designed to test subsets.

Stepwise Regression

One standard procedure for searching for the "optimum subset" of variables in the absence of orthogonality is a technique called **stepwise regression**. It is based on the procedure of sequentially introducing the variables into the model one at a time. Given a predetermined size α, the description of the stepwise routine will be better understood if the methods of **forward selection** and **backward elimination** are described first.

Forward selection is based on the notion that variables should be inserted one at a time until a satisfactory regression equation is found. The procedure is as follows:

STEP 1. Choose the variable that gives the largest regression sum of squares when performing a simple linear regression with y or, equivalently, that which gives the largest value of R^2. We shall call this initial variable x_1. If x_1 is insignificant, the procedure is terminated.

STEP 2. Choose the variable that, when inserted in the model, gives the largest increase in R^2, in the presence of x_1, over the R^2 found in step 1. This, of course, is the variable x_j for which

$$R(\beta_j|\beta_1) = R(\beta_1, \beta_j) - R(\beta_1)$$

is largest. Let us call this variable x_2. The regression model with x_1 and x_2 is then fitted and R^2 observed. If x_2 is insignificant, the procedure is terminated.

STEP 3. Choose the variable x_j that gives the largest value of

$$R(\beta_j \mid \beta_1, \beta_2) = R(\beta_1, \beta_2, \beta_j) - R(\beta_1, \beta_2),$$

again resulting in the largest increase of R^2 over that given in step 2. Calling this variable x_3, we now have a regression model involving x_1, x_2, and x_3. If x_3 is insignificant, the procedure is terminated.

This process is continued until the most recent variable inserted fails to induce a significant increase in the explained regression. Such an increase can be determined at each step by using the appropriate partial F-test or t-test. For example, in step 2 the value

$$f = \frac{R(\beta_2|\beta_1)}{s^2}$$

can be determined to test the appropriateness of x_2 in the model. Here the value of s^2 is the mean square error for the model containing the variables x_1 and x_2. Similarly, in step 3 the ratio

$$f = \frac{R(\beta_3 \mid \beta_1, \beta_2)}{s^2}$$

tests the appropriateness of x_3 in the model. Now, however, the value for s^2 is the mean square error for the model that contains the three variables x_1, x_2, and x_3. If $f < f_\alpha(1, n-3)$ at step 2, for a prechosen significance level, x_2 is not included

and the process is terminated, resulting in a simple linear equation relating y and x_1. However, if $f > f_\alpha(1, n-3)$, we proceed to step 3. Again, if $f < f_\alpha(1, n-4)$ at step 3, x_3 is not included and the process is terminated with the appropriate regression equation containing the variables x_1 and x_2.

Backward elimination involves the same concepts as forward selection except that one begins with all the variables in the model. Suppose, for example, that there are five variables under consideration. The steps are as follows:

STEP 1. Fit a regression equation with all five variables included in the model. Choose the variable that gives the smallest value of the regression sum of squares **adjusted for the others**. Suppose that this variable is x_2. Remove x_2 from the model if

$$f = \frac{R(\beta_2 \mid \beta_1, \beta_3, \beta_4, \beta_5)}{s^2}$$

is insignificant.

STEP 2. Fit a regression equation using the remaining variables x_1, x_3, x_4, and x_5, and repeat step 1. Suppose that variable x_5 is chosen this time. Once again, if

$$f = \frac{R(\beta_5 \mid \beta_1, \beta_3, \beta_4)}{s^2}$$

is insignificant, the variable x_5 is removed from the model. At each step, the s^2 used in the F-test is the mean square error for the regression model at that stage.

This process is repeated until at some step the variable with the smallest adjusted regression sum of squares results in a significant f-value for some predetermined significance level.

Stepwise regression is accomplished with a slight but important modification of the forward selection procedure. The modification involves further testing at each stage to ensure the continued effectiveness of variables that had been inserted into the model at an earlier stage. This represents an improvement over forward selection, since it is quite possible that a variable entering the regression equation at an early stage might have been rendered unimportant or redundant because of relationships that exist between it and other variables entering at later stages. Therefore, at a stage in which a new variable has been entered into the regression equation through a significant increase in R^2 as determined by the F-test, all the variables already in the model are subjected to F-tests (or, equivalently, to t-tests) in light of this new variable and are deleted if they do not display a significant f-value. The procedure is continued until a stage is reached where no additional variables can be inserted or deleted. We illustrate the stepwise procedure in the following example.

Example 12.11: Using the techniques of stepwise regression, find an appropriate linear regression model for predicting the length of infants for the data of Table 12.8.

***Solution*:** STEP 1. Considering each variable separately, four individual simple linear regression equations are fitted. The following pertinent regression sums of

squares are computed:

$$R(\beta_1) = 288.1468, \qquad R(\beta_2) = 215.3013,$$
$$R(\beta_3) = 186.1065, \qquad R(\beta_4) = 100.8594.$$

Variable x_1 clearly gives the largest regression sum of squares. The mean square error for the equation involving only x_1 is $s^2 = 4.7276$, and since

$$f = \frac{R(\beta_1)}{s^2} = \frac{288.1468}{4.7276} = 60.9500,$$

which exceeds $f_{0.05}(1, 7) = 5.59$, the variable x_1 is significant and is entered into the model.

STEP 2. Three regression equations are fitted at this stage, all containing x_1. The important results for the combinations (x_1, x_2), (x_1, x_3), and (x_1, x_4) are

$$R(\beta_2|\beta_1) = 23.8703, \quad R(\beta_3|\beta_1) = 29.3086, \quad R(\beta_4|\beta_1) = 13.8178.$$

Variable x_3 displays the largest regression sum of squares in the presence of x_1. The regression involving x_1 and x_3 gives a new value of $s^2 = 0.6307$, and since

$$f = \frac{R(\beta_3|\beta_1)}{s^2} = \frac{29.3086}{0.6307} = 46.47,$$

which exceeds $f_{0.05}(1, 6) = 5.99$, the variable x_3 is significant and is included along with x_1 in the model. Now we must subject x_1 in the presence of x_3 to a significance test. We find that $R(\beta_1 \mid \beta_3) = 131.349$, and hence

$$f = \frac{R(\beta_1|\beta_3)}{s^2} = \frac{131.349}{0.6307} = 208.26,$$

which is highly significant. Therefore, x_1 is retained along with x_3.

STEP 3. With x_1 and x_3 already in the model, we now require $R(\beta_2 \mid \beta_1, \beta_3)$ and $R(\beta_4 \mid \beta_1, \beta_3)$ in order to determine which, if any, of the remaining two variables is entered at this stage. From the regression analysis using x_2 along with x_1 and x_3, we find $R(\beta_2 \mid \beta_1, \beta_3) = 0.7948$, and when x_4 is used along with x_1 and x_3, we obtain $R(\beta_4 \mid \beta_1, \beta_3) = 0.1855$. The value of s^2 is 0.5979 for the (x_1, x_2, x_3) combination and 0.7198 for the (x_1, x_2, x_4) combination. Since neither f-value is significant at the $\alpha = 0.05$ level, the final regression model includes only the variables x_1 and x_3. The estimating equation is found to be

$$\hat{y} = 20.1084 + 0.4136x_1 + 2.0253x_3,$$

and the coefficient of determination for this model is $R^2 = 0.9882$.

Although (x_1, x_3) is the combination chosen by stepwise regression, it is not necessarily the combination of two variables that gives the largest value of R^2. In fact, we have already observed that the combination (x_2, x_3) gives $R^2 = 0.9905$. Of course, the stepwise procedure never observed this combination. A rational argument could be made that there is actually a negligible difference in performance

between these two estimating equations, at least in terms of percent variation explained. It is interesting to observe, however, that the backward elimination procedure gives the combination (x_2, x_3) in the final equation (see Exercise 12.49 on page 494).

Summary

The main function of each of the procedures explained in this section is to expose the variables to a systematic methodology designed to ensure the eventual inclusion of the best combinations of the variables. Obviously, there is no assurance that this will happen in all problems, and, of course, it is possible that the multicollinearity is so extensive that one has no alternative but to resort to estimation procedures other than least squares. These estimation procedures are discussed in Myers (1990), listed in the Bibliography.

The sequential procedures discussed here represent three of many such methods that have been put forth in the literature and appear in various regression computer packages that are available. These methods are designed to be computationally efficient but, of course, do not give results for all possible subsets of the variables. As a result, the procedures are most effective for data sets that involve a **large number of variables**. For regression problems involving a relatively small number of variables, modern regression computer packages allow for the computation and summarization of quantitative information on all models for every possible subset of the variables. Illustrations are provided in Section 12.11.

Choice of *P*-Values

As one might expect, the choice of the final model with these procedures may depend dramatically on what P-value is chosen. In addition, a procedure is most successful when it is forced to test a large number of candidate variables. For this reason, any forward procedure will be most useful when a relatively large P-value is used. Thus, some software packages use a default P-value of 0.50.

12.10 Study of Residuals and Violation of Assumptions (Model Checking)

It was suggested earlier in this chapter that the residuals, or errors in the regression fit, often carry information that can be very informative to the data analyst. The $e_i = y_i - \hat{y}_i$, $i = 1, 2, \ldots, n$, which are the numerical counterpart to the ϵ_i, the model errors, often shed light on the possible violation of assumptions or the presence of "suspect" data points. Suppose that we let the vector $\mathbf{x}_i$ denote the values of the regressor variables corresponding to the ith data point, supplemented by a 1 in the initial position. That is,

$$\mathbf{x}_i' = [1, x_{1i}, x_{2i}, \ldots, x_{ki}].$$

Consider the quantity

$$h_{ii} = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i, \qquad i = 1, 2, \ldots, n.$$

Chapter 13

One-Factor Experiments: General

13.1 Analysis-of-Variance Technique

In the estimation and hypothesis testing material covered in Chapters 9 and 10, we were restricted in each case to considering no more than two population parameters. Such was the case, for example, in testing for the equality of two population means using independent samples from normal populations with common but unknown variance, where it was necessary to obtain a pooled estimate of σ^2.

This material dealing in two-sample inference represents a special case of what we call the *one-factor problem*. For example, in Exercise 10.35 on page 357, the survival time was measured for two samples of mice, where one sample received a new serum for leukemia treatment and the other sample received no treatment. In this case, we say that there is *one factor*, namely *treatment*, and the factor is at *two levels*. If several competing treatments were being used in the sampling process, more samples of mice would be necessary. In this case, the problem would involve one factor with more than two levels and thus more than two samples.

In the $k > 2$ sample problem, it will be assumed that there are k samples from k populations. One very common procedure used to deal with testing population means is called the **analysis of variance**, or **ANOVA**.

The analysis of variance is certainly not a new technique to the reader who has followed the material on regression theory. We used the analysis-of-variance approach to partition the total sum of squares into a portion due to regression and a portion due to error.

Suppose in an industrial experiment that an engineer is interested in how the mean absorption of moisture in concrete varies among 5 different concrete aggregates. The samples are exposed to moisture for 48 hours. It is decided that 6 samples are to be tested for each aggregate, requiring a total of 30 samples to be tested. The data are recorded in Table 13.1.

The model for this situation may be set up as follows. There are 6 observations taken from each of 5 populations with means $\mu_1, \mu_2, \ldots, \mu_5$, respectively. We may wish to test

$$H_0: \ \mu_1 = \mu_2 = \cdots = \mu_5,$$
$$H_1: \ \text{At least two of the means are not equal.}$$

Table 13.1: Absorption of Moisture in Concrete Aggregates

Aggregate:	**1**	**2**	**3**	**4**	**5**	
	551	595	639	417	563	
	457	580	615	449	631	
	450	508	511	517	522	
	731	583	573	438	613	
	499	633	648	415	656	
	632	517	677	555	679	
Total	3320	3416	3663	2791	3664	16,854
Mean	553.33	569.33	610.50	465.17	610.67	561.80

In addition, we may be interested in making individual comparisons among these 5 population means.

Two Sources of Variability in the Data

In the analysis-of-variance procedure, it is assumed that whatever variation exists among the aggregate averages is attributed to (1) variation in absorption among observations *within* aggregate types and (2) variation *among* aggregate types, that is, due to differences in the chemical composition of the aggregates. The **within-aggregate variation** is, of course, brought about by various causes. Perhaps humidity and temperature conditions were not kept entirely constant throughout the experiment. It is possible that there was a certain amount of heterogeneity in the batches of raw materials that were used. At any rate, we shall consider the within-sample variation to be **chance or random variation**. Part of the goal of the analysis of variance is to determine if the differences among the 5 sample means are what we would expect due to random variation alone or, rather, due to variation beyond merely random effects, i.e., differences in the chemical composition of the aggregates.

Many pointed questions appear at this stage concerning the preceding problem. For example, how many samples must be tested for each aggregate? This is a question that continually haunts the practitioner. In addition, what if the within-sample variation is so large that it is difficult for a statistical procedure to detect the systematic differences? Can we systematically control extraneous sources of variation and thus remove them from the portion we call random variation? We shall attempt to answer these and other questions in the following sections.

13.2 The Strategy of Experimental Design

In Chapters 9 and 10, the notions of estimation and testing for the two-sample case were covered under the important backdrop of the way the experiment is conducted. This falls into the broad category of design of experiments. For example, for the **pooled *t*-test** discussed in Chapter 10, it is assumed that the factor levels (treatments in the mice example) are assigned randomly to the experimental units (mice). The notion of experimental units was discussed in Chapters 9 and 10 and

illustrated through examples. Simply put, experimental units are the units (mice, patients, concrete specimens, time) that **provide the heterogeneity that leads to experimental error** in a scientific investigation. The random assignment eliminates bias that could result with systematic assignment. The goal is to distribute uniformly among the factor levels the risks brought about by the heterogeneity of the experimental units. Random assignment best simulates the conditions that are assumed by the model. In Section 13.7, we discuss **blocking** in experiments. The notion of blocking was presented in Chapters 9 and 10, when comparisons between means were accomplished with **pairing**, that is, the division of the experimental units into homogeneous pairs called **blocks**. The factor levels or treatments are then assigned randomly within blocks. The purpose of blocking is to reduce the effective experimental error. In this chapter, we naturally extend the pairing to larger block sizes, with analysis of variance being the primary analytical tool.

13.3 One-Way Analysis of Variance: Completely Randomized Design (One-Way ANOVA)

Random samples of size n are selected from each of k populations. The k different populations are classified on the basis of a single criterion such as different treatments or groups. Today the term **treatment** is used generally to refer to the various classifications, whether they be different aggregates, different analysts, different fertilizers, or different regions of the country.

Assumptions and Hypotheses in One-Way ANOVA

It is assumed that the k populations are independent and normally distributed with means $\mu_1, \mu_2, \ldots, \mu_k$ and common variance σ^2. As indicated in Section 13.2, these assumptions are made more palatable by randomization. We wish to derive appropriate methods for testing the hypothesis

$$H_0: \ \mu_1 = \mu_2 = \cdots = \mu_k,$$
$$H_1: \ \text{At least two of the means are not equal.}$$

Let y_{ij} denote the jth observation from the ith treatment and arrange the data as in Table 13.2. Here, $Y_{i.}$ is the total of all observations in the sample from the ith treatment, $\bar{y}_{i.}$ is the mean of all observations in the sample from the ith treatment, $Y_{..}$ is the total of all nk observations, and $\bar{y}_{..}$ is the mean of all nk observations.

Model for One-Way ANOVA

Each observation may be written in the form

$$Y_{ij} = \mu_i + \epsilon_{ij},$$

where ϵ_{ij} measures the deviation of the jth observation of the ith sample from the corresponding treatment mean. The ϵ_{ij}-term represents random error and plays the same role as the error terms in the regression models. An alternative and

Table 13.2: k Random Samples

Treatment:	**1**	**2**	$\cdots$	i	$\cdots$	k	
	y_{11}	y_{21}	$\cdots$	y_{i1}	$\cdots$	y_{k1}	
	y_{12}	y_{22}	$\cdots$	y_{i2}	$\cdots$	y_{k2}	
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	
	y_{1n}	y_{2n}	$\cdots$	y_{in}	$\cdots$	y_{kn}	
Total	$Y_{1.}$	$Y_{2.}$	$\cdots$	$Y_{i.}$	$\cdots$	$Y_{k.}$	$Y_{..}$
Mean	$\bar{y}_{1.}$	$\bar{y}_{2.}$	$\cdots$	$\bar{y}_{i.}$	$\cdots$	$\bar{y}_{k.}$	$\bar{y}_{..}$

preferred form of this equation is obtained by substituting $\mu_i = \mu + \alpha_i$, subject to the constraint $\sum_{i=1}^{k} \alpha_i = 0$. Hence, we may write

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

where μ is just the **grand mean** of all the μ_i, that is,

$$\mu = \frac{1}{k}\sum_{i=1}^{k} \mu_i,$$

and α_i is called the **effect** of the ith treatment.

The null hypothesis that the k population means are equal against the alternative that at least two of the means are unequal may now be replaced by the equivalent hypothesis

H_0: $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$,

H_1: At least one of the α_i is not equal to zero.

Resolution of Total Variability into Components

Our test will be based on a comparison of two independent estimates of the common population variance σ^2. These estimates will be obtained by partitioning the total variability of our data, designated by the double summation

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2,$$

into two components.

Theorem 13.1: **Sum-of-Squares Identity**

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2 = n\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2$$

It will be convenient in what follows to identify the terms of the sum-of-squares identity by the following notation:

Three Important Measures of Variability

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2 = \text{total sum of squares},$$

$$SSA = n\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{y}_{..})^2 = \text{treatment sum of squares},$$

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2 = \text{error sum of squares}.$$

The sum-of-squares identity can then be represented symbolically by the equation

$$SST = SSA + SSE.$$

The identity above expresses how between-treatment and within-treatment variation add to the total sum of squares. However, much insight can be gained by investigating the **expected value of both *SSA* and *SSE***. Eventually, we shall develop variance estimates that formulate the ratio to be used to test the equality of population means.

Theorem 13.2:

$$E(SSA) = (k-1)\sigma^2 + n\sum_{i=1}^{k}\alpha_i^2$$

The proof of the theorem is left as an exercise (see Review Exercise 13.53 on page 556).

If H_0 is true, an estimate of σ^2, based on $k-1$ degrees of freedom, is provided by this expression:

Treatment Mean Square

$$s_1^2 = \frac{SSA}{k-1}$$

If H_0 is true and thus each α_i in Theorem 13.2 is equal to zero, we see that

$$E\left(\frac{SSA}{k-1}\right) = \sigma^2,$$

and s_1^2 is an unbiased estimate of σ^2. However, if H_1 is true, we have

$$E\left(\frac{SSA}{k-1}\right) = \sigma^2 + \frac{n}{k-1}\sum_{i=1}^{k}\alpha_i^2,$$

and s_1^2 estimates σ^2 plus an additional term, which measures variation due to the systematic effects.

A second and independent estimate of σ^2, based on $k(n-1)$ degrees of freedom, is this familiar formula:

Error Mean Square

$$s^2 = \frac{SSE}{k(n-1)}$$

It is instructive to point out the importance of the expected values of the mean squares indicated above. In the next section, we discuss the use of an ***F*-ratio** with the treatment mean square residing in the numerator. It turns out that when H_1 is true, the presence of the condition $E(s_1^2) > E(s^2)$ suggests that the F-ratio be used in the context of a **one-sided upper-tailed test**. That is, when H_1 is true, we would expect the numerator s_1^2 to exceed the denominator.

Use of *F*-Test in ANOVA

The estimate s^2 is unbiased regardless of the truth or falsity of the null hypothesis (see Review Exercise 13.52 on page 556). It is important to note that the sum-of-squares identity has partitioned not only the total variability of the data, but also the total number of degrees of freedom. That is,

$$nk - 1 = k - 1 + k(n - 1).$$

F-Ratio for Testing Equality of Means

When H_0 is true, the ratio $f = s_1^2/s^2$ is a value of the random variable F having the F-distribution with $k-1$ and $k(n-1)$ degrees of freedom (see Theorem 8.8). Since s_1^2 overestimates σ^2 when H_0 is false, we have a one-tailed test with the critical region entirely in the right tail of the distribution.

The null hypothesis H_0 is rejected at the α-level of significance when

$$f > f_\alpha[k - 1, k(n - 1)].$$

Another approach, the P-value approach, suggests that the evidence in favor of or against H_0 is

$$P = P\{f[k - 1, k(n - 1)] > f\}.$$

The computations for an analysis-of-variance problem are usually summarized in tabular form, as shown in Table 13.3.

Table 13.3: Analysis of Variance for the One-Way ANOVA

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	**Computed *f***
Treatments	SSA	$k - 1$	$s_1^2 = \dfrac{SSA}{k-1}$	$\dfrac{s_1^2}{s^2}$
Error	SSE	$k(n - 1)$	$s^2 = \dfrac{SSE}{k(n-1)}$	
Total	SST	$kn - 1$		

Example 13.1: Test the hypothesis $\mu_1 = \mu_2 = \cdots = \mu_5$ at the 0.05 level of significance for the data of Table 13.1 on absorption of moisture by various types of cement aggregates.

