Chapter 9

Regression Problems

INTRODUCTION

Among the most common applications of statistical techniques are those involving some sort of regression analysis. Such procedures are designed to detect and interpret stochastic relationships between a dependent (response) variable and one or more independent (predictor) variables. These regression relationships can vary from that of a simple linear relationship between the dependent variable and a single independent variable to complex, nonlinear relationships involving a large number of predictor variables.

In Sections 9.1–9.4, we present nonparametric procedures designed for the simplest of regression relationships, namely, that of a single stochastic linear relationship between a dependent variable and one independent variable. (Such a relationship is commonly referred to as a *regression line*.) In Section 9.1, we present a distribution-free test of the hypothesis that the slope of the regression line is a specified value. Sections 9.2 and 9.3 provide, respectively, a point estimator and distribution-free confidence intervals and bounds for the slope parameter. In Section 9.4, we complete the analysis for a single regression line by discussing both an estimator of the intercept of the line and the use of the estimated linear relationship to provide predictions of dependent variable responses to additional values of the predictor variable. In Section 9.5, we consider the case of two or more regression lines and describe an asymptotically distribution-free test of the hypothesis that the regression lines have the same slope; that is, that the regression lines are parallel.

In Section 9.6, we present the reader with an introduction to the extensive field of rank-based regression analysis for more complicated regression relationships than that of a straight line. In Section 9.7, we provide short introductions to a number of recent developments in the rapidly expanding area of non-rank-based nonparametric regression, where the goal is to make statistical inferences about the relationship between a dependent variable and one or more independent variables without a priori specification of a formal model describing the regression relationship. These non-rank-based approaches to nonparametric regression are generally more complicated than the level assumed throughout the rest of this text. As a result, our approach in Section 9.7 is simply to give brief descriptions of a variety of statistical techniques that are commonly used to develop such procedures and provide appropriate references for readers interested in more detailed information about them, rather than to concentrate on specific procedures and their application to appropriate data sets.

Nonparametric Statistical Methods, Third Edition. Myles Hollander, Douglas A. Wolfe, Eric Chicken.

Finally, in Section 9.8, we consider the asymptotic relative efficiencies of the straight-line regression procedures discussed in Sections 9.1–9.3 and 9.5–9.6 with respect to their competitors based on classical least squares estimators.

ONE REGRESSION LINE

Data. At each of n fixed values, $x_1, \ldots, x_n$, of the independent (predictor) variable x, we observe the value of the response random variable Y. Thus, we obtain a set of observations $Y_1, \ldots, Y_n$, where Y_i is the value of the response variable when $x = x_i$. The x's are assumed to be distinct and, without loss of generality, we take $x_1 < x_2 < \cdots < x_n$.

Assumptions

A1. Our straight-line model is

$$Y_i = \alpha + \beta x_i + e_i, \quad i = 1, \ldots, n, \tag{9.1}$$

where the x's are known constants and α (the intercept) and β (the slope) are unknown parameters.

A2. The random variables $e_1, \ldots, e_n$ are a random sample from a continuous population that has median 0.

9.1 A DISTRIBUTION-FREE TEST FOR THE SLOPE OF THE REGRESSION LINE (THEIL)

Hypothesis

The null hypothesis of interest here is that the slope, β, of the postulated regression line is some specified value β_0, namely,

$$H_0 : \beta = \beta_0. \tag{9.2}$$

Thus, the null hypothesis asserts that for every unit increase in the value of the independent (predictor) variable x, we would expect an increase (or decrease, depending on the sign of β_0) of roughly β_0 in the value of the dependent (response) variable Y.

Procedure

To compute the Theil (1950a) statistic C, we first form the n differences

$$D_i = Y_i - \beta_0 x_i, \quad i = 1, \ldots, n. \tag{9.3}$$

Let

$$C = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} c(D_j - D_i), \tag