***Solution*:** The hypotheses are

$$H_0\text{: } \mu_1 = \mu_2 = \cdots = \mu_5,$$
$$H_1\text{: At least two of the means are not equal.}$$
$$\alpha = 0.05.$$

Critical region: $f > 2.76$ with $v_1 = 4$ and $v_2 = 25$ degrees of freedom. The sum-of-squares computations give

$$SST = 209{,}377, \quad SSA = 85{,}356,$$
$$SSE = 209{,}377 - 85{,}356 = 124{,}021.$$

These results and the remaining computations are exhibited in Figure 13.1 in the *SAS* ANOVA procedure.

```
                        The GLM Procedure
Dependent Variable: moisture

                                                Sum of
Source                     DF        Squares     Mean Square    F Value    Pr > F
Model                       4     85356.4667      21339.1167       4.30    0.0088
Error                      25    124020.3333       4960.8133
Corrected Total            29    209376.8000

  R-Square     Coeff Var      Root MSE    moisture Mean
  0.407669      12.53703      70.43304         561.8000

Source                     DF       Type I SS     Mean Square    F Value    Pr > F
aggregate                   4     85356.46667     21339.11667       4.30    0.0088
```

Figure 13.1: *SAS* output for the analysis-of-variance procedure.

Decision: Reject H_0 and conclude that the aggregates do not have the same mean absorption. The P-value for $f = 4.30$ is 0.0088, which is smaller than 0.05. ∎

In addition to the ANOVA, a box plot was constructed for each aggregate. The plots are shown in Figure 13.2. From these plots it is evident that the absorption is not the same for all aggregates. In fact, it appears as if aggregate 4 stands out from the rest. A more formal analysis showing this result will appear in Exercise 13.21 on page 531.

During experimental work, one often loses some of the desired observations. Experimental animals may die, experimental material may be damaged, or human subjects may drop out of a study. The previous analysis for equal sample size will still be valid if we slightly modify the sum of squares formulas. We now assume the k random samples to be of sizes $n_1, n_2, \ldots, n_k$, respectively.

Sum of Squares, Unequal Sample Sizes

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2, \quad SSA = \sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2, \quad SSE = SST - SSA$$

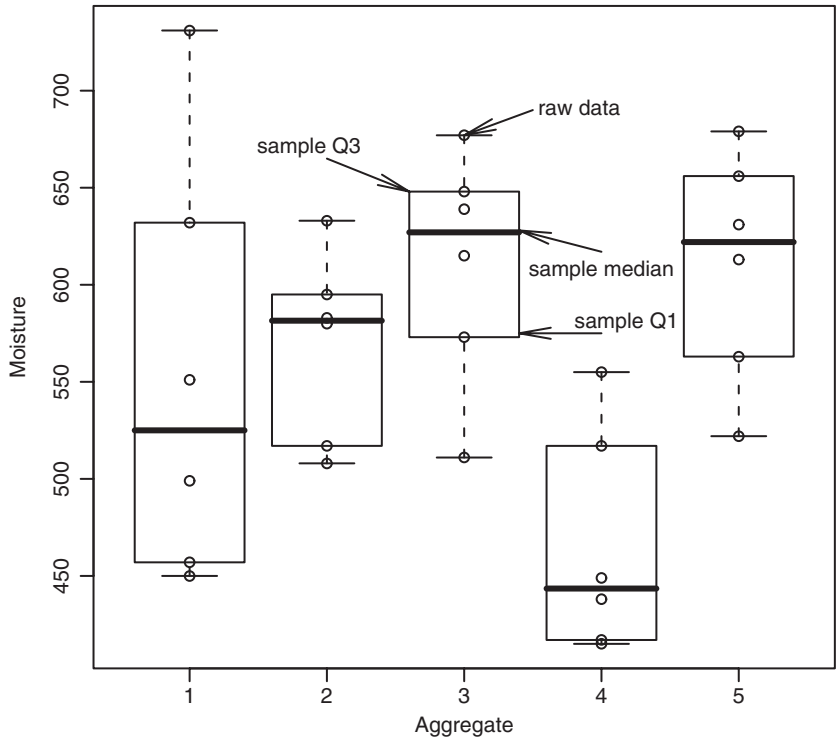


Figure 13.2: Box plots for the absorption of moisture in concrete aggregates.

The degrees of freedom are then partitioned as before: $N-1$ for SST, $k-1$ for SSA, and $N-1-(k-1)=N-k$ for SSE, where $N=\sum_{i=1}^{k} n_i$.

Example 13.2: Part of a study conducted at Virginia Tech was designed to measure serum alkaline phosphatase activity levels (in Bessey-Lowry units) in children with seizure disorders who were receiving anticonvulsant therapy under the care of a private physician. Forty-five subjects were found for the study and categorized into four drug groups:

G-1: Control (not receiving anticonvulsants and having no history of seizure disorders)

G-2: Phenobarbital

G-3: Carbamazepine

G-4: Other anticonvulsants

From blood samples collected from each subject, the serum alkaline phosphatase activity level was determined and recorded as shown in Table 13.4. Test the hypothesis at the 0.05 level of significance that the average serum alkaline phosphatase activity level is the same for the four drug groups.

Table 13.4: Serum Alkaline Phosphatase Activity Level

G-1		G-2	G-3	G-4
49.20	97.50	97.07	62.10	110.60
44.54	105.00	73.40	94.95	57.10
45.80	58.05	68.50	142.50	117.60
95.84	86.60	91.85	53.00	77.71
30.10	58.35	106.60	175.00	150.00
36.50	72.80	0.57	79.50	82.90
82.30	116.70	0.79	29.50	111.50
87.85	45.15	0.77	78.40	
105.00	70.35	0.81	127.50	
95.22	77.40			

***Solution*:** With the level of significance at 0.05, the hypotheses are

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$,
H_1: At least two of the means are not equal.

Critical region: $f > 2.836$, from interpolating in Table A.6.
Computations: $Y_{1.} = 1460.25$, $Y_{2.} = 440.36$, $Y_{3.} = 842.45$, $Y_{4.} = 707.41$, and $Y_{..} = 3450.47$. The analysis of variance is shown in the *MINITAB* output of Figure 13.3.

```
One-way ANOVA: G-1, G-2, G-3, G-4

Source  DF     SS    MS     F      P
Factor   3  13939  4646  3.57  0.022
Error   41  53376  1302
Total   44  67315

S = 36.08   R-Sq = 20.71%   R-Sq(adj) = 14.90%

                          Individual 95% CIs For Mean Based on
                          Pooled StDev
Level   N    Mean  StDev  --+---------+---------+---------+-------
G-1    20   73.01  25.75             (----*-----)
G-2     9   48.93  47.11  (-------*-------)
G-3     9   93.61  46.57                  (-------*-------)
G-4     7  101.06  30.76                    (--------*--------)
                          --+---------+---------+---------+-------
                          30        60        90       120

Pooled StDev = 36.08
```

Figure 13.3: *MINITAB* analysis of data in Table 13.4.

Decision: Reject H_0 and conclude that the average serum alkaline phosphatase activity levels for the four drug groups are not all the same. The calculated P-value is 0.022. ∎

In concluding our discussion on the analysis of variance for the one-way classification, we state the advantages of choosing equal sample sizes over the choice of unequal sample sizes. The first advantage is that the f-ratio is insensitive to slight departures from the assumption of equal variances for the k populations when the samples are of equal size. Second, the choice of equal sample sizes minimizes the probability of committing a type II error.

13.4 Tests for the Equality of Several Variances

Although the f-ratio obtained from the analysis-of-variance procedure is insensitive to departures from the assumption of equal variances for the k normal populations when the samples are of equal size, we may still prefer to exercise caution and run a preliminary test for homogeneity of variances. Such a test would certainly be advisable in the case of unequal sample sizes if there was a reasonable doubt concerning the homogeneity of the population variances. Suppose, therefore, that we wish to test the null hypothesis

$$H_0\colon\ \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$$

against the alternative

$$H_1\colon\ \text{The variances are not all equal.}$$

The test that we shall use, called **Bartlett's test**, is based on a statistic whose sampling distribution provides exact critical values when the sample sizes are equal. These critical values for equal sample sizes can also be used to yield highly accurate approximations to the critical values for unequal sample sizes.

First, we compute the k sample variances $s_1^2, s_2^2, \ldots, s_k^2$ from samples of size $n_1, n_2, \ldots, n_k$, with $\sum_{i=1}^{k} n_i = N$. Second, we combine the sample variances to give the pooled estimate

$$s_p^2 = \frac{1}{N-k}\sum_{i=1}^{k}(n_i - 1)s_i^2.$$

Now

$$b = \frac{[(s_1^2)^{n_1-1}(s_2^2)^{n_2-1}\cdots(s_k^2)^{n_k-1}]^{1/(N-k)}}{s_p^2}$$

is a value of a random variable B having the **Bartlett distribution**. For the special case where $n_1 = n_2 = \cdots = n_k = n$, we reject H_0 at the α-level of significance if

$$b < b_k(\alpha; n),$$

where $b_k(\alpha; n)$ is the critical value leaving an area of size α in the left tail of the Bartlett distribution. Table A.10 gives the critical values, $b_k(\alpha; n)$, for $\alpha = 0.01$ and 0.05; $k = 2, 3, \ldots, 10$; and selected values of n from 3 to 100.

When the sample sizes are unequal, the null hypothesis is rejected at the α-level of significance if

$$b < b_k(\alpha; n_1, n_2, \ldots, n_k),$$

where

$$b_k(\alpha; n_1, n_2, \ldots, n_k) \approx \frac{n_1 b_k(\alpha; n_1) + n_2 b_k(\alpha; n_2) + \cdots + n_k b_k(\alpha; n_k)}{N}.$$

As before, all the $b_k(\alpha; n_i)$ for sample sizes $n_1, n_2, \ldots, n_k$ are obtained from Table A.10.

Example 13.3: Use Bartlett's test to test the hypothesis at the 0.01 level of significance that the population variances of the four drug groups of Example 13.2 are equal.

***Solution*:** We have the hypotheses

$$H_0: \ \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2,$$
$$H_1: \ \text{The variances are not equal,}$$
$$\text{with } \alpha = 0.01.$$

Critical region: Referring to Example 13.2, we have $n_1 = 20$, $n_2 = 9$, $n_3 = 9$, $n_4 = 7$, $N = 45$, and $k = 4$. Therefore, we reject when

$$\begin{aligned} b &< b_4(0.01; 20, 9, 9, 7) \\ &\approx \frac{(20)(0.8586) + (9)(0.6892) + (9)(0.6892) + (7)(0.6045)}{45} \\ &= 0.7513. \end{aligned}$$

Computations: First compute

$$s_1^2 = 662.862, \ s_2^2 = 2219.781, \ s_3^2 = 2168.434, \ s_4^2 = 946.032,$$

and then

$$\begin{aligned} s_p^2 &= \frac{(19)(662.862) + (8)(2219.781) + (8)(2168.434) + (6)(946.032)}{41} \\ &= 1301.861. \end{aligned}$$

Now

$$b = \frac{[(662.862)^{19}(2219.781)^8(2168.434)^8(946.032)^6]^{1/41}}{1301.861} = 0.8557.$$

Decision: Do not reject the hypothesis, and conclude that the population variances of the four drug groups are not significantly different. ∎

Although Bartlett's test is most often used for testing of homogeneity of variances, other methods are available. A method due to Cochran provides a computationally simple procedure, but it is restricted to situations in which the sample

sizes are equal. **Cochran's test** is particularly useful for detecting if one variance is much larger than the others. The statistic that is used is

$$G = \frac{\text{largest } S_i^2}{\sum_{i=1}^{k} S_i^2},$$

and the hypothesis of equality of variances is rejected if $g > g_\alpha$, where the value of g_α is obtained from Table A.11.

To illustrate Cochran's test, let us refer again to the data of Table 13.1 on moisture absorption in concrete aggregates. Were we justified in assuming equal variances when we performed the analysis of variance in Example 13.1? We find that

$$s_1^2 = 12{,}134,\ s_2^2 = 2303,\ s_3^2 = 3594,\ s_4^2 = 3319,\ s_5^2 = 3455.$$

Therefore,

$$g = \frac{12{,}134}{24{,}805} = 0.4892,$$

which does not exceed the table value $g_{0.05} = 0.5065$. Hence, we conclude that the assumption of equal variances is reasonable.

Exercises

13.1 Six different machines are being considered for use in manufacturing rubber seals. The machines are being compared with respect to tensile strength of the product. A random sample of four seals from each machine is used to determine whether the mean tensile strength varies from machine to machine. The following are the tensile-strength measurements in kilograms per square centimeter $\times 10^{-1}$:

Machine					
1	**2**	**3**	**4**	**5**	**6**
17.5	16.4	20.3	14.6	17.5	18.3
16.9	19.2	15.7	16.7	19.2	16.2
15.8	17.7	17.8	20.8	16.5	17.5
18.6	15.4	18.9	18.9	20.5	20.1

Perform the analysis of variance at the 0.05 level of significance and indicate whether or not the mean tensile strengths differ significantly for the six machines.

13.2 The data in the following table represent the number of hours of relief provided by five different brands of headache tablets administered to 25 subjects experiencing fevers of 38°C or more. Perform the analysis of variance and test the hypothesis at the 0.05 level of significance that the mean number of hours of relief provided by the tablets is the same for all five brands. Discuss the results.

Tablet				
A	***B***	***C***	***D***	***E***
5.2	9.1	3.2	2.4	7.1
4.7	7.1	5.8	3.4	6.6
8.1	8.2	2.2	4.1	9.3
6.2	6.0	3.1	1.0	4.2
3.0	9.1	7.2	4.0	7.6

13.3 In an article "Shelf-Space Strategy in Retailing," published in *Proceedings: Southern Marketing Association*, the effect of shelf height on the supermarket sales of canned dog food is investigated. An experiment was conducted at a small supermarket for a period of 8 days on the sales of a single brand of dog food, referred to as Arf dog food, involving three levels of shelf height: knee level, waist level, and eye level. During each day, the shelf height of the canned dog food was randomly changed on three different occasions. The remaining sections of the gondola that housed the given brand were filled with a mixture of dog food brands that were both familiar and unfamiliar to customers in this particular geographic area. Sales, in hundreds of dollars, of Arf dog food per day for the three shelf heights are given. Based on the data, is there a significant difference in the average daily sales of this dog food based on shelf height? Use a 0.01 level of significance.

Shelf Height		
Knee Level	**Waist Level**	**Eye Level**
77	88	85
82	94	85
86	93	87
78	90	81
81	91	80
86	94	79
77	90	87
81	87	93

13.4 Immobilization of free-ranging white-tailed deer by drugs allows researchers the opportunity to closely examine the deer and gather valuable physiological information. In the study *Influence of Physical Restraint and Restraint Facilitating Drugs on Blood Measurements of White-Tailed Deer and Other Selected Mammals*, conducted at Virginia Tech, wildlife biologists tested the "knockdown" time (time from injection to immobilization) of three different immobilizing drugs. Immobilization, in this case, is defined as the point where the animal no longer has enough muscle control to remain standing. Thirty male white-tailed deer were randomly assigned to each of three treatments. Group *A* received 5 milligrams of liquid succinylcholine chloride (SCC); group *B* received 8 milligrams of powdered SCC; and group *C* received 200 milligrams of phencyclidine hydrochloride. Knockdown times, in minutes, were recorded. Perform an analysis of variance at the 0.01 level of significance and determine whether or not the average knockdown time for the three drugs is the same.

Group		
A	***B***	***C***
11	10	4
5	7	4
14	16	6
7	7	3
10	7	5
7	5	6
23	10	8
4	10	3
11	6	7
11	12	3

13.5 The mitochondrial enzyme NADPH:NAD transhydrogenase of the common rat tapeworm (*Hymenolepiasis diminuta*) catalyzes hydrogen in the transfer from NADPH to NAD, producing NADH. This enzyme is known to serve a vital role in the tapeworm's anaerobic metabolism, and it has recently been hypothesized that it may serve as a proton exchange pump, transferring protons across the mitochondrial membrane. A study on *Effect of Various Substrate Concentrations on the Conformational Variation of the NADPH:NAD Transhydrogenase of Hymenolepiasis diminuta*, conducted at Bowling Green State University, was designed to assess the ability of this enzyme to undergo conformation or shape changes. Changes in the specific activity of the enzyme caused by variations in the concentration of NADP could be interpreted as supporting the theory of conformational change. The enzyme in question is located in the inner membrane of the tapeworm's mitochondria. Tapeworms were homogenized, and through a series of centrifugations, the enzyme was isolated. Various concentrations of NADP were then added to the isolated enzyme solution, and the mixture was then incubated in a water bath at 56°C for 3 minutes. The enzyme was then analyzed on a dual-beam spectrophotometer, and the results shown were calculated, with the specific activity of the enzyme given in nanomoles per minute per milligram of protein. Test the hypothesis at the 0.01 level that the average specific activity is the same for the four concentrations.

NADP Concentration (nm)				
0	**80**	**160**	**360**	
11.01	11.38	11.02	6.04	10.31
12.09	10.67	10.67	8.65	8.30
10.55	12.33	11.50	7.76	9.48
11.26	10.08	10.31	10.13	8.89
			9.36	

13.6 A study measured the sorption (either absorption or adsorption) rates of three different types of organic chemical solvents. These solvents are used to clean industrial fabricated-metal parts and are potential hazardous waste. Independent samples from each type of solvent were tested, and their sorption rates were recorded as a mole percentage. (See McClave, Dietrich, and Sincich, 1997.)

Aromatics		**Chloroalkanes**		**Esters**		
1.06	0.95	1.58	1.12	0.29	0.43	0.06
0.79	0.65	1.45	0.91	0.06	0.51	0.09
0.82	1.15	0.57	0.83	0.44	0.10	0.17
0.89	1.12	1.16	0.43	0.55	0.53	0.17
1.05				0.61	0.34	0.60

Is there a significant difference in the mean sorption rates for the three solvents? Use a *P*-value for your conclusions. Which solvent would you use?

13.7 It has been shown that the fertilizer magnesium ammonium phosphate, $MgNH_4PO_4$, is an effective supplier of the nutrients necessary for plant growth. The compounds supplied by this fertilizer are highly soluble in water, allowing the fertilizer to be applied directly on the soil surface or mixed with the growth substrate during the potting process. A study on the *Effect of Magnesium Ammonium Phosphate on Height of Chrysanthemums* was conducted at George Mason University to determine a possible optimum level of fertilization, based on the enhanced vertical growth response of the chrysanthemums. Forty chrysanthemum

seedlings were divided into four groups, each containing 10 plants. Each was planted in a similar pot containing a uniform growth medium. To each group of plants an increasing concentration of $MgNH_4PO_4$, measured in grams per bushel, was added. The four groups of plants were grown under uniform conditions in a greenhouse for a period of four weeks. The treatments and the respective changes in heights, measured in centimeters, are shown next.

Treatment							
50 g/bu		**100 g/bu**		**200 g/bu**		**400 g/bu**	
13.2	12.4	16.0	12.6	7.8	14.4	21.0	14.8
12.8	17.2	14.8	13.0	20.0	15.8	19.1	15.8
13.0	14.0	14.0	23.6	17.0	27.0	18.0	26.0
14.2	21.6	14.0	17.0	19.6	18.0	21.1	22.0
15.0	20.0	22.2	24.4	20.2	23.2	25.0	18.2

Can we conclude at the 0.05 level of significance that different concentrations of $MgNH_4PO_4$ affect the average attained height of chrysanthemums? How much $MgNH_4PO_4$ appears to be best?

13.8 For the data set in Exercise 13.7, use Bartlett's test to check whether the variances are equal. Use $\alpha = 0.05$.

13.9 Use Bartlett's test at the 0.01 level of significance to test for homogeneity of variances in Exercise 13.5 on page 519.

13.10 Use Cochran's test at the 0.01 level of significance to test for homogeneity of variances in Exercise 13.4 on page 519.

13.11 Use Bartlett's test at the 0.05 level of significance to test for homogeneity of variances in Exercise 13.6 on page 519.

13.5 Single-Degree-of-Freedom Comparisons

The analysis of variance in a one-way classification, or a one-factor experiment, as it is often called, merely indicates whether or not the hypothesis of equal treatment means can be rejected. Usually, an experimenter would prefer his or her analysis to probe deeper. For instance, in Example 13.1, by rejecting the null hypothesis we concluded that the means are not all equal, but we still do not know where the differences exist among the aggregates. The engineer might have the feeling *a priori* that aggregates 1 and 2 should have similar absorption properties and that the same is true for aggregates 3 and 5. However, it is of interest to study the difference between the two groups. It would seem, then, appropriate to test the hypothesis

$$H_0: \ \mu_1 + \mu_2 - \mu_3 - \mu_5 = 0,$$
$$H_1: \ \mu_1 + \mu_2 - \mu_3 - \mu_5 \neq 0.$$

We notice that the hypothesis is a linear function of the population means where the coefficients sum to zero.

Definition 13.1:

Any linear function of the form

$$\omega = \sum_{i=1}^{k} c_i \mu_i,$$

where $\sum_{i=1}^{k} c_i = 0$, is called a **comparison** or **contrast** in the treatment means.

The experimenter can often make multiple comparisons by testing the significance of contrasts in the treatment means, that is, by testing a hypothesis of the following type:

Hypothesis for a Contrast

$$H_0: \sum_{i=1}^{k} c_i\mu_i = 0,$$

$$H_1: \sum_{i=1}^{k} c_i\mu_i \neq 0,$$

where $\sum_{i=1}^{k} c_i = 0$.

The test is conducted by first computing a similar contrast in the sample means,

$$w = \sum_{i=1}^{k} c_i\bar{y}_{i\cdot}.$$

Since $\bar{Y}_{1\cdot}, \bar{Y}_{2\cdot}, \ldots, \bar{Y}_{k\cdot}$ are independent random variables having normal distributions with means $\mu_1, \mu_2, \ldots, \mu_k$ and variances $\sigma_1^2/n_1, \sigma_2^2/n_2, \ldots, \sigma_k^2/n_k$, respectively, Theorem 7.11 assures us that w is a value of the normal random variable W with

$$\text{mean } \mu_W = \sum_{i=1}^{k} c_i\mu_i \text{ and variance } \sigma_W^2 = \sigma^2 \sum_{i=1}^{k} \frac{c_i^2}{n_i}.$$

Therefore, when H_0 is true, $\mu_W = 0$ and, by Example 7.5, the statistic

$$\frac{W^2}{\sigma_W^2} = \frac{\left(\sum_{i=1}^{k} c_i\bar{Y}_{i\cdot}\right)^2}{\sigma^2 \sum_{i=1}^{k} (c_i^2/n_i)}$$

is distributed as a chi-squared random variable with 1 degree of freedom.

Test Statistic for Testing a Contrast

Our hypothesis is tested at the α-level of significance by computing

$$f = \frac{\left(\sum_{i=1}^{k} c_i\bar{y}_{i\cdot}\right)^2}{s^2 \sum_{i=1}^{k} (c_i^2/n_i)} = \frac{\left[\sum_{i=1}^{k} (c_iY_{i\cdot}/n_i)\right]^2}{s^2 \sum_{i=1}^{k} (c_i^2/n_i)} = \frac{SSw}{s^2}.$$

Here f is a value of the random variable F having the F-distribution with 1 and $N - k$ degrees of freedom.

When the sample sizes are all equal to n,

$$SSw = \frac{\left(\sum_{i=1}^{k} c_iY_{i\cdot}\right)^2}{n \sum_{i=1}^{k} c_i^2}.$$

The quantity SSw, called the **contrast sum of squares**, indicates the portion of SSA that is explained by the contrast in question.

This sum of squares will be used to test the hypothesis that

$$\sum_{i=1}^{k} c_i \mu_i = 0.$$

It is often of interest to test multiple contrasts, particularly contrasts that are linearly independent or orthogonal. As a result, we need the following definition:

Definition 13.2:

The two contrasts

$$\omega_1 = \sum_{i=1}^{k} b_i \mu_i \qquad \text{and} \qquad \omega_2 = \sum_{i=1}^{k} c_i \mu_i$$

are said to be **orthogonal** if $\sum_{i=1}^{k} b_i c_i / n_i = 0$ or, when the n_i are all equal to n, if

$$\sum_{i=1}^{k} b_i c_i = 0.$$

If ω_1 and ω_2 are orthogonal, then the quantities SSw_1 and SSw_2 are components of SSA, each with a single degree of freedom. The treatment sum of squares with $k-1$ degrees of freedom can be partitioned into at most $k-1$ independent single-degree-of-freedom contrast sums of squares satisfying the identity

$$SSA = SSw_1 + SSw_2 + \cdots + SSw_{k-1},$$

if the contrasts are orthogonal to each other.

Example 13.4: Referring to Example 13.1, find the contrast sum of squares corresponding to the orthogonal contrasts

$$\omega_1 = \mu_1 + \mu_2 - \mu_3 - \mu_5, \qquad \omega_2 = \mu_1 + \mu_2 + \mu_3 - 4\mu_4 + \mu_5,$$

and carry out appropriate tests of significance. In this case, it is of interest *a priori* to compare the two groups $(1, 2)$ and $(3, 5)$. An important and independent contrast is the comparison between the set of aggregates $(1, 2, 3, 5)$ and aggregate 4.

***Solution*:** It is obvious that the two contrasts are orthogonal, since

$$(1)(1) + (1)(1) + (-1)(1) + (0)(-4) + (-1)(1) = 0.$$

The second contrast indicates a comparison between aggregates $(1, 2, 3,$ and $5)$ and aggregate 4. We can write two additional contrasts orthogonal to the first two, namely

$$\omega_3 = \mu_1 - \mu_2 \qquad \text{(aggregate 1 versus aggregate 2)},$$
$$\omega_4 = \mu_3 - \mu_5 \qquad \text{(aggregate 3 versus aggregate 5)}.$$

From the data of Table 13.1, we have

$$SSw_1 = \frac{(3320 + 3416 - 3663 - 3664)^2}{6[(1)^2 + (1)^2 + (-1)^2 + (-1)^2]} = 14,553,$$

$$SSw_2 = \frac{[3320 + 3416 + 3663 + 3664 - 4(2791)]^2}{6[(1)^2 + (1)^2 + (1)^2 + (1)^2 + (-4)^2]} = 70,035.$$

A more extensive analysis-of-variance table is shown in Table 13.5. We note that the two contrast sums of squares account for nearly all the aggregate sum of squares. There is a significant difference between aggregates in their absorption properties, and the contrast ω_1 is marginally significant. However, the *f*-value of 14.12 for ω_2 is highly significant, and the hypothesis

$$H_0: \ \mu_1 + \mu_2 + \mu_3 + \mu_5 = 4\mu_4$$

is rejected.

Table 13.5: Analysis of Variance Using Orthogonal Contrasts

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Computed *f*
Aggregates	85,356	4	21,339	4.30
$(1,2)$ vs. $(3,5)$	14,553	1	14,553	2.93
$(1,2,3,5)$ vs. 4	70,035	1	70,035	14.12
Error	124,021	25	4961	
Total	209,377	29		

Orthogonal contrasts allow the practitioner to partition the treatment variation into independent components. Normally, the experimenter would have certain contrasts that were of interest to him or her. Such was the case in our example, where *a priori* considerations suggested that aggregates $(1,2)$ and $(3,5)$ constituted distinct groups with different absorption properties, a postulation that was not strongly supported by the significance test. However, the second comparison supported the conclusion that aggregate 4 seemed to "stand out" from the rest. In this case, the complete partitioning of SSA was not necessary, since two of the four possible independent comparisons accounted for a majority of the variation in treatments.

Figure 13.4 shows a *SAS* GLM procedure that displays a complete set of orthogonal contrasts. Note that the sums of squares for the four contrasts add to the aggregate sum of squares. Also, note that the latter two contrasts (1 versus 2, 3 versus 5) reveal insignificant comparisons. ◢

13.6 Multiple Comparisons

The analysis of variance is a powerful procedure for testing the homogeneity of a set of means. However, if we reject the null hypothesis and accept the stated alternative—that the means are not all equal—we still do not know which of the population means are equal and which are different.

Chapter 14

Factorial Experiments (Two or More Factors)

14.1 Introduction

Consider a situation where it is of interest to study the effects of **two factors**, A and B, on some response. For example, in a chemical experiment, we would like to vary simultaneously the reaction pressure and reaction time and study the effect of each on the yield. In a biological experiment, it is of interest to study the effects of drying time and temperature on the amount of solids (percent by weight) left in samples of yeast. As in Chapter 13, the term **factor** is used in a general sense to denote any feature of the experiment such as temperature, time, or pressure that may be varied from trial to trial. We define the **levels** of a factor to be the actual values used in the experiment.

For each of these cases, it is important to determine not only if each of the two factors has an influence on the response, but also if there is a significant interaction between the two factors. As far as terminology is concerned, the experiment described here is a two-factor experiment and the experimental design may be either a completely randomized design, in which the various treatment combinations are assigned randomly to all the experimental units, or a randomized complete block design, in which factor combinations are assigned randomly within blocks. In the case of the yeast example, the various treatment combinations of temperature and drying time would be assigned randomly to the samples of yeast if we were using a completely randomized design.

Many of the concepts studied in Chapter 13 are extended in this chapter to two and three factors. The main thrust of this material is the use of the completely randomized design with a *factorial experiment*. A factorial experiment in two factors involves experimental trials (or a single trial) with all factor combinations. For example, in the temperature-drying-time example with, say, 3 levels of each and $n = 2$ runs at each of the 9 combinations, we have a *two-factor factorial experiment in a completely randomized design*. Neither factor is a blocking factor; we are interested in how each influences percent solids in the samples and whether or not they interact. The biologist would have available 18 physical samples of

material which are experimental units. These would then be assigned randomly to the 18 combinations (9 treatment combinations, each duplicated).

Before we launch into analytical details, sums of squares, and so on, it may be of interest for the reader to observe the obvious connection between what we have described and the situation with the one-factor problem. Consider the yeast experiment. Explanation of degrees of freedom aids the reader or the analyst in visualizing the extension. We should initially view the 9 treatment combinations as if they represented one factor with 9 levels (8 degrees of freedom). Thus, an initial look at degrees of freedom gives

Treatment combinations	8
Error	9
Total	17

Main Effects and Interaction

The experiment could be analyzed as described in the above table. However, the F-test for combinations would probably not give the analyst the information he or she desires, namely, that which considers the role of temperature and drying time. Three drying times have 2 associated degrees of freedom; three temperatures have 2 degrees of freedom. The main factors, temperature and drying time, are called **main effects**. The main effects represent 4 of the 8 degrees of freedom for *factor combinations*. The additional 4 degrees of freedom are associated with *interaction* between the two factors. As a result, the analysis involves

Combinations		8
Temperature	2	
Drying time	2	
Interaction	4	
Error		9
Total		17

Recall from Chapter 13 that factors in an analysis of variance may be viewed as fixed or random, depending on the type of inference desired and how the levels were chosen. Here we must consider fixed effects, random effects, and even cases where effects are mixed. Most attention will be directed toward expected mean squares when we advance to these topics. In the following section, we focus on the concept of interaction.

14.2 Interaction in the Two-Factor Experiment

In the randomized block model discussed previously, it was assumed that one observation on each treatment is taken in each block. If the model assumption is correct, that is, if blocks and treatments are the only real effects and interaction does not exist, the expected value of the mean square error is the experimental error variance σ^2. Suppose, however, that there is interaction occurring between treatments and blocks as indicated by the model

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

of Section 13.8. The expected value of the mean square error is then given as

$$E\left[\frac{SSE}{(b-1)(k-1)}\right] = \sigma^2 + \frac{1}{(b-1)(k-1)}\sum_{i=1}^{k}\sum_{j=1}^{b}(\alpha\beta)_{ij}^2.$$

The treatment and block effects do not appear in the expected mean square error, but the interaction effects do. Thus, if there is interaction in the model, the mean square error reflects variation due to experimental error plus an interaction contribution, and for this experimental plan, there is no way of separating them.

Interaction and the Interpretation of Main Effects

From an experimenter's point of view it should seem necessary to arrive at a significance test on the existence of interaction by separating true error variation from that due to interaction. The main effects, A and B, take on a different meaning in the presence of interaction. In the previous biological example, the effect that drying time has on the amount of solids left in the yeast might very well depend on the temperature to which the samples are exposed. In general, there could be experimental situations in which factor A has a positive effect on the response at one level of factor B, while at a different level of factor B the effect of A is negative. We use the term **positive effect** here to indicate that the yield or response increases as the levels of a given factor increase according to some defined order. In the same sense, a **negative effect** corresponds to a decrease in response for increasing levels of the factor.

Consider, for example, the following data on temperature (factor A at levels t_1, t_2, and t_3 in increasing order) and drying time d_1, d_2, and d_3 (also in increasing order). The response is percent solids. These data are completely hypothetical and given to illustrate a point.

	B			
A	d_1	d_2	d_3	**Total**
t_1	4.4	8.8	5.2	18.4
t_2	7.5	8.5	2.4	18.4
t_3	9.7	7.9	0.8	18.4
Total	21.6	25.2	8.4	55.2

Clearly the effect of temperature on percent solids is positive at the low drying time d_1 but negative for high drying time d_3. This **clear interaction** between temperature and drying time is obviously of interest to the biologist, but, based on the totals of the responses for temperatures t_1, t_2, and t_3, the temperature sum of squares, SSA, will yield a value of zero. We say then that the presence of interaction is **masking** the effect of temperature. Thus, if we consider the average effect of temperature, averaged over drying time, **there is no effect**. This then defines the main effect. But, of course, this is likely not what is pertinent to the biologist.

Before drawing any final conclusions resulting from tests of significance on the main effects and interaction effects, the **experimenter should first observe whether or not the test for interaction is significant**. If interaction is

not significant, then the results of the tests on the main effects are meaningful. However, if interaction should be significant, then only those tests on the main effects that turn out to be significant are meaningful. Nonsignificant main effects in the presence of interaction might well be a result of masking and dictate the need to observe the influence of each factor at fixed levels of the other.

A Graphical Look at Interaction

The presence of interaction as well as its scientific impact can be interpreted nicely through the use of **interaction plots**. The plots clearly give a pictorial view of the tendency in the data to show the effect of changing one factor as one moves from one level to another of a second factor. Figure 14.1 illustrates the strong temperature by drying time interaction. The interaction is revealed in nonparallel lines.

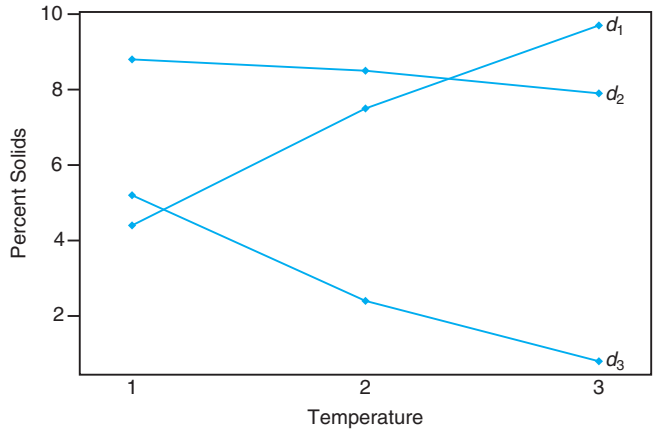


Figure 14.1: Interaction plot for temperature–drying time data.

The relatively strong *temperature effect* on percent solids at the lower drying time is reflected in the steep slope at d_1. At the middle drying time d_2 the temperature has very little effect, while at the high drying time d_3 the negative slope illustrates a negative effect of temperature. Interaction plots such as this set give the scientist a quick and meaningful interpretation of the interaction that is present. It should be apparent that **parallelism** in the plots signals an **absence of interaction**.

Need for Multiple Observations

Interaction and experimental error are separated in the two-factor experiment only if multiple observations are taken at the various treatment combinations. For maximum efficiency, there should be the same number n of observations at each combination. These should be true replications, not just repeated measurements. For

example, in the yeast illustration, if we take $n = 2$ observations at each combination of temperature and drying time, there should be two separate samples and not merely repeated measurements on the same sample. This allows variability due to experimental units to appear in "error," so the variation is not merely measurement error.

14.3 Two-Factor Analysis of Variance

To present general formulas for the analysis of variance of a two-factor experiment using repeated observations in a completely randomized design, we shall consider the case of n replications of the treatment combinations determined by a levels of factor A and b levels of factor B. The observations may be classified by means of a rectangular array where the rows represent the levels of factor A and the columns represent the levels of factor B. Each treatment combination defines a cell in our array. Thus, we have ab cells, each cell containing n observations. Denoting the kth observation taken at the ith level of factor A and the jth level of factor B by y_{ijk}, Table 14.1 shows the abn observations.

Table 14.1: Two-Factor Experiment with n Replications

	B					
A	**1**	**2**	$\cdots$	***b***	**Total**	**Mean**
1	y_{111}	y_{121}	$\cdots$	y_{1b1}	$Y_{1..}$	$\bar{y}_{1..}$
	y_{112}	y_{122}	$\cdots$	y_{1b2}		
	$\vdots$	$\vdots$		$\vdots$		
	y_{11n}	y_{12n}	$\cdots$	y_{1bn}		
2	y_{211}	y_{221}	$\cdots$	y_{2b1}	$Y_{2..}$	$\bar{y}_{2..}$
	y_{212}	y_{222}	$\cdots$	y_{2b2}		
	$\vdots$	$\vdots$		$\vdots$		
	y_{21n}	y_{22n}	$\cdots$	y_{2bn}		
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
a	y_{a11}	y_{a21}	$\cdots$	y_{ab1}	$Y_{a..}$	$\bar{y}_{a..}$
	y_{a12}	y_{a22}	$\cdots$	y_{ab2}		
	$\vdots$	$\vdots$		$\vdots$		
	y_{a1n}	y_{a2n}	$\cdots$	y_{abn}		
Total	$Y_{.1.}$	$Y_{.2.}$	$\cdots$	$Y_{.b.}$	$Y_{...}$	
Mean	$\bar{y}_{.1.}$	$\bar{y}_{.2.}$	$\cdots$	$\bar{y}_{.b.}$		$\bar{y}_{...}$

The observations in the (ij)th cell constitute a random sample of size n from a population that is assumed to be normally distributed with mean μ_{ij} and variance σ^2. All ab populations are assumed to have the same variance σ^2. Let us define

the following useful symbols, some of which are used in Table 14.1:

$Y_{ij.}$ = sum of the observations in the (ij)th cell,
$Y_{i..}$ = sum of the observations for the ith level of factor A,
$Y_{.j.}$ = sum of the observations for the jth level of factor B,
$Y_{...}$ = sum of all abn observations,
$\bar{y}_{ij.}$ = mean of the observations in the (ij)th cell,
$\bar{y}_{i..}$ = mean of the observations for the ith level of factor A,
$\bar{y}_{.j.}$ = mean of the observations for the jth level of factor B,
$\bar{y}_{...}$ = mean of all abn observations.

Unlike in the one-factor situation covered at length in Chapter 13, here we are assuming that the **populations**, where n independent identically distributed observations are taken, are **combinations** of factors. Also we will assume throughout that an equal number (n) of observations are taken at each factor combination. In cases in which the sample sizes per combination are unequal, the computations are more complicated but the concepts are transferable.

Model and Hypotheses for the Two-Factor Problem

Each observation in Table 14.1 may be written in the form

$$y_{ijk} = \mu_{ij} + \epsilon_{ijk},$$

where ϵ_{ijk} measures the deviations of the observed y_{ijk} values in the (ij)th cell from the population mean μ_{ij}. If we let $(\alpha\beta)_{ij}$ denote the interaction effect of the ith level of factor A and the jth level of factor B, α_i the effect of the ith level of factor A, β_j the effect of the jth level of factor B, and μ the overall mean, we can write

$$\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij},$$

and then

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

on which we impose the restrictions

$$\sum_{i=1}^{a} \alpha_i = 0, \qquad \sum_{j=1}^{b} \beta_j = 0, \qquad \sum_{i=1}^{a} (\alpha\beta)_{ij} = 0, \qquad \sum_{j=1}^{b} (\alpha\beta)_{ij} = 0.$$

The three hypotheses to be tested are as follows:

1. H_0': $\alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$,
H_1': At least one of the α_i is not equal to zero.

2. H_0'': $\beta_1 = \beta_2 = \cdots = \beta_b = 0$,
H_1'': At least one of the β_j is not equal to zero.

3. H_0''': $(\alpha\beta)_{11} = (\alpha\beta)_{12} = \cdots = (\alpha\beta)_{ab} = 0$,
H_1''': At least one of the $(\alpha\beta)_{ij}$ is not equal to zero.

We warned the reader about the problem of masking of main effects when interaction is a heavy contributor in the model. It is recommended that the interaction test result be considered first. The interpretation of the main effect test follows, and the nature of the scientific conclusion depends on whether interaction is found. If interaction is ruled out, then hypotheses 1 and 2 above can be tested and the interpretation is quite simple. However, if interaction is found to be present the interpretation can be more complicated, as we have seen from the discussion of the drying time and temperature in the previous section. In what follows, the structure of the tests of hypotheses 1, 2, and 3 will be discussed. Interpretation of results will be incorporated in the discussion of the analysis in Example 14.1.

The tests of the hypotheses above will be based on a comparison of independent estimates of σ^2 provided by splitting the total sum of squares of our data into four components by means of the following identity.

Partitioning of Variability in the Two-Factor Case

Theorem 14.1: **Sum-of-Squares Identity**

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk}-\bar{y}_{...})^2 = bn\sum_{i=1}^{a}(\bar{y}_{i..}-\bar{y}_{...})^2 + an\sum_{j=1}^{b}(\bar{y}_{.j.}-\bar{y}_{...})^2 + n\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y}_{...})^2 + \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk}-\bar{y}_{ij.})^2$$

Symbolically, we write the sum-of-squares identity as

$$SST = SSA + SSB + SS(AB) + SSE,$$

where SSA and SSB are called the sums of squares for the main effects A and B, respectively, $SS(AB)$ is called the interaction sum of squares for A and B, and SSE is the error sum of squares. The degrees of freedom are partitioned according to the identity

$$abn - 1 = (a-1) + (b-1) + (a-1)(b-1) + ab(n-1).$$

Formation of Mean Squares

If we divide each of the sums of squares on the right side of the sum-of-squares identity by its corresponding number of degrees of freedom, we obtain the four statistics

$$S_1^2 = \frac{SSA}{a-1}, \qquad S_2^2 = \frac{SSB}{b-1}, \qquad S_3^2 = \frac{SS(AB)}{(a-1)(b-1)}, \qquad S^2 = \frac{SSE}{ab(n-1)}.$$

All of these variance estimates are independent estimates of σ^2 under the condition that there are no effects α_i, β_j, and, of course, $(\alpha\beta)_{ij}$. If we interpret the sums of

squares as functions of the independent random variables $y_{111}, y_{112}, \ldots, y_{abn}$, it is not difficult to verify that

$$E(S_1^2) = E\left[\frac{SSA}{a-1}\right] = \sigma^2 + \frac{nb}{a-1}\sum_{i=1}^{a}\alpha_i^2,$$
$$E(S_2^2) = E\left[\frac{SSB}{b-1}\right] = \sigma^2 + \frac{na}{b-1}\sum_{j=1}^{b}\beta_j^2,$$
$$E(S_3^2) = E\left[\frac{SS(AB)}{(a-1)(b-1)}\right] = \sigma^2 + \frac{n}{(a-1)(b-1)}\sum_{i=1}^{a}\sum_{j=1}^{b}(\alpha\beta)_{ij}^2,$$
$$E(S^2) = E\left[\frac{SSE}{ab(n-1)}\right] = \sigma^2,$$

from which we immediately observe that all four estimates of σ^2 are unbiased when H_0', H_0'', and H_0''' are true.

To test the hypothesis H_0', that the effects of factors A are all equal to zero, we compute the following ratio:

F-Test for Factor *A*

$$f_1 = \frac{s_1^2}{s^2},$$

which is a value of the random variable F_1 having the F-distribution with $a-1$ and $ab(n-1)$ degrees of freedom when H_0' is true. The null hypothesis is rejected at the α-level of significance when $f_1 > f_\alpha[a-1, ab(n-1)]$.

Similarly, to test the hypothesis H_0'' that the effects of factor B are all equal to zero, we compute the following ratio:

F-Test for Factor *B*

$$f_2 = \frac{s_2^2}{s^2},$$

which is a value of the random variable F_2 having the F-distribution with $b-1$ and $ab(n-1)$ degrees of freedom when H_0'' is true. This hypothesis is rejected at the α-level of significance when $f_2 > f_\alpha[b-1, ab(n-1)]$.

Finally, to test the hypothesis H_0''', that the interaction effects are all equal to zero, we compute the following ratio:

F-Test for Interaction

$$f_3 = \frac{s_3^2}{s^2},$$

which is a value of the random variable F_3 having the F-distribution with $(a-1)(b-1)$ and $ab(n-1)$ degrees of freedom when H_0''' is true. We conclude that, at the α-level of significance, interaction is present when $f_3 > f_\alpha[(a-1)(b-1), ab(n-1)]$.

As indicated in Section 14.2, it is advisable to interpret the test for interaction before attempting to draw inferences on the main effects. If interaction is not significant, there is certainly evidence that the tests on main effects are interpretable. Rejection of hypothesis 1 on page 566 implies that the response means at the levels

of factor A are significantly different, while rejection of hypothesis 2 implies a similar condition for the means at levels of factor B. However, a significant interaction could very well imply that the data should be analyzed in a somewhat different manner—**perhaps observing the effect of factor A at fixed levels of factor B**, and so forth.

The computations in an analysis-of-variance problem, for a two-factor experiment with n replications, are usually summarized as in Table 14.2.

Table 14.2: Analysis of Variance for the Two-Factor Experiment with n Replications

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	**Computed f**
Main effect:				
A	SSA	$a-1$	$s_1^2 = \frac{SSA}{a-1}$	$f_1 = \frac{s_1^2}{s^2}$
B	SSB	$b-1$	$s_2^2 = \frac{SSB}{b-1}$	$f_2 = \frac{s_2^2}{s^2}$
Two-factor interactions:				
AB	$SS(AB)$	$(a-1)(b-1)$	$s_3^2 = \frac{SS(AB)}{(a-1)(b-1)}$	$f_3 = \frac{s_3^2}{s^2}$
Error	SSE	$ab(n-1)$	$s^2 = \frac{SSE}{ab(n-1)}$	
Total	SST	$abn-1$		

Example 14.1: In an experiment conducted to determine which of 3 different missile systems is preferable, the propellant burning rate for 24 static firings was measured. Four different propellant types were used. The experiment yielded duplicate observations of burning rates at each combination of the treatments.

The data, after coding, are given in Table 14.3. Test the following hypotheses: (a) H_0': there is no difference in the mean propellant burning rates when different missile systems are used, (b) H_0'': there is no difference in the mean propellant burning rates of the 4 propellant types, (c) H_0''': there is no interaction between the different missile systems and the different propellant types.

Table 14.3: Propellant Burning Rates

Missile System	**Propellant Type**			
	b_1	b_2	b_3	b_4
a_1	34.0	30.1	29.8	29.0
	32.7	32.8	26.7	28.9
a_2	32.0	30.2	28.7	27.6
	33.2	29.8	28.1	27.8
a_3	28.4	27.3	29.7	28.8
	29.3	28.9	27.3	29.1

Solution: 1. (a) H_0': $\alpha_1 = \alpha_2 = \alpha_3 = 0$.
(b) H_0'': $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$.

(c) H_0''': $(\alpha\beta)_{11} = (\alpha\beta)_{12} = \cdots = (\alpha\beta)_{34} = 0$.

2. (a) H_1': At least one of the α_i is not equal to zero.
   (b) H_1'': At least one of the β_j is not equal to zero.
   (c) H_1''': At least one of the $(\alpha\beta)_{ij}$ is not equal to zero.

The sum-of-squares formula is used as described in Theorem 14.1. The analysis of variance is shown in Table 14.4.

Table 14.4: Analysis of Variance for the Data of Table 14.3

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	**Computed f**
Missile system	14.52	2	7.26	5.84
Propellant type	40.08	3	13.36	10.75
Interaction	22.16	6	3.69	2.97
Error	14.91	12	1.24	
Total	91.68	23		

The reader is directed to a *SAS* GLM Procedure (General Linear Models) for analysis of the burning rate data in Figure 14.2. Note how the "model" (11 degrees of freedom) is initially tested and the system, type, and system by type interaction are tested separately. The F-test on the model ($P = 0.0030$) is testing the accumulation of the two main effects and the interaction.

(a) Reject H_0' and conclude that different missile systems result in different mean propellant burning rates. The P-value is approximately 0.0169.

(b) Reject H_0'' and conclude that the mean propellant burning rates are not the same for the four propellant types. The P-value is approximately 0.0010.

(c) Interaction is barely insignificant at the 0.05 level, but the P-value of approximately 0.0513 would indicate that interaction must be taken seriously.

At this point we should draw some type of interpretation of the interaction. It should be emphasized that statistical significance of a main effect merely implies that *marginal means are significantly different*. However, consider the two-way table of averages in Table 14.5.

Table 14.5: Interpretation of Interaction

	b_1	b_2	b_3	b_4	**Average**
a_1	33.35	31.45	28.25	28.95	30.50
a_2	32.60	30.00	28.40	27.70	29.68
a_3	28.85	28.10	28.50	28.95	28.60
Average	31.60	29.85	28.38	28.53	

It is apparent that more important information exists in the body of the table—trends that are inconsistent with the trend depicted by marginal averages. Table 14.5 certainly suggests that the effect of propellant type depends on the system

```
                          The GLM Procedure
Dependent Variable: rate
                                 Sum of
 Source             DF       Squares    Mean Square   F Value   Pr > F
 Model              11   76.76833333     6.97893939      5.62   0.0030
 Error              12   14.91000000     1.24250000
 Corrected Total    23   91.67833333

 R-Square     Coeff Var      Root MSE     rate Mean
 0.837366      3.766854      1.114675      29.59167

 Source             DF   Type III SS   Mean Square  F Value    Pr > F
 system              2   14.52333333    7.26166667     5.84    0.0169
 type                3   40.08166667   13.36055556    10.75    0.0010
 system*type         6   22.16333333    3.69388889     2.97    0.0512
```

Figure 14.2: *SAS* printout of the analysis of the propellant rate data of Table 14.3.

being used. For example, for system 3 the propellant-type effect does not appear to be important, although it does have a large effect if either system 1 or system 2 is used. This explains the "significant" interaction between these two factors. More will be revealed subsequently concerning this interaction. ◢

Example 14.2: Referring to Example 14.1, choose two orthogonal contrasts to partition the sum of squares for the missile systems into single-degree-of-freedom components to be used in comparing systems 1 and 2 versus 3, and system 1 versus system 2.

***Solution*:** The contrast for comparing systems 1 and 2 with 3 is

$$w_1 = \mu_{1.} + \mu_{2.} - 2\mu_{3.}.$$

A second contrast, orthogonal to w_1, for comparing system 1 with system 2, is given by $w_2 = \mu_{1.} - \mu_{2.}$. The single-degree-of-freedom sums of squares are

$$SSw_1 = \frac{[244.0 + 237.4 - (2)(228.8)]^2}{(8)[(1)^2 + (1)^2 + (-2)^2]} = 11.80$$

and

$$SSw_2 = \frac{(244.0 - 237.4)^2}{(8)[(1)^2 + (-1)^2]} = 2.72.$$

Notice that $SSw_1 + SSw_2 = SSA$, as expected. The computed f-values corresponding to w_1 and w_2 are, respectively,

$$f_1 = \frac{11.80}{1.24} = 9.5 \qquad \text{and} \qquad f_2 = \frac{2.72}{1.24} = 2.2.$$

Compared to the critical value $f_{0.05}(1, 12) = 4.75$, we find f_1 to be significant. In fact, the P-value is less than 0.01. Thus, the first contrast indicates that the

hypothesis

$$H_0\colon\ \frac{1}{2}(\mu_{1.} + \mu_{2.}) = \mu_{3.}$$

is rejected. Since $f_2 < 4.75$, the mean burning rates of the first and second systems are not significantly different.

Impact of Significant Interaction in Example 14.1

If the hypothesis of no interaction in Example 14.1 is true, we could make the *general* comparisons of Example 14.2 regarding our missile systems rather than separate comparisons for each propellant. Similarly, we might make general comparisons among the propellants rather than separate comparisons for each missile system. For example, we could compare propellants 1 and 2 with 3 and 4 and also propellant 1 versus propellant 2. The resulting *f*-ratios, each with 1 and 12 degrees of freedom, turn out to be 24.81 and 7.39, respectively, and both are quite significant at the 0.05 level.

From propellant averages there appears to be evidence that propellant 1 gives the highest mean burning rate. A prudent experimenter might be somewhat cautious in drawing overall conclusions in a problem such as this one, where the *f*-ratio for interaction is barely below the 0.05 critical value. For example, the overall evidence, 31.60 versus 29.85 on the average for the two propellants, certainly indicates that propellant 1 is superior, in terms of a higher burning rate, to propellant 2. However, if we restrict ourselves to system 3, where we have an average of 28.85 for propellant 1 as opposed to 28.10 for propellant 2, there appears to be little or no difference between these two propellants. In fact, there appears to be a stabilization of burning rates for the different propellants if we operate with system 3. There is certainly overall evidence which indicates that system 1 gives a higher burning rate than system 3, but if we restrict ourselves to propellant 4, this conclusion does not appear to hold.

The analyst can conduct a simple *t*-test using average burning rates for system 3 in order to display conclusive evidence that interaction is *producing considerable difficulty in allowing broad conclusions on main effects*. Consider a comparison of propellant 1 against propellant 2 only using system 3. Borrowing an estimate of σ^2 from the overall analysis, that is, using $s^2 = 1.24$ with 12 degrees of freedom, we have

$$|t| = \frac{0.75}{\sqrt{2s^2/n}} = \frac{0.75}{\sqrt{1.24}} = 0.67,$$

which is not even close to being significant. This illustration suggests that one must be cautious about strict interpretation of main effects in the presence of interaction.

Graphical Analysis for the Two-Factor Problem of Example 14.1

Many of the same types of graphical displays that were suggested in the one-factor problems certainly apply in the two-factor case. Two-dimensional plots of cell means or treatment combination means can provide insight into the presence of

interactions between the two factors. In addition, a plot of residuals against fitted values may well provide an indication of whether or not the homogeneous variance assumption holds. Often, of course, a violation of the homogeneous variance assumption involves an increase in the error variance as *the response numbers get larger*. As a result, this plot may point out the violation.

Figure 14.3 shows the plot of cell means in the case of the missile system propellant illustration in Example 14.1. Notice how graphically (in this case) the lack of parallelism shows through. Note the flatness of the part of the figure showing the propellant effect for system 3. This illustrates interaction among the factors. Figure 14.4 shows the plot of residuals against fitted values for the same data. There is no apparent sign of difficulty with the homogeneous variance assumption.

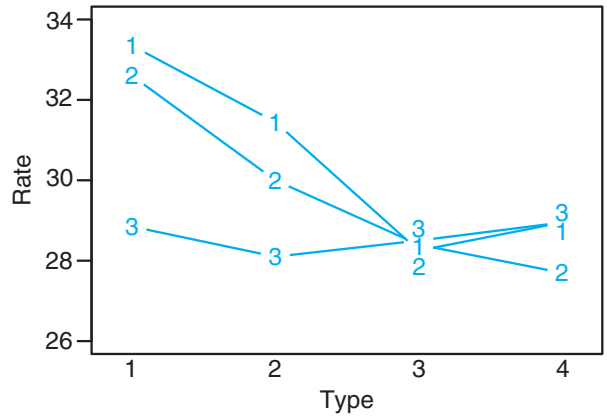


Figure 14.3: Plot of cell means for data of Example 14.1. Numbers represent missile systems.

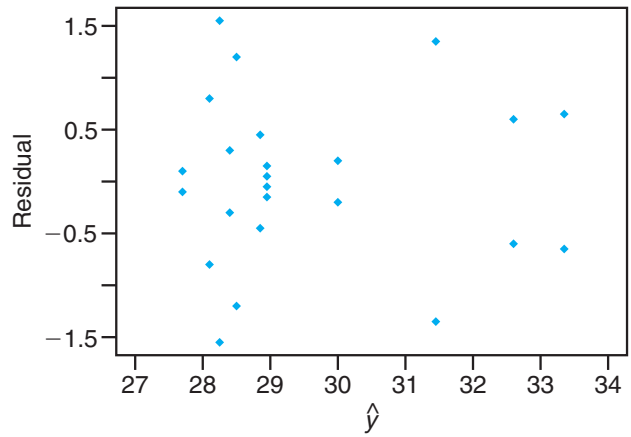


Figure 14.4: Residual plot of data of Example 14.1.

Example 14.3: An electrical engineer is investigating a plasma etching process used in semiconductor manufacturing. It is of interest to study the effects of two factors, the C_2F_6 gas flow rate (A) and the power applied to the cathode (B). The response is the etch rate. Each factor is run at 3 levels, and 2 experimental runs on etch rate are made for each of the 9 combinations. The setup is that of a completely randomized design. The data are given in Table 14.6. The etch rate is in A°/min.

Table 14.6: Data for Example 14.3

	Power Supplied		
C_2F_6 Flow Rate	**1**	**2**	**3**
1	288	488	670
	360	465	720
2	385	482	692
	411	521	724
3	488	595	761
	462	612	801

The levels of the factors are in ascending order, with level 1 being low level and level 3 being the highest.

(a) Show an analysis of variance table and draw conclusions, beginning with the test on interaction.

(b) Do tests on main effects and draw conclusions.

***Solution*:** A *SAS* output is given in Figure 14.5. From the output we learn the following.

```
                          The GLM Procedure
Dependent Variable: etchrate
                                 Sum of
Source                     DF       Squares   Mean Square   F Value   Pr > F
Model                       8   379508.7778    47438.5972     61.00   <.0001
Error                       9     6999.5000      777.7222
Corrected Total            17   386508.2778

R-Square     Coeff Var      Root MSE    etchrate Mean
0.981890      5.057714      27.88767          551.3889

Source                     DF   Type III SS   Mean Square   F Value   Pr > F
c2f6                        2    46343.1111    23171.5556     29.79   0.0001
power                       2   330003.4444   165001.7222    212.16   <.0001
c2f6*power                  4     3162.2222      790.5556      1.02   0.4485
```

Figure 14.5: *SAS* printout for Example 14.3.

(a) The P-value for the test of interaction is 0.4485. We can conclude that there is no significant interaction.

(b) There is a significant difference in mean etch rate for the 3 levels of C_2F_6 flow rate. Duncan's test shows that the mean etch rate for level 3 is significantly

higher than that for level 2 and the rate for level 2 is significantly higher than that for level 1. See Figure 14.6(a).

There is a significant difference in mean etch rate based on the level of power to the cathode. Duncan's test revealed that the etch rate for level 3 is significantly higher than that for level 2 and the rate for level 2 is significantly higher than that for level 1. See Figure 14.6(b).

Duncan Grouping	Mean	N	c2f6
A	619.83	6	3
B	535.83	6	2
C	498.50	6	1

(a)

Duncan Grouping	Mean	N	power
A	728.00	6	3
B	527.17	6	2
C	399.00	6	1

(b)

Figure 14.6: *SAS* output, for Example 14.3. (a) Duncan's test on gas flow rate; (b) Duncan's test on power.

Exercises

14.1 An experiment was conducted to study the effects of temperature and type of oven on the life of a particular component. Four types of ovens and 3 temperature levels were used in the experiment. Twenty-four pieces were assigned randomly, two to each combination of treatments, and the following results recorded.

	Oven			
Temperature (°F)	O_1	O_2	O_3	O_4
500	227	214	225	260
	221	259	236	229
550	187	181	232	246
	208	179	198	273
600	174	198	178	206
	202	194	213	219

Using a 0.05 level of significance, test the hypothesis that

(a) different temperatures have no effect on the life of the component;

(b) different ovens have no effect on the life of the component;

(c) the type of oven and temperature do not interact.

14.2 To ascertain the stability of vitamin C in reconstituted frozen orange juice concentrate stored in a refrigerator for a period of up to one week, the study *Vitamin C Retention in Reconstituted Frozen Orange Juice* was conducted by the Department of Human Nutrition and Foods at Virginia Tech. Three types of frozen orange juice concentrate were tested using 3 different time periods. The time periods refer to the number of days from when the orange juice was blended until it was tested. The results, in milligrams of ascorbic acid per liter, were recorded. Use a 0.05 level of significance to test the hypothesis that

(a) there is no difference in ascorbic acid contents among the different brands of orange juice concentrate;

(b) there is no difference in ascorbic acid contents for the different time periods;

(c) the brands of orange juice concentrate and the number of days from the time the juice was blended until it was tested do not interact.

	Time (days)					
Brand	**0**		**3**		**7**	
Richfood	52.6	54.2	49.4	49.2	42.7	48.8
	49.8	46.5	42.8	53.2	40.4	47.6
Sealed-Sweet	56.0	48.0	48.8	44.0	49.2	44.0
	49.6	48.4	44.0	42.4	42.0	43.2
Minute Maid	52.5	52.0	48.0	47.0	48.5	43.3
	51.8	53.6	48.2	49.6	45.2	47.6

14.3 Three strains of rats were studied under 2 environmental conditions for their performance in a maze test. The error scores for the 48 rats were recorded.

	Strain					
Environment	**Bright**		**Mixed**		**Dull**	
Free	28	12	33	83	101	94
	22	23	36	14	33	56
	25	10	41	76	122	83
	36	86	22	58	35	23
Restricted	72	32	60	89	136	120
	48	93	35	126	38	153
	25	31	83	110	64	128
	91	19	99	118	87	140

Use a 0.01 level of significance to test the hypothesis that

(a) there is no difference in error scores for different environments;

(b) there is no difference in error scores for different strains;

(c) the environments and strains of rats do not interact.

14.4 Corrosion fatigue in metals has been defined as the simultaneous action of cyclic stress and chemical attack on a metal structure. A widely used technique for minimizing corrosion fatigue damage in aluminum involves the application of a protective coating. A study conducted by the Department of Mechanical Engineering at Virginia Tech used 3 different levels of humidity

Low: 20–25% relative humidity

Medium: 55–60% relative humidity

High: 86–91% relative humidity

and 3 types of surface coatings

Uncoated: no coating

Anodized: sulfuric acid anodic oxide coating

Conversion: chromate chemical conversion coating

The corrosion fatigue data, expressed in thousands of cycles to failure, were recorded as follows:

	Relative Humidity					
Coating	**Low**		**Medium**		**High**	
Uncoated	361	469	314	522	1344	1216
	466	937	244	739	1027	1097
	1069	1357	261	134	1011	1011
Anodized	114	1032	322	471	78	466
	1236	92	306	130	387	107
	533	211	68	398	130	327
Conversion	130	1482	252	874	586	524
	841	529	105	755	402	751
	1595	754	847	573	846	529

(a) Perform an analysis of variance with $\alpha = 0.05$ to test for significant main and interaction effects.

(b) Use Duncan's multiple-range test at the 0.05 level of significance to determine which humidity levels result in different corrosion fatigue damage.

14.5 To determine which muscles need to be subjected to a conditioning program in order to improve one's performance on the flat serve used in tennis, a study was conducted by the Department of Health, Physical Education and Recreation at Virginia Tech. Five different muscles

1: anterior deltoid 4: middle deltoid

2: pectorial major 5: triceps

3: posterior deltoid

were tested on each of 3 subjects, and the experiment was carried out 3 times for each treatment combination. The electromyographic data, recorded during the serve, are presented here.

	Muscle				
Subject	**1**	**2**	**3**	**4**	**5**
1	32	5	58	10	19
	59	1.5	61	10	20
	38	2	66	14	23
2	63	10	64	45	43
	60	9	78	61	61
	50	7	78	71	42
3	43	41	26	63	61
	54	43	29	46	85
	47	42	23	55	95

Use a 0.01 level of significance to test the hypothesis that

(a) different subjects have equal electromyographic measurements;

(b) different muscles have no effect on electromyographic measurements;

(c) subjects and types of muscle do not interact.

14.6 An experiment was conducted to determine whether additives increase the adhesiveness of rubber products. Sixteen products were made with the new additive and another 16 without the new additive. The observed adhesiveness was as recorded below.

	Temperature (°C)			
	50	**60**	**70**	**80**
Without Additive	2.3	3.4	3.8	3.9
	2.9	3.7	3.9	3.2
	3.1	3.6	4.1	3.0
	3.2	3.2	3.8	2.7
With Additive	4.3	3.8	3.9	3.5
	3.9	3.8	4.0	3.6
	3.9	3.9	3.7	3.8
	4.2	3.5	3.6	3.9

Perform an analysis of variance to test for significant main and interaction effects.

14.7 The extraction rate of a certain polymer is known to depend on the reaction temperature and the amount of catalyst used. An experiment was conducted at four levels of temperature and five levels of the catalyst, and the extraction rate was recorded in the following table.

	Amount of Catalyst				
	0.5%	**0.6%**	**0.7%**	**0.8%**	**0.9%**
50°C	38	45	57	59	57
	41	47	59	61	58
60°C	44	56	70	73	61
	43	57	69	72	58
70°C	44	56	70	73	61
	47	60	67	61	59
80°C	49	62	70	62	53
	47	65	55	69	58

Perform an analysis of variance. Test for significant main and interaction effects.

14.8 In Myers, Montgomery, and Anderson-Cook (2009), a scenario is discussed involving an auto bumper plating process. The response is the thickness of the material. Factors that may impact the thickness include amount of nickel (A) and pH (B). A two-factor experiment is designed. The plan is a completely randomized design in which the individual bumpers are assigned randomly to the factor combinations. Three levels of pH and two levels of nickel content are involved in the experiment. The thickness data, in cm $\times$ 10^{-3}, are as follows:

Nickel Content	**pH**		
(grams)	**5**	**5.5**	**6**
18	250	211	221
	195	172	150
	188	165	170
10	115	88	69
	165	112	101
	142	108	72

(a) Display the analysis-of-variance table with tests for both main effects and interaction. Show P-values.

(b) Give engineering conclusions. What have you learned from the analysis of the data?

(c) Show a plot that depicts either a presence or an absence of interaction.

14.9 An engineer is interested in the effects of cutting speed and tool geometry on the life in hours of a machine tool. Two cutting speeds and two different geometries are used. Three experimental tests are accomplished at each of the four combinations. The data are as follows.

Tool	**Cutting Speed**					
Geometry	**Low**			**High**		
1	22	28	20	34	37	29
2	18	15	16	11	10	10

(a) Show an analysis-of-variance table with tests on interaction and main effects.

(b) Comment on the effect that interaction has on the test on cutting speed.

(c) Do secondary tests that will allow the engineer to learn the true impact of cutting speed.

(d) Show a plot that graphically displays the interaction effect.

14.10 Two factors in a manufacturing process for an integrated circuit are studied in a two-factor experiment. The purpose of the experiment is to learn their effect on the resistivity of the wafer. The factors are implant dose (2 levels) and furnace position (3 levels). Experimentation is costly so only one experimental run is made at each combination. The data are as follows.

Dose	**Position**		
1	15.5	14.8	21.3
2	27.2	24.9	26.1

It is to be assumed that no interaction exists between these two factors.

(a) Write the model and explain terms.

(b) Show the analysis-of-variance table.

(c) Explain the 2 "error" degrees of freedom.

(d) Use Tukey's test to do multiple-comparison tests on furnace position. Explain what the results show.

14.11 A study was done to determine the impact of two factors, method of analysis and the laboratory doing the analysis, on the level of sulfur content in coal. Twenty-eight coal specimens were randomly assigned to 14 factor combinations, the structure of the experimental units represented by combinations of seven laboratories and two methods of analysis with two specimens per factor combination. The data, expressed in percent of sulfur, are as follows:

	Method			
Laboratory	**1**		**2**	
1	0.109	0.105	0.105	0.108
2	0.129	0.122	0.127	0.124
3	0.115	0.112	0.109	0.111
4	0.108	0.108	0.117	0.118
5	0.097	0.096	0.110	0.097
6	0.114	0.119	0.116	0.122
7	0.155	0.145	0.164	0.160

(The data are taken from G. Taguchi, "Signal to Noise Ratio and Its Applications to Testing Material," *Reports of Statistical Application Research*, Union of Japanese Scientists and Engineers, Vol. 18, No. 4, 1971.)

(a) Do an analysis of variance and show results in an analysis-of-variance table.

(b) Is interaction significant? If so, discuss what it means to the scientist. Use a P-value in your conclusion.

(c) Are the individual main effects, laboratory, and method of analysis statistically significant? Discuss

what is learned and let your answer be couched in the context of any significant interaction.

(d) Do an interaction plot that illustrates the effect of interaction.

(e) Do a test comparing methods 1 and 2 at laboratory 1 and do the same test at laboratory 7. Comment on what these results illustrate.

14.12 In an experiment conducted in the Civil Engineering Department at Virginia Tech, growth of a certain type of algae in water was observed as a function of time and the dosage of copper added to the water. The data are as follows. Response is in units of algae.

	Time in Days		
Copper	**5**	**12**	**18**
1	0.30	0.37	0.25
	0.34	0.36	0.23
	0.32	0.35	0.24
2	0.24	0.30	0.27
	0.23	0.32	0.25
	0.22	0.31	0.25
3	0.20	0.30	0.27
	0.28	0.31	0.29
	0.24	0.30	0.25

(a) Do an analysis of variance and show the analysis-of-variance table.

(b) Comment concerning whether the data are sufficient to show a time effect on algae concentration.

(c) Do the same for copper content. Does the level of copper impact algae concentration?

(d) Comment on the results of the test for interaction. How is the effect of copper content influenced by time?

14.13 In Myers, *Classical and Modern Regression with Applications* (Duxbury Classic Series, 2nd edition, 1990), an experiment is described in which the Environmental Protection Agency seeks to determine the effect of two water treatment methods on magnesium uptake. Magnesium levels in grams per cubic centimeter (cc) are measured, and two different time levels are incorporated into the experiment. The data are as follows:

	Treatment					
Time (hr)	**1**			**2**		
1	2.19	2.15	2.16	2.03	2.01	2.04
2	2.01	2.03	2.04	1.88	1.86	1.91

(a) Do an interaction plot. What is your impression?

(b) Do an analysis of variance and show tests for the main effects and interaction.

(c) Give scientific findings regarding how time and treatment influence magnesium uptake.

(d) Fit the appropriate regression model with treatment as a categorical variable. Include interaction in the model.

(e) Is interaction significant in the regression model?

14.14 Consider the data set in Exercise 14.12 and answer the following questions.

(a) Both factors, copper and time, are quantitative in nature. As a result, a regression model may be of interest. Describe what might be an appropriate model using x_1 = copper content and x_2 = time. Fit the model to the data, showing regression coefficients and a t-test on each.

(b) Fit the model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \epsilon,$$

and compare it to the one you chose in (a). Which is more appropriate? Use R^2_{adj} as a criterion.

14.15 The purpose of the study *The Incorporation of a Chelating Agent into a Flame Retardant Finish of a Cotton Flannelette and the Evaluation of Selected Fabric Properties*, conducted at Virginia Tech, was to evaluate the use of a chelating agent as part of the flame retardant finish of cotton flannelette by determining its effect upon flammability after the fabric is laundered under specific conditions. There were two treatments at two levels. Two baths were prepared, one with carboxymethyl cellulose (bath I) and one without (bath II). Half of the fabric was laundered 5 times and half was laundered 10 times. There were 12 pieces of fabric in each bath/number of launderings combination. After the washings, the lengths of fabric that burned and the burn times were measured. Burn times (in seconds) were recorded as follows:

Launderings	**Bath I**			**Bath II**		
5	13.7	23.0	15.7	6.2	5.4	5.0
	25.5	15.8	14.8	4.4	5.0	3.3
	14.0	29.4	9.7	16.0	2.5	1.6
	14.0	12.3	12.3	3.9	2.5	7.1
10	27.2	16.8	12.9	18.2	8.8	14.5
	14.9	17.1	13.0	14.7	17.1	13.9
	10.8	13.5	25.5	10.6	5.8	7.3
	14.2	27.4	11.5	17.7	18.3	9.9

(a) Perform an analysis of variance. Is there a significant interaction term?

(b) Are there main effect differences? Discuss.

14.4 Three-Factor Experiments

In this section, we consider an experiment with three factors, A, B, and C, at a, b, and c levels, respectively, in a completely randomized experimental design. Assume again that we have n observations for each of the abc treatment combinations. We shall proceed to outline significance tests for the three main effects and interactions involved. It is hoped that the reader can then use the description given here to generalize the analysis to $k > 3$ factors.

Model for the Three-Factor Experiment

The model for the three-factor experiment is

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijkl},$$

$i = 1, 2, \ldots, a$; $j = 1, 2, \ldots, b$; $k = 1, 2, \ldots, c$; and $l = 1, 2, \ldots, n$, where α_i, β_j, and γ_k are the main effects and $(\alpha\beta)_{ij}$, $(\alpha\gamma)_{ik}$, and $(\beta\gamma)_{jk}$ are the two-factor interaction effects that have the same interpretation as in the two-factor experiment.

The term $(\alpha\beta\gamma)_{ijk}$ is called the **three-factor interaction effect**, a term that represents a nonadditivity of the $(\alpha\beta)_{ij}$ over the different levels of the factor C. As before, the sum of all main effects is zero and the sum over any subscript of the two- and three-factor interaction effects is zero. In many experimental situations, these higher-order interactions are insignificant and their mean squares reflect only random variation, but we shall outline the analysis in its most general form.

Again, in order that valid significance tests can be made, we must assume that the errors are values of independent and normally distributed random variables, each with mean 0 and common variance σ^2.

The general philosophy concerning the analysis is the same as that discussed for the one- and two-factor experiments. The sum of squares is partitioned into eight terms, each representing a source of variation from which we obtain independent estimates of σ^2 when all the main effects and interaction effects are zero. If the effects of any given factor or interaction are not all zero, then the mean square will estimate the error variance plus a component due to the systematic effect in question.

Sum of Squares for a Three-Factor Experiment

$$SSA = bcn \sum_{i=1}^{a} (\bar{y}_{i\ldots} - \bar{y}_{\ldots})^2 \quad SS(AB) = cn \sum_i \sum_j (\bar{y}_{ij\cdot} - \bar{y}_{i\ldots} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\ldots})^2$$

$$SSB = acn \sum_{j=1}^{b} (\bar{y}_{\cdot j\cdot} - \bar{y}_{\ldots})^2 \quad SS(AC) = bn \sum_i \sum_k (\bar{y}_{i\cdot k} - \bar{y}_{i\ldots} - \bar{y}_{\cdot\cdot k} + \bar{y}_{\ldots})^2$$

$$SSC = abn \sum_{k=1}^{c} (\bar{y}_{\cdot\cdot k} - \bar{y}_{\ldots})^2 \quad SS(BC) = an \sum_j \sum_k (\bar{y}_{\cdot jk} - \bar{y}_{\cdot j\cdot} - \bar{y}_{\cdot\cdot k} + \bar{y}_{\ldots})^2$$

$$SS(ABC) = n \sum_i \sum_j \sum_k (\bar{y}_{ijk} - \bar{y}_{ij\cdot} - \bar{y}_{i\cdot k} - \bar{y}_{\cdot jk} + \bar{y}_{i\ldots} + \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdot\cdot k} - \bar{y}_{\ldots})^2$$

$$SST = \sum_i \sum_j \sum_k \sum_l (y_{ijkl} - \bar{y}_{\ldots})^2 \quad SSE = \sum_i \sum_j \sum_k \sum_l (y_{ijkl} - \bar{y}_{ijk\cdot})^2$$

Chapter 16

Nonparametric Statistics

16.1 Nonparametric Tests

Most of the hypothesis-testing procedures discussed in previous chapters are based on the assumption that the random samples are selected from normal populations. Fortunately, most of these tests are still reliable when we experience slight departures from normality, particularly when the sample size is large. Traditionally, these testing procedures have been referred to as **parametric methods**. In this chapter, we consider a number of alternative test procedures, called **nonparametric** or **distribution-free methods**, that often assume no knowledge whatsoever about the distributions of the underlying populations, except perhaps that they are continuous.

Nonparametric, or distribution-free procedures, are used with increasing frequency by data analysts. There are many applications in science and engineering where the data are reported as values not on a continuum but rather on an **ordinal scale** such that it is quite natural to assign ranks to the data. In fact, the reader may notice quite early in this chapter that the distribution-free methods described here involve an *analysis of ranks*. Most analysts find the computations involved in nonparametric methods to be very appealing and intuitive.

For an example where a nonparametric test is applicable, consider the situation in which two judges rank five brands of premium beer by assigning a rank of 1 to the brand believed to have the best overall quality, a rank of 2 to the second best, and so forth. A nonparametric test could then be used to determine whether there is any agreement between the two judges.

We should also point out that there are a number of disadvantages associated with nonparametric tests. Primarily, they do not utilize all the information provided by the sample, and thus a nonparametric test will be less efficient than the corresponding parametric procedure when both methods are applicable. Consequently, to achieve the same power, a nonparametric test will require a larger sample size than will the corresponding parametric test.

As we indicated earlier, slight departures from normality result in minor deviations from the ideal for the standard parametric tests. This is particularly true for the t-test and the F-test. In the case of the t-test and the F-test, the P-value

quoted may be slightly in error if there is a moderate violation of the normality assumption.

In summary, if a parametric and a nonparametric test are both applicable to the same set of data, we should carry out the more efficient parametric technique. However, we should recognize that the assumptions of normality often cannot be justified and that we do not always have quantitative measurements. It is fortunate that statisticians have provided us with a number of useful nonparametric procedures. Armed with nonparametric techniques, the data analyst has more ammunition to accommodate a wider variety of experimental situations. It should be pointed out that even under the standard normal theory assumptions, the efficiencies of the nonparametric techniques are remarkably close to those of the corresponding parametric procedure. On the other hand, serious departures from normality will render the nonparametric method much more efficient than the parametric procedure.

Sign Test

The reader should recall that the procedures discussed in Section 10.4 for testing the null hypothesis that $\mu = \mu_0$ are valid only if the population is approximately normal or if the sample is large. If $n < 30$ and the population is decidedly nonnormal, we must resort to a nonparametric test.

The sign test is used to test hypotheses on a population *median*. In the case of many of the nonparametric procedures, the mean is replaced by the median as the pertinent **location parameter** under test. Recall that the sample median was defined in Section 1.3. The population counterpart, denoted by $\tilde{\mu}$, has an analogous definition. Given a random variable X, $\tilde{\mu}$ is defined such that $P(X > \tilde{\mu}) \leq 0.5$ and $P(X < \tilde{\mu}) \leq 0.5$. In the continuous case,

$$P(X > \tilde{\mu}) = P(X < \tilde{\mu}) = 0.5.$$

Of course, if the distribution is symmetric, the population mean and median are equal. In testing the null hypothesis H_0 that $\tilde{\mu} = \tilde{\mu}_0$ against an appropriate alternative, on the basis of a random sample of size n, we replace each sample value exceeding $\tilde{\mu}_0$ with a *plus* sign and each sample value less than $\tilde{\mu}_0$ with a *minus* sign. If the null hypothesis is true and the population is symmetric, the sum of the plus signs should be approximately equal to the sum of the minus signs. When one sign appears more frequently than it should based on chance alone, we reject the hypothesis that the population median $\tilde{\mu}$ is equal to $\tilde{\mu}_0$.

In theory, the sign test is applicable only in situations where $\tilde{\mu}_0$ cannot equal the value of any of the observations. Although there is a zero probability of obtaining a sample observation exactly equal to $\tilde{\mu}_0$ when the population is continuous, nevertheless, in practice a sample value equal to $\tilde{\mu}_0$ will often occur from a lack of precision in recording the data. When sample values equal to $\tilde{\mu}_0$ are observed, they are excluded from the analysis and the sample size is correspondingly reduced.

The appropriate test statistic for the sign test is the binomial random variable X, representing the number of plus signs in our random sample. If the null hypothesis that $\tilde{\mu} = \tilde{\mu}_0$ is true, the probability that a sample value results in either a plus or a minus sign is equal to $1/2$. Therefore, to test the null hypothesis that

$\tilde{\mu} = \tilde{\mu}_0$, we actually test the null hypothesis that the number of plus signs is a value of a random variable having the binomial distribution with the parameter $p = 1/2$. P-values for both one-sided and two-sided alternatives can then be calculated using this binomial distribution. For example, in testing

$$H_0\text{: } \tilde{\mu} = \tilde{\mu}_0,$$
$$H_1\text{: } \tilde{\mu} < \tilde{\mu}_0,$$

we shall reject H_0 in favor of H_1 only if the proportion of plus signs is sufficiently less than $1/2$, that is, when the value x of our random variable is small. Hence, if the computed P-value

$$P = P(X \leq x \text{ when } p = 1/2)$$

is less than or equal to some preselected significance level α, we reject H_0 in favor of H_1. For example, when $n = 15$ and $x = 3$, we find from Table A.1 that

$$P = P(X \leq 3 \text{ when } p = 1/2) = \sum_{x=0}^{3} b\left(x; 15, \frac{1}{2}\right) = 0.0176,$$

so the null hypothesis $\tilde{\mu} = \tilde{\mu}_0$ can certainly be rejected at the 0.05 level of significance but not at the 0.01 level.

To test the hypothesis

$$H_0\text{: } \tilde{\mu} = \tilde{\mu}_0,$$
$$H_1\text{: } \tilde{\mu} > \tilde{\mu}_0,$$

we reject H_0 in favor of H_1 only if the proportion of plus signs is sufficiently greater than $1/2$, that is, when x is large. Hence, if the computed P-value

$$P = P(X \geq x \text{ when } p = 1/2)$$

is less than α, we reject H_0 in favor of H_1. Finally, to test the hypothesis

$$H_0\text{: } \tilde{\mu} = \tilde{\mu}_0,$$
$$H_1\text{: } \tilde{\mu} \neq \tilde{\mu}_0,$$

we reject H_0 in favor of H_1 when the proportion of plus signs is significantly less than or greater than $1/2$. This, of course, is equivalent to x being sufficiently small or sufficiently large. Therefore, if $x < n/2$ and the computed P-value

$$P = 2P(X \leq x \text{ when } p = 1/2)$$

is less than or equal to α, or if $x > n/2$ and the computed P-value

$$P = 2P(X \geq x \text{ when } p = 1/2)$$

is less than or equal to α, we reject H_0 in favor of H_1.

Whenever $n > 10$, binomial probabilities with $p = 1/2$ can be approximated from the normal curve, since $np = nq > 5$. Suppose, for example, that we wish to test the hypothesis

$$H_0: \ \tilde{\mu} = \tilde{\mu}_0,$$
$$H_1: \ \tilde{\mu} < \tilde{\mu}_0,$$

at the $\alpha = 0.05$ level of significance, for a random sample of size $n = 20$ that yields $x = 6$ plus signs. Using the normal curve approximation with

$$\tilde{\mu} = np = (20)(0.5) = 10$$

and

$$\sigma = \sqrt{npq} = \sqrt{(20)(0.5)(0.5)} = 2.236,$$

we find that

$$z = \frac{6.5 - 10}{2.236} = -1.57.$$

Therefore,

$$P = P(X \leq 6) \approx P(Z < -1.57) = 0.0582,$$

which leads to the nonrejection of the null hypothesis.

Example 16.1: The following data represent the number of hours that a rechargeable hedge trimmer operates before a recharge is required:

$$1.5, 2.2, 0.9, 1.3, 2.0, 1.6, 1.8, 1.5, 2.0, 1.2, 1.7.$$

Use the sign test to test the hypothesis, at the 0.05 level of significance, that this particular trimmer operates a median of 1.8 hours before requiring a recharge.

Solution:

1. H_0: $\tilde{\mu} = 1.8$.
2. H_1: $\tilde{\mu} \neq 1.8$.
3. $\alpha = 0.05$.
4. Test statistic: Binomial variable X with $p = \frac{1}{2}$.
5. Computations: Replacing each value by the symbol "+" if it exceeds 1.8 and by the symbol "−" if it is less than 1.8 and discarding the one measurement that equals 1.8, we obtain the sequence

   $$- \ + \ - \ - \ + \ - \ - \ + \ - \ -$$

   for which $n = 10$, $x = 3$, and $n/2 = 5$. Therefore, from Table A.1 the computed P-value is

   $$P = 2P\left(X \leq 3 \text{ when } p = \frac{1}{2}\right) = 2\sum_{x=0}^{3} b\left(x; 10, \frac{1}{2}\right) = 0.3438 > 0.05.$$

6. Decision: Do not reject the null hypothesis and conclude that the median operating time is not significantly different from 1.8 hours. ∎

We can also use the sign test to test the null hypothesis $\tilde{\mu}_1 - \tilde{\mu}_2 = d_0$ for paired observations. Here we replace each difference, d_i, with a plus or minus sign depending on whether the adjusted difference, $d_i - d_0$, is positive or negative. Throughout this section, we have assumed that the populations are symmetric. However, even if populations are skewed, we can carry out the same test procedure, but the hypotheses refer to the population medians rather than the means.

Example 16.2: A taxi company is trying to decide whether the use of radial tires instead of regular belted tires improves fuel economy. Sixteen cars are equipped with radial tires and driven over a prescribed test course. Without changing drivers, the same cars are then equipped with the regular belted tires and driven once again over the test course. The gasoline consumption, in kilometers per liter, is given in Table 16.1. Can we conclude at the 0.05 level of significance that cars equipped with radial tires obtain better fuel economy than those equipped with regular belted tires?

Table 16.1: Data for Example 16.2

Car	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Radial Tires	4.2	4.7	6.6	7.0	6.7	4.5	5.7	6.0
Belted Tires	4.1	4.9	6.2	6.9	6.8	4.4	5.7	5.8
Car	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
Radial Tires	7.4	4.9	6.1	5.2	5.7	6.9	6.8	4.9
Belted Tires	6.9	4.9	6.0	4.9	5.3	6.5	7.1	4.8

***Solution*:** Let $\tilde{\mu}_1$ and $\tilde{\mu}_2$ represent the median kilometers per liter for cars equipped with radial and belted tires, respectively.

1. H_0: $\tilde{\mu}_1 - \tilde{\mu}_2 = 0$.
2. H_1: $\tilde{\mu}_1 - \tilde{\mu}_2 > 0$.
3. $\alpha = 0.05$.
4. Test statistic: Binomial variable X with $p = 1/2$.
5. Computations: After replacing each positive difference by a "+" symbol and each negative difference by a "−" symbol and then discarding the two zero differences, we obtain the sequence

$$+ \quad - \quad + \quad + \quad - \quad + \quad + \quad + \quad + \quad + \quad + \quad + \quad - \quad +$$

for which $n = 14$ and $x = 11$. Using the normal curve approximation, we find

$$z = \frac{10.5 - 7}{\sqrt{(14)(0.5)(0.5)}} = 1.87,$$

and then

$$P = P(X \geq 11) \approx P(Z > 1.87) = 0.0307.$$

6. Decision: Reject H_0 and conclude that, on the average, radial tires do improve fuel economy. ∎

Not only is the sign test one of the simplest nonparametric procedures to apply; it has the additional advantage of being applicable to dichotomous data that cannot be recorded on a numerical scale but can be represented by positive and negative responses. For example, the sign test is applicable in experiments where a qualitative response such as "hit" or "miss" is recorded, and in sensory-type experiments where a plus or minus sign is recorded depending on whether the taste tester correctly or incorrectly identifies the desired ingredient.

We shall attempt to make comparisons between many of the nonparametric procedures and the corresponding parametric tests. In the case of the sign test the competition is, of course, the *t*-test. If we are sampling from a normal distribution, the use of the *t*-test will result in a larger power for the test. If the distribution is merely symmetric, though not normal, the *t*-test is preferred in terms of power unless the distribution has extremely "heavy tails" compared to the normal distribution.

16.2 Signed-Rank Test

The reader should note that the sign test utilizes only the plus and minus signs of the differences between the observations and $\tilde{\mu}_0$ in the one-sample case, or the plus and minus signs of the differences between the pairs of observations in the paired-sample case; it does not take into consideration the magnitudes of these differences. A test utilizing both direction and magnitude, proposed in 1945 by Frank Wilcoxon, is now commonly referred to as the **Wilcoxon signed-rank test**.

The analyst can extract more information from the data in a nonparametric fashion if it is reasonable to invoke an additional restriction on the distribution from which the data were taken. The Wilcoxon signed-rank test applies in the case of a **symmetric continuous distribution**. Under this condition, we can test the null hypothesis $\tilde{\mu} = \tilde{\mu}_0$. We first subtract $\tilde{\mu}_0$ from each sample value, discarding all differences equal to zero. The remaining differences are then ranked without regard to sign. A rank of 1 is assigned to the smallest absolute difference (i.e., without sign), a rank of 2 to the next smallest, and so on. When the absolute value of two or more differences is the same, assign to each the average of the ranks that would have been assigned if the differences were distinguishable. For example, if the fifth and sixth smallest differences are equal in absolute value, each is assigned a rank of 5.5. If the hypothesis $\tilde{\mu} = \tilde{\mu}_0$ is true, the total of the ranks corresponding to the positive differences should nearly equal the total of the ranks corresponding to the negative differences. Let us represent these totals by w_+ and w_-, respectively. We designate the smaller of w_+ and w_- by w.

In selecting repeated samples, we would expect w_+ and w_-, and therefore w, to vary. Thus, we may think of w_+, w_-, and w as values of the corresponding random variables W_+, W_-, and W. The null hypothesis $\tilde{\mu} = \tilde{\mu}_0$ can be rejected in favor of the alternative $\tilde{\mu} < \tilde{\mu}_0$ only if w_+ is small and w_- is large. Likewise, the alternative $\tilde{\mu} > \tilde{\mu}_0$ can be accepted only if w_+ is large and w_- is small. For a two-sided alternative, we may reject H_0 in favor of H_1 if either w_+ or w_-, and hence w, is sufficiently small. Therefore, no matter what the alternative hypothesis

may be, we reject the null hypothesis when the value of the appropriate statistic W_+, W_-, or W is sufficiently small.

Two Samples with Paired Observations

To test the null hypothesis that we are sampling two continuous symmetric populations with $\tilde{\mu}_1 = \tilde{\mu}_2$ for the paired-sample case, we rank the differences of the paired observations without regard to sign and proceed as in the single-sample case. The various test procedures for both the single- and paired-sample cases are summarized in Table 16.2.

Table 16.2: Signed-Rank Test

H_0	H_1	Compute
$\tilde{\mu} = \tilde{\mu}_0$	$\tilde{\mu} < \tilde{\mu}_0$	w_+
	$\tilde{\mu} > \tilde{\mu}_0$	w_-
	$\tilde{\mu} \neq \tilde{\mu}_0$	w
$\tilde{\mu}_1 = \tilde{\mu}_2$	$\tilde{\mu}_1 < \tilde{\mu}_2$	w_+
	$\tilde{\mu}_1 > \tilde{\mu}_2$	w_-
	$\tilde{\mu}_1 \neq \tilde{\mu}_2$	w

It is not difficult to show that whenever $n < 5$ and the level of significance does not exceed 0.05 for a one-tailed test or 0.10 for a two-tailed test, all possible values of w_+, w_-, or w will lead to the acceptance of the null hypothesis. However, when $5 \leq n \leq 30$, Table A.16 shows approximate critical values of W_+ and W_- for levels of significance equal to 0.01, 0.025, and 0.05 for a one-tailed test and critical values of W for levels of significance equal to 0.02, 0.05, and 0.10 for a two-tailed test. The null hypothesis is rejected if the computed value w_+, w_-, or w is **less than or equal to** the appropriate tabled value. For example, when $n = 12$, Table A.16 shows that a value of $w_+ \leq 17$ is required for the one-sided alternative $\tilde{\mu} < \tilde{\mu}_0$ to be significant at the 0.05 level.

Example 16.3: Rework Example 16.1 by using the signed-rank test.

***Solution*:**

1. H_0: $\tilde{\mu} = 1.8$.
2. H_1: $\tilde{\mu} \neq 1.8$.
3. $\alpha = 0.05$.
4. Critical region: Since $n = 10$ after discarding the one measurement that equals 1.8, Table A.16 shows the critical region to be $w \leq 8$.
5. Computations: Subtracting 1.8 from each measurement and then ranking the differences without regard to sign, we have

d_i	−0.3	0.4	−0.9	−0.5	0.2	−0.2	−0.3	0.2	−0.6	−0.1
Ranks	5.5	7	10	8	3	3	5.5	3	9	1

Now $w_+ = 13$ and $w_- = 42$, so $w = 13$, the smaller of w_+ and w_-.

6. Decision: As before, do not reject H_0 and conclude that the median operating time is not significantly different from 1.8 hours.

The signed-rank test can also be used to test the null hypothesis that $\tilde{\mu}_1 - \tilde{\mu}_2 = d_0$. In this case, the populations need not be symmetric. As with the sign test, we subtract d_0 from each difference, rank the adjusted differences without regard to sign, and apply the same procedure as above.

Example 16.4: It is claimed that a college senior can increase his or her score in the major field area of the graduate record examination by at least 50 points if he or she is provided with sample problems in advance. To test this claim, 20 college seniors are divided into 10 pairs such that the students in each matched pair have almost the same overall grade-point averages for their first 3 years in college. Sample problems and answers are provided at random to one member of each pair 1 week prior to the examination. The examination scores are given in Table 16.3.

Table 16.3: Data for Example 16.4

	Pair									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
With Sample Problems	531	621	663	579	451	660	591	719	543	575
Without Sample Problems	509	540	688	502	424	683	568	748	530	524

Test the null hypothesis, at the 0.05 level of significance, that sample problems increase scores by 50 points against the alternative hypothesis that the increase is less than 50 points.

***Solution*:** Let $\tilde{\mu}_1$ and $\tilde{\mu}_2$ represent the median scores of all students taking the test in question with and without sample problems, respectively.

1. H_0: $\tilde{\mu}_1 - \tilde{\mu}_2 = 50$.
2. H_1: $\tilde{\mu}_1 - \tilde{\mu}_2 < 50$.
3. $\alpha = 0.05$.
4. Critical region: Since $n = 10$, Table A.16 shows the critical region to be $w_+ \leq 11$.
5. Computations:

	Pair									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
d_i	22	81	−25	77	27	−23	23	−29	13	51
$d_i - d_0$	−28	31	−75	27	−23	−73	−27	−79	−37	1
Ranks	5	6	9	3.5	2	8	3.5	10	7	1

Now we find that $w_+ = 6 + 3.5 + 1 = 10.5$.

6. Decision: Reject H_0 and conclude that sample problems do not, on average, increase one's graduate record score by as much as 50 points.

Normal Approximation for Large Samples

When $n \geq 15$, the sampling distribution of W_+ (or W_-) approaches the normal distribution with mean and variance given by

$$\mu_{W_+} = \frac{n(n+1)}{4} \text{ and } \sigma^2_{W_+} = \frac{n(n+1)(2n+1)}{24}.$$

Therefore, when n exceeds the largest value in Table A.16, the statistic

$$Z = \frac{W_+ - \mu_{W_+}}{\sigma_{W_+}}$$

can be used to determine the critical region for the test.

Exercises

16.1 The following data represent the time, in minutes, that a patient has to wait during 12 visits to a doctor's office before being seen by the doctor:

17	15	20	20	32	28
12	26	25	25	35	24

Use the sign test at the 0.05 level of significance to test the doctor's claim that the median waiting time for her patients is not more than 20 minutes.

16.2 The following data represent the number of hours of flight training received by 18 student pilots from a certain instructor prior to their first solo flight:

9	12	18	14	12	14	12	10	16
11	9	11	13	11	13	15	13	14

Using binomial probabilities from Table A.1, perform a sign test at the 0.02 level of significance to test the instructor's claim that the median time required before his students' solo is 12 hours of flight training.

16.3 A food inspector examined 16 jars of a certain brand of jam to determine the percent of foreign impurities. The following data were recorded:

2.4	2.3	3.1	2.2	2.3	1.2	1.0	2.4
1.7	1.1	4.2	1.9	1.7	3.6	1.6	2.3

Using the normal approximation to the binomial distribution, perform a sign test at the 0.05 level of significance to test the null hypothesis that the median percent of impurities in this brand of jam is 2.5% against the alternative that the median percent of impurities is not 2.5%.

16.4 A paint supplier claims that a new additive will reduce the drying time of its acrylic paint. To test this claim, 12 panels of wood were painted, one-half of each panel with paint containing the regular additive and the other half with paint containing the new additive. The drying times, in hours, were recorded as follows:

	Drying Time (hours)	
Panel	**New Additive**	**Regular Additive**
1	6.4	6.6
2	5.8	5.8
3	7.4	7.8
4	5.5	5.7
5	6.3	6.0
6	7.8	8.4
7	8.6	8.8
8	8.2	8.4
9	7.0	7.3
10	4.9	5.8
11	5.9	5.8
12	6.5	6.5

Use the sign test at the 0.05 level to test the null hypothesis that the new additive is no better than the regular additive in reducing the drying time of this kind of paint.

16.5 It is claimed that a new diet will reduce a person's weight by 4.5 kilograms, on average, in a period of 2 weeks. The weights of 10 women were recorded before and after a 2-week period during which they followed this diet, yielding the following data:

Woman	**Weight Before**	**Weight After**
1	58.5	60.0
2	60.3	54.9
3	61.7	58.1
4	69.0	62.1
5	64.0	58.5
6	62.6	59.9
7	56.7	54.4
8	63.6	60.2
9	68.2	62.3
10	59.4	58.7

Use the sign test at the 0.05 level of significance to test the hypothesis that the diet reduces the median

weight by 4.5 kilograms against the alternative hypothesis that the median weight loss is less than 4.5 kilograms.

16.6 Two types of instruments for measuring the amount of sulfur monoxide in the atmosphere are being compared in an air-pollution experiment. The following readings were recorded daily for a period of 2 weeks:

	Sulfur Monoxide	
Day	**Instrument *A***	**Instrument *B***
1	0.96	0.87
2	0.82	0.74
3	0.75	0.63
4	0.61	0.55
5	0.89	0.76
6	0.64	0.70
7	0.81	0.69
8	0.68	0.57
9	0.65	0.53
10	0.84	0.88
11	0.59	0.51
12	0.94	0.79
13	0.91	0.84
14	0.77	0.63

Using the normal approximation to the binomial distribution, perform a sign test to determine whether the different instruments lead to different results. Use a 0.05 level of significance.

16.7 The following figures give the systolic blood pressure of 16 joggers before and after an 8-kilometer run:

Jogger	**Before**	**After**
1	158	164
2	149	158
3	160	163
4	155	160
5	164	172
6	138	147
7	163	167
8	159	169
9	165	173
10	145	147
11	150	156
12	161	164
13	132	133
14	155	161
15	146	154
16	159	170

Use the sign test at the 0.05 level of significance to test the null hypothesis that jogging 8 kilometers increases the median systolic blood pressure by 8 points against the alternative that the increase in the median is less than 8 points.

16.8 Analyze the data of Exercise 16.1 by using the signed-rank test.

16.9 Analyze the data of Exercise 16.2 by using the signed-rank test.

16.10 The weights of 5 people before they stopped smoking and 5 weeks after they stopped smoking, in kilograms, are as follows:

	Individual				
	1	**2**	**3**	**4**	**5**
Before	66	80	69	52	75
After	71	82	68	56	73

Use the signed-rank test for paired observations to test the hypothesis, at the 0.05 level of significance, that giving up smoking has no effect on a person's weight against the alternative that one's weight increases if he or she quits smoking.

16.11 Rework Exercise 16.5 by using the signed-rank test.

16.12 The following are the numbers of prescriptions filled by two pharmacies over a 20-day period:

Day	**Pharmacy *A***	**Pharmacy *B***
1	19	17
2	21	15
3	15	12
4	17	12
5	24	16
6	12	15
7	19	11
8	14	13
9	20	14
10	18	21
11	23	19
12	21	15
13	17	11
14	12	10
15	16	20
16	15	12
17	20	13
18	18	17
19	14	16
20	22	18

Use the signed-rank test at the 0.01 level of significance to determine whether the two pharmacies, on average, fill the same number of prescriptions against the alternative that pharmacy *A* fills more prescriptions than pharmacy *B*.

16.13 Rework Exercise 16.7 by using the signed-rank test.

16.14 Rework Exercise 16.6 by using the signed-rank test.

16.3 Wilcoxon Rank-Sum Test

As we indicated earlier, the nonparametric procedure is generally an appropriate alternative to the normal theory test when the normality assumption does not hold. When we are interested in testing equality of means of two continuous distributions that are obviously nonnormal, and samples are independent (i.e., there is no pairing of observations), the **Wilcoxon rank-sum test** or **Wilcoxon two-sample test** is an appropriate alternative to the two-sample t-test described in Chapter 10.

We shall test the null hypothesis H_0 that $\tilde{\mu}_1 = \tilde{\mu}_2$ against some suitable alternative. First we select a random sample from each of the populations. Let n_1 be the number of observations in the smaller sample, and n_2 the number of observations in the larger sample. When the samples are of equal size, n_1 and n_2 may be randomly assigned. Arrange the $n_1 + n_2$ observations of the combined samples in ascending order and substitute a rank of $1, 2, \ldots, n_1 + n_2$ for each observation. In the case of ties (identical observations), we replace the observations by the mean of the ranks that the observations would have if they were distinguishable. For example, if the seventh and eighth observations were identical, we would assign a rank of 7.5 to each of the two observations.

The sum of the ranks corresponding to the n_1 observations in the smaller sample is denoted by w_1. Similarly, the value w_2 represents the sum of the n_2 ranks corresponding to the larger sample. The total $w_1 + w_2$ depends only on the number of observations in the two samples and is in no way affected by the results of the experiment. Hence, if $n_1 = 3$ and $n_2 = 4$, then $w_1 + w_2 = 1 + 2 + \cdots + 7 = 28$, regardless of the numerical values of the observations. In general,

$$w_1 + w_2 = \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2},$$

the arithmetic sum of the integers $1, 2, \ldots, n_1 + n_2$. Once we have determined w_1, it may be easier to find w_2 by the formula

$$w_2 = \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2} - w_1.$$

In choosing repeated samples of sizes n_1 and n_2, we would expect w_1, and therefore w_2, to vary. Thus, we may think of w_1 and w_2 as values of the random variables W_1 and W_2, respectively. The null hypothesis $\tilde{\mu}_1 = \tilde{\mu}_2$ will be rejected in favor of the alternative $\tilde{\mu}_1 < \tilde{\mu}_2$ only if w_1 is small and w_2 is large. Likewise, the alternative $\tilde{\mu}_1 > \tilde{\mu}_2$ can be accepted only if w_1 is large and w_2 is small. For a two-tailed test, we may reject H_0 in favor of H_1 if w_1 is small and w_2 is large or if w_1 is large and w_2 is small. In other words, the alternative $\tilde{\mu}_1 < \tilde{\mu}_2$ is accepted if w_1 is sufficiently small; the alternative $\tilde{\mu}_1 > \tilde{\mu}_2$ is accepted if w_2 is sufficiently small; and the alternative $\tilde{\mu}_1 \neq \tilde{\mu}_2$ is accepted if the minimum of w_1 and w_2 is sufficiently small. In actual practice, we usually base our decision on the value

$$u_1 = w_1 - \frac{n_1(n_1 + 1)}{2} \qquad \text{or} \qquad u_2 = w_2 - \frac{n_2(n_2 + 1)}{2}$$

of the related statistic U_1 or U_2 or on the value u of the statistic U, the minimum of U_1 and U_2. These statistics simplify the construction of tables of critical values,

since both U_1 and U_2 have symmetric sampling distributions and assume values in the interval from 0 to n_1n_2 such that $u_1 + u_2 = n_1n_2$.

From the formulas for u_1 and u_2 we see that u_1 will be small when w_1 is small and u_2 will be small when w_2 is small. Consequently, the null hypothesis will be rejected whenever the appropriate statistic U_1, U_2, or U assumes a value less than or equal to the desired critical value given in Table A.17. The various test procedures are summarized in Table 16.4.

Table 16.4: Rank-Sum Test

H_0	H_1	Compute
$\tilde{\mu}_1 = \tilde{\mu}_2$	$\tilde{\mu}_1 < \tilde{\mu}_2$	u_1
	$\tilde{\mu}_1 > \tilde{\mu}_2$	u_2
	$\tilde{\mu}_1 \neq \tilde{\mu}_2$	u

Table A.17 gives critical values of U_1 and U_2 for levels of significance equal to 0.001, 0.01, 0.025, and 0.05 for a one-tailed test, and critical values of U for levels of significance equal to 0.002, 0.02, 0.05, and 0.10 for a two-tailed test. If the observed value of u_1, u_2, or u is **less than or equal** to the tabled critical value, the null hypothesis is rejected at the level of significance indicated by the table. Suppose, for example, that we wish to test the null hypothesis that $\tilde{\mu}_1 = \tilde{\mu}_2$ against the one-sided alternative that $\tilde{\mu}_1 < \tilde{\mu}_2$ at the 0.05 level of significance for random samples of sizes $n_1 = 3$ and $n_2 = 5$ that yield the value $w_1 = 8$. It follows that

$$u_1 = 8 - \frac{(3)(4)}{2} = 2.$$

Our one-tailed test is based on the statistic U_1. Using Table A.17, we reject the null hypothesis of equal means when $u_1 \leq 1$. Since $u_1 = 2$ does not fall in the rejection region, the null hypothesis cannot be rejected.

Example 16.5: The nicotine content of two brands of cigarettes, measured in milligrams, was found to be as follows:

Brand A	2.1	4.0	6.3	5.4	4.8	3.7	6.1	3.3		
Brand B	4.1	0.6	3.1	2.5	4.0	6.2	1.6	2.2	1.9	5.4

Test the hypothesis, at the 0.05 level of significance, that the median nicotine contents of the two brands are equal against the alternative that they are unequal.

Solution:

1. H_0: $\tilde{\mu}_1 = \tilde{\mu}_2$.
2. H_1: $\tilde{\mu}_1 \neq \tilde{\mu}_2$.
3. $\alpha = 0.05$.
4. Critical region: $u \leq 17$ (from Table A.17).
5. Computations: The observations are arranged in ascending order and ranks from 1 to 18 assigned.

Original Data	Ranks	Original Data	Ranks
0.6	1	4.0	10.5*
1.6	2	4.0	10.5
1.9	3	4.1	12
2.1	4*	4.8	13*
2.2	5	5.4	14.5*
2.5	6	5.4	14.5
3.1	7	6.1	16*
3.3	8*	6.2	17
3.7	9*	6.3	18*

*The ranks marked with an asterisk belong to sample A.

Now

$$w_1 = 4 + 8 + 9 + 10.5 + 13 + 14.5 + 16 + 18 = 93$$

and

$$w_2 = \frac{(18)(19)}{2} - 93 = 78.$$

Therefore,

$$u_1 = 93 - \frac{(8)(9)}{2} = 57, \qquad u_2 = 78 - \frac{(10)(11)}{2} = 23.$$

6. Decision: Do not reject the null hypothesis H_0 and conclude that there is no significant difference in the median nicotine contents of the two brands of cigarettes.

Normal Theory Approximation for Two Samples

When both n_1 and n_2 exceed 8, the sampling distribution of U_1 (or U_2) approaches the normal distribution with mean and variance given by

$$\mu_{U_1} = \frac{n_1 n_2}{2} \text{ and } \sigma^2_{U_1} = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}.$$

Consequently, when n_2 is greater than 20, the maximum value in Table A.17, and n_1 is at least 9, we can use the statistic

$$Z = \frac{U_1 - \mu_{U_1}}{\sigma_{U_1}}$$

for our test, with the critical region falling in either or both tails of the standard normal distribution, depending on the form of H_1.

The use of the Wilcoxon rank-sum test is not restricted to nonnormal populations. It can be used in place of the two-sample t-test when the populations are normal, although the power will be smaller. The Wilcoxon rank-sum test is always superior to the t-test for decidedly nonnormal populations.

16.4 Kruskal-Wallis Test

In Chapters 13, 14, and 15, the technique of analysis of variance was prominent as an analytical technique for testing equality of $k \geq 2$ population means. Again, however, the reader should recall that normality must be assumed in order for the F-test to be theoretically correct. In this section, we investigate a nonparametric alternative to analysis of variance.

The **Kruskal-Wallis test**, also called the **Kruskal-Wallis *H* test**, is a generalization of the rank-sum test to the case of $k > 2$ samples. It is used to test the null hypothesis H_0 that k independent samples are from identical populations. Introduced in 1952 by W. H. Kruskal and W. A. Wallis, the test is a nonparametric procedure for testing the equality of means in the one-factor analysis of variance when the experimenter wishes to avoid the assumption that the samples were selected from normal populations.

Let n_i $(i = 1, 2, \ldots, k)$ be the number of observations in the ith sample. First, we combine all k samples and arrange the $n = n_1 + n_2 + \cdots + n_k$ observations in ascending order, substituting the appropriate rank from $1, 2, \ldots, n$ for each observation. In the case of ties (identical observations), we follow the usual procedure of replacing the observations by the mean of the ranks that the observations would have if they were distinguishable. The sum of the ranks corresponding to the n_i observations in the ith sample is denoted by the random variable R_i. Now let us consider the statistic

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1),$$

which is approximated very well by a chi-squared distribution with $k-1$ degrees of freedom when H_0 is true, provided each sample consists of at least 5 observations. The fact that h, the assumed value of H, is large when the independent samples come from populations that are not identical allows us to establish the following decision criterion for testing H_0:

Kruskal-Wallis Test

To test the null hypothesis H_0 that k independent samples are from identical populations, compute

$$h = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{r_i^2}{n_i} - 3(n+1),$$

where r_i is the assumed value of R_i, for $i = 1, 2, \ldots, k$. If h falls in the critical region $H > \chi^2_\alpha$ with $v = k - 1$ degrees of freedom, reject H_0 at the α-level of significance; otherwise, fail to reject H_0.

Example 16.6: In an experiment to determine which of three different missile systems is preferable, the propellant burning rate is measured. The data, after coding, are given in Table 16.5. Use the Kruskal-Wallis test and a significance level of $\alpha = 0.05$ to test the hypothesis that the propellant burning rates are the same for the three missile systems.

Table 16.5: Propellant Burning Rates

Missile System								
1			**2**			**3**		
24.0	16.7	22.8	23.2	19.8	18.1	18.4	19.1	17.3
19.8	18.9		17.6	20.2	17.8	17.3	19.7	18.9
						18.8	19.3	

***Solution*:** 1. H_0: $\mu_1 = \mu_2 = \mu_3$.

2. H_1: The three means are not all equal.

3. $\alpha = 0.05$.

4. Critical region: $h > \chi^2_{0.05} = 5.991$, for $v = 2$ degrees of freedom.

5. Computations: In Table 16.6, we convert the 19 observations to ranks and sum the ranks for each missile system.

Table 16.6: Ranks for Propellant Burning Rates

Missile System		
1	**2**	**3**
19	18	7
1	14.5	11
17	6	2.5
14.5	4	2.5
9.5	16	13
$r_1 = 61.0$	5	9.5
	$r_2 = 63.5$	8
		12
		$r_3 = 65.5$

Now, substituting $n_1 = 5$, $n_2 = 6$, $n_3 = 8$ and $r_1 = 61.0$, $r_2 = 63.5$, $r_3 = 65.5$, our test statistic H assumes the value

$$h = \frac{12}{(19)(20)}\left(\frac{61.0^2}{5} + \frac{63.5^2}{6} + \frac{65.5^2}{8}\right) - (3)(20) = 1.66.$$

6. Decision: Since $h = 1.66$ does not fall in the critical region $h > 5.991$, we have insufficient evidence to reject the hypothesis that the propellant burning rates are the same for the three missile systems. ◢

Exercises

16.15 A cigarette manufacturer claims that the tar content of brand B cigarettes is lower than that of brand A cigarettes. To test this claim, the following determinations of tar content, in milligrams, were recorded:

Brand A	1	12	9	13	11	14
Brand B	8	10	7			

Use the rank-sum test with $\alpha = 0.05$ to test whether the claim is valid.

16.16 To find out whether a new serum will arrest leukemia, nine patients, who have all reached an advanced stage of the disease, are selected. Five patients receive the treatment and four do not. The survival times, in years, from the time the experiment commenced are

Treatment	2.1	5.3	1.4	4.6	0.9
No treatment	1.9	0.5	2.8	3.1	

Use the rank-sum test, at the 0.05 level of significance, to determine if the serum is effective.

16.17 The following data represent the number of hours that two different types of scientific pocket calculators operate before a recharge is required.

Calculator A	5.5	5.6	6.3	4.6	5.3	5.0	6.2	5.8	5.1
Calculator B	3.8	4.8	4.3	4.2	4.0	4.9	4.5	5.2	4.5

Use the rank-sum test with $\alpha = 0.01$ to determine if calculator A operates longer than calculator B on a full battery charge.

16.18 A fishing line is being manufactured by two processes. To determine if there is a difference in the mean breaking strength of the lines, 10 pieces manufactured by each process are selected and then tested for breaking strength. The results are as follows:

Process 1	10.4	9.8	11.5	10.0	9.9
	9.6	10.9	11.8	9.3	10.7
Process 2	8.7	11.2	9.8	10.1	10.8
	9.5	11.0	9.8	10.5	9.9

Use the rank-sum test with $\alpha = 0.1$ to determine if there is a difference between the mean breaking strengths of the lines manufactured by the two processes.

16.19 From a mathematics class of 12 equally capable students using programmed materials, 5 students are selected at random and given additional instruction by the teacher. The results on the final examination are as follows:

	Grade						
Additional Instruction	87	69	78	91	80		
No Additional Instruction	75	88	64	82	93	79	67

Use the rank-sum test with $\alpha = 0.05$ to determine if the additional instruction affects the average grade.

16.20 The following data represent the weights, in kilograms, of personal luggage carried on various flights by a member of a baseball team and a member of a basketball team.

Luggage Weight (kilograms)				
Baseball Player			**Basketball Player**	
16.3	20.0	18.6	15.4	16.3
18.1	15.0	15.4	17.7	18.1
15.9	18.6	15.6	18.6	16.8
14.1	14.5	18.3	12.7	14.1
17.7	19.1	17.4	15.0	13.6
16.3	13.6	14.8	15.9	16.3
13.2	17.2	16.5		

Use the rank-sum test with $\alpha = 0.05$ to test the null hypothesis that the two athletes carry the same amount of luggage on the average against the alternative hypothesis that the average weights of luggage for the two athletes are different.

16.21 The following data represent the operating times in hours for three types of scientific pocket calculators before a recharge is required:

Calculator								
A			***B***			***C***		
4.9	6.1	4.3	5.5	5.4	6.2	6.4	6.8	5.6
4.6	5.2		5.8	5.5	5.2	6.5	6.3	6.6
				4.8				

Use the Kruskal-Wallis test, at the 0.01 level of significance, to test the hypothesis that the operating times for all three calculators are equal.

16.22 In Exercise 13.6 on page 519, use the Kruskal-Wallis test at the 0.05 level of significance to determine if the organic chemical solvents differ significantly in sorption rate.

16.5 Runs Test

In applying the many statistical concepts discussed throughout this book, it was always assumed that the sample data had been collected by some randomization procedure. The **runs test**, based on the order in which the sample observations are obtained, is a useful technique for testing the null hypothesis H_0 that the observations have indeed been drawn at random.

To illustrate the runs test, let us suppose that 12 people are polled to find out if they use a certain product. We would seriously question the assumed randomness of the sample if all 12 people were of the same sex. We shall designate a male and a female by the symbols M and F, respectively, and record the outcomes according to their sex in the order in which they occur. A typical sequence for the experiment might be

$$\underbrace{M\ M}\ \underbrace{F\ F\ F}\ \underbrace{M}\ \underbrace{F\ F}\ \underbrace{M\ M\ M\ M},$$

where we have grouped subsequences of identical symbols. Such groupings are called **runs**.

Definition 16.1: A **run** is a subsequence of one or more identical symbols representing a common property of the data.

Regardless of whether the sample measurements represent qualitative or quantitative data, the runs test divides the data into two mutually exclusive categories: male or female; defective or nondefective; heads or tails; above or below the median; and so forth. Consequently, a sequence will always be limited to two distinct symbols. Let n_1 be the number of symbols associated with the category that occurs the least and n_2 be the number of symbols that belong to the other category. Then the sample size $n = n_1 + n_2$.

For the $n = 12$ symbols in our poll, we have five runs, with the first containing two M's, the second containing three F's, and so on. If the number of runs is larger or smaller than what we would expect by chance, the hypothesis that the sample was drawn at random should be rejected. Certainly, a sample resulting in only two runs,

$$M\ M\ M\ M\ M\ M\ M\ F\ F\ F\ F\ F$$

or the reverse, is most unlikely to occur from a random selection process. Such a result indicates that the first 7 people interviewed were all males, followed by 5 females. Likewise, if the sample resulted in the maximum number of 12 runs, as in the alternating sequence

$$M\ F\ M\ F\ M\ F\ M\ F\ M\ F\ M\ F,$$

we would again be suspicious of the order in which the individuals were selected for the poll.

The runs test for randomness is based on the random variable V, the total number of runs that occur in the complete sequence of the experiment. In Table A.18, values of $P(V \leq v^*$ when H_0 is true) are given for $v^* = 2, 3, \ldots, 20$ runs and

16.6 Tolerance Limits

Tolerance limits for a normal distribution of measurements were discussed in Chapter 9. In this section, we consider a method for constructing tolerance intervals that is independent of the shape of the underlying distribution. As we might suspect, for a reasonable degree of confidence they will be substantially longer than those constructed assuming normality, and the sample size required is generally very large. Nonparametric tolerance limits are stated in terms of the smallest and largest observations in our sample.

Two-Sided Tolerance Limits

For any distribution of measurements, two-sided tolerance limits are indicated by the smallest and largest observations in a sample of size n, where n is determined so that one can assert with $100(1-\gamma)\%$ confidence that **at least** the proportion $1-\alpha$ of the distribution is included between the sample extremes.

Table A.19 gives required sample sizes for selected values of γ and $1-\alpha$. For example, when $\gamma = 0.01$ and $1-\alpha = 0.95$, we must choose a random sample of size $n = 130$ in order to be 99% confident that at least 95% of the distribution of measurements is included between the sample extremes.

Instead of determining the sample size n such that a specified proportion of measurements is contained between the sample extremes, it is desirable in many industrial processes to determine the sample size such that a fixed proportion of the population falls below the largest (or above the smallest) observation in the sample. Such limits are called one-sided tolerance limits.

One-Sided Tolerance Limits

For any distribution of measurements, a one-sided tolerance limit is determined by the smallest (largest) observation in a sample of size n, where n is determined so that one can assert with $100(1-\gamma)\%$ confidence that **at least** the proportion $1-\alpha$ of the distribution will exceed the smallest (be less than the largest) observation in the sample.

Table A.20 shows required sample sizes corresponding to selected values of γ and $1-\alpha$. Hence, when $\gamma = 0.05$ and $1-\alpha = 0.70$, we must choose a sample of size $n = 9$ in order to be 95% confident that 70% of our distribution of measurements will exceed the smallest observation in the sample.

16.7 Rank Correlation Coefficient

In Chapter 11, we used the sample correlation coefficient r to measure the population correlation coefficient ρ, the linear relationship between two continuous variables X and Y. If ranks $1, 2, \ldots, n$ are assigned to the x observations in order of magnitude and similarly to the y observations, and if these ranks are then substituted for the actual numerical values in the formula for the correlation coefficient in Chapter 11, we obtain the nonparametric counterpart of the conventional correlation coefficient. A correlation coefficient calculated in this manner is known as the **Spearman rank correlation coefficient** and is denoted by r_s. When there are no ties among either set of measurements, the formula for r_s reduces to a much simpler expression involving the differences d_i between the ranks assigned to the n pairs of x's and y's, which we now state.

Rank Correlation Coefficient

A nonparametric measure of association between two variables X and Y is given by the **rank correlation coefficient**

$$r_s = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} d_i^2,$$

where d_i is the difference between the ranks assigned to x_i and y_i and n is the number of pairs of data.

In practice, the preceding formula is also used when there are ties among either the x or y observations. The ranks for tied observations are assigned as in the signed-rank test by averaging the ranks that would have been assigned if the observations were distinguishable.

The value of r_s will usually be close to the value obtained by finding r based on numerical measurements and is interpreted in much the same way. As before, the value of r_s will range from -1 to $+1$. A value of $+1$ or -1 indicates perfect association between X and Y, the plus sign occurring for identical rankings and the minus sign occurring for reverse rankings. When r_s is close to zero, we conclude that the variables are uncorrelated.

Example 16.8: The figures listed in Table 16.7, released by the Federal Trade Commission, show the milligrams of tar and nicotine found in 10 brands of cigarettes. Calculate the rank correlation coefficient to measure the degree of relationship between tar and nicotine content in cigarettes.

Table 16.7: Tar and Nicotine Contents

Cigarette Brand	Tar Content	Nicotine Content
Viceroy	14	0.9
Marlboro	17	1.1
Chesterfield	28	1.6
Kool	17	1.3
Kent	16	1.0
Raleigh	13	0.8
Old Gold	24	1.5
Philip Morris	25	1.4
Oasis	18	1.2
Players	31	2.0

***Solution*:** Let X and Y represent the tar and nicotine contents, respectively. First we assign ranks to each set of measurements, with the rank of 1 assigned to the lowest number in each set, the rank of 2 to the second lowest number in each set, and so forth, until the rank of 10 is assigned to the largest number. Table 16.8 shows the individual rankings of the measurements and the differences in ranks for the 10 pairs of observations.

Table 16.8: Rankings for Tar and Nicotine Content

Cigarette Brand	x_i	y_i	d_i
Viceroy	2.0	2.0	0.0
Marlboro	4.5	4.0	0.5
Chesterfield	9.0	9.0	0.0
Kool	4.5	6.0	−1.5
Kent	3.0	3.0	0.0
Raleigh	1.0	1.0	0.0
Old Gold	7.0	8.0	−1.0
Philip Morris	8.0	7.0	1.0
Oasis	6.0	5.0	1.0
Players	10.0	10.0	0.0

Substituting into the formula for r_s, we find that

$$r_s = 1 - \frac{(6)(5.50)}{(10)(100-1)} = 0.967,$$

indicating a high positive correlation between the amounts of tar and nicotine found in cigarettes.

Some advantages to using r_s rather than r do exist. For instance, we no longer assume the underlying relationship between X and Y to be linear and therefore, when the data possess a distinct curvilinear relationship, the rank correlation coefficient will likely be more reliable than the conventional measure. A second advantage to using the rank correlation coefficient is the fact that no assumptions of normality are made concerning the distributions of X and Y. Perhaps the greatest advantage occurs when we are unable to make meaningful numerical measurements but nevertheless can establish rankings. Such is the case, for example, when different judges rank a group of individuals according to some attribute. The rank correlation coefficient can be used in this situation as a measure of the consistency of the two judges.

To test the hypothesis that $\rho = 0$ by using a rank correlation coefficient, one needs to consider the sampling distribution of the r_s-values under the assumption of no correlation. Critical values for $\alpha = 0.05, 0.025, 0.01$, and 0.005 have been calculated and appear in Table A.21. The setup of this table is similar to that of the table of critical values for the t-distribution except for the left column, which now gives the number of pairs of observations rather than the degrees of freedom. Since the distribution of the r_s-values is symmetric about zero when $\rho = 0$, the r_s-value that leaves an area of α to the left is equal to the negative of the r_s-value that leaves an area of α to the right. For a two-sided alternative hypothesis, the critical region of size α falls equally in the two tails of the distribution. For a test in which the alternative hypothesis is negative, the critical region is entirely in the left tail of the distribution, and when the alternative is positive, the critical region is placed entirely in the right tail.

Example 16.9: Refer to Example 16.8 and test the hypothesis that the correlation between the amounts of tar and nicotine found in cigarettes is zero against the alternative that it is greater than zero. Use a 0.01 level of significance.

Solution:

1. H_0: $\rho = 0$.
2. H_1: $\rho > 0$.
3. $\alpha = 0.01$.
4. Critical region: $r_s > 0.745$ from Table A.21.
5. Computations: From Example 16.8, $r_s = 0.967$.
6. Decision: Reject H_0 and conclude that there is a significant correlation between the amounts of tar and nicotine found in cigarettes.

Under the assumption of no correlation, it can be shown that the distribution of the r_s-values approaches a normal distribution with a mean of 0 and a standard deviation of $1/\sqrt{n-1}$ as n increases. Consequently, when n exceeds the values given in Table A.21, one can test for a significant correlation by computing

$$z = \frac{r_s - 0}{1/\sqrt{n-1}} = r_s\sqrt{n-1}$$

and comparing with critical values of the standard normal distribution shown in Table A.3.

Exercises

16.23 A random sample of 15 adults living in a small town were selected to estimate the proportion of voters favoring a certain candidate for mayor. Each individual was also asked if he or she was a college graduate. By letting Y and N designate the responses of "yes" and "no" to the education question, the following sequence was obtained:

$$N\ N\ N\ N\ Y\ Y\ N\ Y\ Y\ N\ Y\ N\ N\ N\ N$$

Use the runs test at the 0.1 level of significance to determine if the sequence supports the contention that the sample was selected at random.

16.24 A silver-plating process is used to coat a certain type of serving tray. When the process is in control, the thickness of the silver on the trays will vary randomly following a normal distribution with a mean of 0.02 millimeter and a standard deviation of 0.005 millimeter. Suppose that the next 12 trays examined show the following thicknesses of silver: 0.019, 0.021, 0.020, 0.019, 0.020, 0.018, 0.023, 0.021, 0.024, 0.022, 0.023, 0.022. Use the runs test to determine if the fluctuations in thickness from one tray to another are random. Let $\alpha = 0.05$.

16.25 Use the runs test to test, at level 0.01, whether there is a difference in the average operating time for the two calculators of Exercise 16.17 on page 670.

16.26 In an industrial production line, items are inspected periodically for defectives. The following is a sequence of defective items, D, and nondefective items, N, produced by this production line:

$$D\ D\ N\ N\ N\ D\ N\ N\ D\ D\ N\ N\ N\ N$$
$$N\ D\ D\ D\ N\ N\ D\ N\ N\ N\ N\ D\ N\ D$$

Use the large-sample theory for the runs test, with a significance level of 0.05, to determine whether the defectives are occurring at random.

16.27 Assuming that the measurements of Exercise 1.14 on page 30 were recorded successively from left to right as they were collected, use the runs test, with $\alpha = 0.05$, to test the hypothesis that the data represent a random sequence.

16.28 How large a sample is required to be 95% confident that at least 85% of the distribution of measurements is included between the sample extremes?

16.29 What is the probability that the range of a random sample of size 24 includes at least 90% of the population?

16.30 How large a sample is required to be 99% confident that at least 80% of the population will be less than the largest observation in the sample?

16.31 What is the probability that at least 95% of a population will exceed the smallest value in a random sample of size $n = 135$?

16.32 The following table gives the recorded grades for 10 students on a midterm test and the final examination in a calculus course:

Student	Midterm Test	Final Examination
L.S.A.	84	73
W.P.B.	98	63
R.W.K.	91	87
J.R.L.	72	66
J.K.L.	86	78
D.L.P.	93	78
B.L.P.	80	91
D.W.M.	0	0
M.N.M.	92	88
R.H.S.	87	77

(a) Calculate the rank correlation coefficient.

(b) Test the null hypothesis that $\rho = 0$ against the alternative that $\rho > 0$. Use $\alpha = 0.025$.

16.33 With reference to the data of Exercise 11.1 on page 398,

(a) calculate the rank correlation coefficient;

(b) test the null hypothesis, at the 0.05 level of significance, that $\rho = 0$ against the alternative that $\rho \neq 0$. Compare your results with those obtained in Exercise 11.44 on page 435.

16.34 Calculate the rank correlation coefficient for the daily rainfall and amount of particulate removed in Exercise 11.13 on page 400.

16.35 With reference to the weights and chest sizes of infants in Exercise 11.47 on page 436,

(a) calculate the rank correlation coefficient;

(b) test the hypothesis, at the 0.025 level of significance, that $\rho = 0$ against the alternative that $\rho > 0$.

16.36 A consumer panel tests nine brands of microwave ovens for overall quality. The ranks assigned by the panel and the suggested retail prices are as follows:

Manufacturer	Panel Rating	Suggested Price
A	6	\$480
B	9	395
C	2	575
D	8	550
E	5	510
F	1	545
G	7	400
H	4	465
I	3	420

Is there a significant relationship between the quality and the price of a microwave oven? Use a 0.05 level of significance.

16.37 Two judges at a college homecoming parade rank eight floats in the following order:

	Float							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Judge *A*	5	8	4	3	6	2	7	1
Judge *B*	7	5	4	2	8	1	6	3

(a) Calculate the rank correlation coefficient.

(b) Test the null hypothesis that $\rho = 0$ against the alternative that $\rho > 0$. Use $\alpha = 0.05$.

16.38 In the article called "Risky Assumptions" by Paul Slovic, Baruch Fischoff, and Sarah Lichtenstein, published in *Psychology Today* (June 1980), the risk of dying in the United States from 30 activities and technologies is ranked by members of the League of Women Voters and also by experts who are professionally involved in assessing risks. The rankings are as shown in Table 16.9.

(a) Calculate the rank correlation coefficient.

(b) Test the null hypothesis of zero correlation between the rankings of the League of Women Voters and the experts against the alternative that the correlation is not zero. Use a 0.05 level of significance.

Table 16.9: The Ranking Data for Exercise 16.38

Activity or Technology Risk	Voters	Experts	Activity or Technology Risk	Voters	Experts
Nuclear power	1	20	Motor vehicles	2	1
Handguns	3	4	Smoking	4	2
Motorcycles	5	6	Alcoholic beverages	6	3
Private aviation	7	12	Police work	8	17
Pesticides	9	8	Surgery	10	5
Fire fighting	11	18	Large construction	12	13
Hunting	13	23	Spray cans	14	26
Mountain climing	15	29	Bicycles	16	15
Commercial aviation	17	16	Electric power	18	9
Swimming	19	10	Contraceptives	20	11
Skiing	21	30	X-rays	22	7
Football	23	27	Railroads	24	19
Food preservatives	25	14	Food coloring	26	21
Power mowers	27	28	Antibiotics	28	24
Home appliances	29	22	Vaccinations	30	25

Review Exercises

16.39 A study by a chemical company compared the drainage properties of two different polymers. Ten different sludges were used, and both polymers were allowed to drain in each sludge. The free drainage was measured in mL/min.

Sludge Type	Polymer *A*	Polymer *B*
1	12.7	12.0
2	14.6	15.0
3	18.6	19.2
4	17.5	17.3
5	11.8	12.2
6	16.9	16.6
7	19.9	20.1
8	17.6	17.6
9	15.6	16.0
10	16.0	16.1

(a) Use the sign test at the 0.05 level to test the null hypothesis that polymer A has the same median drainage as polymer B.

(b) Use the signed-rank test to test the hypotheses of part (a).

16.40 In Review Exercise 13.45 on page 555, use the Kruskal-Wallis test, at the 0.05 level of significance, to determine if the chemical analyses performed by the four laboratories give, on average, the same results.

16.41 Use the data from Exercise 13.14 on page 530 to see if the median amount of nitrogen lost in perspiration is different for the three levels of dietary protein.

Chapter 18

Bayesian Statistics

18.1 Bayesian Concepts

The classical methods of estimation that we have studied in this text are based solely on information provided by the random sample. These methods essentially interpret probabilities as relative frequencies. For example, in arriving at a 95% confidence interval for μ, we interpret the statement

$$P(-1.96 < Z < 1.96) = 0.95$$

to mean that 95% of the time in repeated experiments Z will fall between -1.96 and 1.96. Since

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

for a normal sample with known variance, the probability statement here means that 95% of the random intervals $(\bar{X} - 1.96\sigma/\sqrt{n}, \bar{X} + 1.96\sigma/\sqrt{n})$ contain the true mean μ. Another approach to statistical methods of estimation is called **Bayesian methodology**. The main idea of the method comes from Bayes' rule, described in Section 2.7. The key difference between the Bayesian approach and the classical or frequentist approach is that in Bayesian concepts, the parameters are viewed as random variables.

Subjective Probability

Subjective probability is the foundation of Bayesian concepts. In Chapter 2, we discussed two possible approaches to probability, namely the relative frequency and the indifference approaches. The first one determines a probability as a consequence of repeated experiments. For instance, to decide the free-throw percentage of a basketball player, we can record the number of shots made and the total number of attempts this player has made. The probability of hitting a free-throw for this player can be calculated as the ratio of these two numbers. On the other hand, if we have no knowledge of any bias in a die, the probability that a 3 will appear in the next throw will be 1/6. Such an approach to probability interpretation is based on the indifference rule.

However, in many situations, the preceding probability interpretations cannot be applied. For instance, consider the questions "What is the probability that it will rain tomorrow?" "How likely is it that this stock will go up by the end of the month?" and "What is the likelihood that two companies will be merged together?" They can hardly be interpreted by the aforementioned approaches, and the answers to these questions may be different for different people. Yet these questions are constantly asked in daily life, and the approach used to explain these probabilities is called *subjective probability*, which reflects one's subjective opinion.

Conditional Perspective

Recall that in Chapters 9 through 17, all statistical inferences were based on the fact that the parameters are unknown but fixed quantities, apart from those in Section 9.14, in which the parameters were treated as variables and the maximum likelihood estimates (MLEs) were calculated conditioning on the observed sample data. In Bayesian statistics, not only are the parameters treated as variables as in MLE calculation, but also they are treated as random.

Because the observed data are the only experimental results for the practitioner, statistical inference is based on the actual observed data from a given experiment. Such a view is called a *conditional perspective*. Furthermore, in Bayesian concepts, since the parameters are treated as random, a probability distribution can be specified, generally by using the *subjective probability* for the parameter. Such a distribution is called a *prior distribution* and it usually reflects the experimenter's prior belief about the parameter. In the Bayesian perspective, once an experiment is conducted and data are observed, all knowledge about the parameter is contained in the actual observed data and in the prior information.

Bayesian Applications

Although Bayes' rule is credited to Thomas Bayes, Bayesian applications were first introduced by French scientist Pierre Simon Laplace, who published a paper on using Bayesian inference on the unknown binomial proportions (for binomial distribution, see Section 5.2).

Since the introduction of the Markov chain Monte Carlo (MCMC) computational tools for Bayesian analysis in the early 1990s, Bayesian statistics has become more and more popular in statistical modeling and data analysis. Meanwhile, methodology developments using Bayesian concepts have progressed dramatically, and they are applied in fields such as bioinformatics, biology, business, engineering, environmental and ecology science, life science and health, medicine, and many others.

18.2 Bayesian Inferences

Consider the problem of finding a point estimate of the parameter θ for the population with distribution $f(x|\ \theta)$, given θ. Denote by $\pi(\theta)$ the prior distribution of θ. Suppose that a random sample of size n, denoted by $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, is observed.

Definition 18.1: The distribution of θ, given $\mathbf{x}$, which is called the posterior distribution, is given by

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{g(\mathbf{x})},$$

where $g(\mathbf{x})$ is the marginal distribution of $\mathbf{x}$.

The marginal distribution of $\mathbf{x}$ in the above definition can be calculated using the following formula:

$$g(\mathbf{x}) = \begin{cases} \sum\limits_{\theta} f(\mathbf{x}|\theta)\pi(\theta), & \theta \text{ is discrete,} \\ \int_{-\infty}^{\infty} f(\mathbf{x}|\theta)\pi(\theta)\, d\theta, & \theta \text{ is continuous.} \end{cases}$$

Example 18.1: Assume that the prior distribution for the proportion of defectives produced by a machine is

p	0.1	0.2
$\pi(p)$	0.6	0.4

Denote by x the number of defectives among a random sample of size 2. Find the posterior probability distribution of p, given that x is observed.

***Solution*:** The random variable X follows a binomial distribution

$$f(x|p) = b(x; 2, p) = \binom{2}{x} p^x q^{2-x}, \quad x = 0, 1, 2.$$

The marginal distribution of x can be calculated as

$$\begin{aligned} g(x) &= f(x|0.1)\pi(0.1) + f(x|0.2)\pi(0.2) \\ &= \binom{2}{x}[(0.1)^x(0.9)^{2-x}(0.6) + (0.2)^x(0.8)^{2-x}(0.4)]. \end{aligned}$$

Hence, for $x = 0, 1, 2$, we obtain the marginal probabilities as

x	0	1	2
$g(x)$	0.742	0.236	0.022

The posterior probability of $p = 0.1$, given x, is

$$\pi(0.1|x) = \frac{f(x|0.1)\pi(0.1)}{g(x)} = \frac{(0.1)^x(0.9)^{2-x}(0.6)}{(0.1)^x(0.9)^{2-x}(0.6) + (0.2)^x(0.8)^{2-x}(0.4)},$$

and $\pi(0.2|x) = 1 - \pi(0.1|x)$.

Suppose that $x = 0$ is observed.

$$\pi(0.1|0) = \frac{f(0 \mid 0.1)\pi(0.1)}{g(0)} = \frac{(0.1)^0(0.9)^{2-0}(0.6)}{0.742} = 0.6550,$$

and $\pi(0.2|0) = 0.3450$. If $x = 1$ is observed, $\pi(0.1|1) = 0.4576$, and $\pi(0.2|1) = 0.5424$. Finally, $\pi(0.1|2) = 0.2727$, and $\pi(0.2|2) = 0.7273$. ∎

The prior distribution for Example 18.1 is discrete, although the natural range of p is from 0 to 1. Consider the following example, where we have a prior distribution covering the whole space for p.

Example 18.2: Suppose that the prior distribution of p is uniform (i.e., $\pi(p) = 1$, for $0 < p < 1$). Use the same random variable X as in Example 18.1 to find the posterior distribution of p.

Solution: As in Example 18.1, we have

$$f(x|p) = b(x; 2, p) = \binom{2}{x} p^x q^{2-x}, \quad x = 0, 1, 2.$$

The marginal distribution of x can be calculated as

$$g(x) = \int_0^1 f(x|p)\pi(p)\, dp = \binom{2}{x} \int_0^1 p^x (1-p)^{2-x}\, dp.$$

The integral above can be evaluated at each x directly as $g(0) = 1/3$, $g(1) = 1/3$, and $g(2) = 1/3$. Therefore, the posterior distribution of p, given x, is

$$\pi(p|x) = \frac{\binom{2}{x} p^x (1-p)^{2-x}}{1/3} = 3\binom{2}{x} p^x (1-p)^{2-x}, \quad 0 < p < 1.$$

The posterior distribution above is actually a beta distribution (see Section 6.8) with parameters $\alpha = x + 1$ and $\beta = 3 - x$. So, if $x = 0$ is observed, the posterior distribution of p is a beta distribution with parameters $(1, 3)$. The posterior mean is $\mu = \frac{1}{1+3} = \frac{1}{4}$ and the posterior variance is $\sigma^2 = \frac{(1)(3)}{(1+3)^2(1+3+1)} = \frac{3}{80}$. ∎

Using the posterior distribution, we can estimate the parameter(s) in a population in a straightforward fashion. In computing posterior distributions, it is very helpful if one is familiar with the distributions in Chapters 5 and 6. Note that in Definition 18.1, the *variable* in the posterior distribution is θ, while $\mathbf{x}$ is given. Thus, we can treat $g(\mathbf{x})$ as a constant as we calculate the posterior distribution of θ. Then the posterior distribution can be expressed as

$$\pi(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)\pi(\theta),$$

where the symbol "$\propto$" stands for *is proportional to*. In the calculation of the posterior distribution above, we can leave the factors that do not depend on θ out of the normalization constant, i.e., the marginal density $g(\mathbf{x})$.

Example 18.3: Suppose that random variables $X_1, \ldots, X_n$ are independent and from a Poisson distribution with mean λ. Assume that the prior distribution of λ is exponential with mean 1. Find the posterior distribution of λ when $\bar{x} = 3$ with $n = 10$.

Solution: The density function of $\mathbf{X} = (X_1, \ldots, X_n)$ is

$$f(\mathbf{x}|\lambda) = \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!},$$

and the prior distribution is

$$\pi(\theta) = e^{-\lambda}, \text{ for } \lambda > 0.$$

Hence, using Definition 18.1 we obtain the posterior distribution of λ as

$$\pi(\lambda|\mathbf{x}) \propto f(\mathbf{x}|\lambda)\pi(\lambda) = e^{-n\lambda}\frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}e^{-\lambda} \propto e^{-(n+1)\lambda}\lambda^{\sum_{i=1}^{n} x_i}.$$

Referring to the gamma distribution in Section 6.6, we conclude that the posterior distribution of λ follows a gamma distribution with parameters $1+\sum_{i=1}^{n} x_i$ and $\frac{1}{n+1}$. Hence, we have the posterior mean and variance of λ as $\frac{\sum_{i=1}^{n} x_i+1}{n+1}$ and $\frac{\sum_{i=1}^{n} x_i+1}{(n+1)^2}$. So, when $\bar{x} = 3$ with $n = 10$, we have $\sum_{i=1}^{10} x_i = 30$. Hence, the posterior distribution of λ is a gamma distribution with parameters 31 and 1/11. ∎

From Example 18.3 we observe that sometimes it is quite convenient to use the "proportional to" technique in calculating the posterior distribution, especially when the result can be formed to a commonly used distribution as described in Chapters 5 and 6.

Point Estimation Using the Posterior Distribution

Once the posterior distribution is derived, we can easily use the summary of the posterior distribution to make inferences on the population parameters. For instance, the posterior mean, median, and mode can all be used to estimate the parameter.

Example 18.4: Suppose that $x = 1$ is observed for Example 18.2. Find the posterior mean and the posterior mode.

***Solution*:** When $x = 1$, the posterior distribution of p can be expressed as

$$\pi(p|1) = 6p(1-p), \qquad \text{for} \quad 0 < p < 1.$$

To calculate the mean of this distribution, we need to find

$$\int_0^1 6p^2(1-p)\,dp = 6\left(\frac{1}{3} - \frac{1}{4}\right) = \frac{1}{2}.$$

To find the posterior mode, we need to obtain the value of p such that the posterior distribution is maximized. Taking derivative of $\pi(p)$ with respect to p, we obtain $6 - 12p$. Solving for p in $6 - 12p = 0$, we obtain $p = 1/2$. The second derivative is -12, which implies that the posterior mode is achieved at $p = 1/2$. ∎

Bayesian methods of estimation concerning the mean μ of a normal population are based on the following example.

Example 18.5: If $\bar{x}$ is the mean of a random sample of size n from a normal population with known variance σ^2, and the prior distribution of the population mean is a normal distribution with known mean μ_0 and known variance σ_0^2, then show that the posterior distribution of the population mean is also a normal distribution with

mean μ^* and standard deviation σ^*, where

$$\mu^* = \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2/n}\bar{x} + \frac{\sigma^2/n}{\sigma_0^2 + \sigma^2/n}\mu_0 \qquad \text{and} \qquad \sigma^* = \sqrt{\frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}}.$$

Solution: The density function of our sample is

$$f(x_1, x_2, \ldots, x_n \mid \mu) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2\right],$$

for $-\infty < x_i < \infty$ and $i = 1, 2, \ldots, n$, and the prior is

$$\pi(\mu) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{1}{2}\left(\frac{\mu - \mu_0}{\sigma_0}\right)^2\right], \quad -\infty < \mu < \infty.$$

Then the posterior distribution of μ is

$$\begin{aligned}\pi(\mu|\mathbf{x}) &\propto \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2 + \left(\frac{\mu - \mu_0}{\sigma_0}\right)^2\right]\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left[\frac{n(\bar{x} - \mu)^2}{\sigma^2} + \frac{(\mu - \mu_0)^2}{\sigma_0^2}\right]\right\},\end{aligned}$$

due to

$$\sum_{i=1}^{n}(x_i - \mu)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2$$

from Section 8.5. Completing the squares for μ yields the posterior distribution

$$\pi(\mu|\mathbf{x}) \propto \exp\left[-\frac{1}{2}\left(\frac{\mu - \mu^*}{\sigma^*}\right)^2\right],$$

where

$$\mu^* = \frac{n\bar{x}\sigma_0^2 + \mu_0\sigma^2}{n\sigma_0^2 + \sigma^2}, \qquad \sigma^* = \sqrt{\frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}}.$$

This is a normal distribution with mean μ^* and standard deviation σ^*. ∎

The Central Limit Theorem allows us to use Example 18.5 also when we select sufficiently large random samples ($n \geq 30$ for many engineering experimental cases) from nonnormal populations (the distribution is not very far from symmetric), and when the prior distribution of the mean is approximately normal.

Several comments need to be made about Example 18.5. The posterior mean μ^* can also be written as

$$\mu^* = \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2/n}\bar{x} + \frac{\sigma^2/n}{\sigma_0^2 + \sigma^2/n}\mu_0,$$

which is a weighted average of the sample mean $\bar{x}$ and the prior mean μ_0. Since both coefficients are between 0 and 1 and they sum to 1, the posterior mean μ^* is always

between $\bar{x}$ and μ_0. This means that the posterior estimation of μ is influenced by both $\bar{x}$ and μ_0. Furthermore, the weight of $\bar{x}$ depends on the prior variance as well as the variance of the sample mean. For a large sample problem ($n \to \infty$), the posterior mean $\mu^* \to \bar{x}$. This means that the prior mean does not play any role in estimating the population mean μ using the posterior distribution. This is very reasonable since it indicates that when the amount of data is substantial, information from the data will dominate the information on μ provided by the prior. On the other hand, when the prior variance is large ($\sigma_0^2 \to \infty$), the posterior mean μ^* also goes to $\bar{x}$. Note that for a normal distribution, the larger the variance, the flatter the density function. The flatness of the normal distribution in this case means that there is almost no subjective prior information available on the parameter μ before the data are collected. Thus, it is reasonable that the posterior estimation μ^* only depends on the data value $\bar{x}$.

Now consider the posterior standard deviation σ^*. This value can also be written as

$$\sigma^* = \sqrt{\frac{\sigma_0^2 \sigma^2/n}{\sigma_0^2 + \sigma^2/n}}.$$

It is obvious that the value σ^* is smaller than both σ_0 and $\sigma/\sqrt{n}$, the prior standard deviation and the standard deviation of $\bar{x}$, respectively. This suggests that the posterior estimation is more accurate than both the prior and the sample data. Hence, incorporating both the data and prior information results in better posterior information than using any of the data or prior alone. This is a common phenomenon in Bayesian inference. Furthermore, to compute μ^* and σ^* by the formulas in Example 18.5, we have assumed that σ^2 is known. Since this is generally not the case, we shall replace σ^2 by the sample variance s^2 whenever $n \geq 30$.

Bayesian Interval Estimation

Similar to the classical confidence interval, in Bayesian analysis we can calculate a $100(1-\alpha)\%$ Bayesian interval using the posterior distribution.

Definition 18.2:

The interval $a < \theta < b$ will be called a $100(1-\alpha)\%$ **Bayesian interval** for θ if

$$\int_{-\infty}^{a} \pi(\theta|x)\, d\theta = \int_{b}^{\infty} \pi(\theta|x)\, d\theta = \frac{\alpha}{2}.$$

Recall that under the frequentist approach, the probability of a confidence interval, say 95%, is interpreted as a coverage probability, which means that if an experiment is repeated again and again (with considerable unobserved data), the probability that the intervals calculated according to the rule will cover the true parameter is 95%. However, in Bayesian interval interpretation, say for a 95% interval, we can state that the probability of the unknown parameter falling into the calculated interval (which only depends on the observed data) is 95%.

Example 18.6: Supposing that $X \sim b(x; n, p)$, with known $n = 2$, and the prior distribution of p is uniform $\pi(p) = 1$, for $0 < p < 1$, find a 95% Bayesian interval for p.

***Solution*:** As in Example 18.2, when $x = 0$, the posterior distribution is a beta distribution with parameters 1 and 3, i.e., $\pi(p|0) = 3(1-p)^2$, for $0 < p < 1$. Thus, we need to solve for a and b using Definition 18.2, which yields the following:

$$0.025 = \int_0^a 3(1-p)^2 \, dp = 1 - (1-a)^3$$

and

$$0.025 = \int_b^1 3(1-p)^2 \, dp = (1-b)^3.$$

The solutions to the above equations result in $a = 0.0084$ and $b = 0.7076$. Therefore, the probability that p falls into $(0.0084, 0.7076)$ is 95%.

For the normal population and normal prior case described in Example 18.5, the posterior mean μ^* is the Bayes estimate of the population mean μ, and a $100(1-\alpha)\%$ **Bayesian interval** for μ can be constructed by computing the interval

$$\mu^* - z_{\alpha/2}\sigma^* < \mu < \mu^* + z_{\alpha/2}\sigma^*,$$

which is centered at the posterior mean and contains $100(1-\alpha)\%$ of the posterior probability.

Example 18.7: An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed with a standard deviation of 100 hours. Prior experience leads us to believe that μ is a value of a normal random variable with a mean $\mu_0 = 800$ hours and a standard deviation $\sigma_0 = 10$ hours. If a random sample of 25 bulbs has an average life of 780 hours, find a 95% Bayesian interval for μ.

***Solution*:** According to Example 18.5, the posterior distribution of the mean is also a normal distribution with mean

$$\mu^* = \frac{(25)(780)(10)^2 + (800)(100)^2}{(25)(10)^2 + (100)^2} = 796$$

and standard deviation

$$\sigma^* = \sqrt{\frac{(10)^2(100)^2}{(25)(10)^2 + (100)^2}} = \sqrt{80}.$$

The 95% Bayesian interval for μ is then given by

$$796 - 1.96\sqrt{80} < \mu < 796 + 1.96\sqrt{80},$$

or

$$778.5 < \mu < 813.5.$$

Hence, we are 95% sure that μ will be between 778.5 and 813.5.

On the other hand, ignoring the prior information about μ, we could proceed as in Section 9.4 and construct the classical 95% confidence interval

$$780 - (1.96)\left(\frac{100}{\sqrt{25}}\right) < \mu < 780 + (1.96)\left(\frac{100}{\sqrt{25}}\right),$$

or $740.8 < \mu < 819.2$, which is seen to be wider than the corresponding Bayesian interval.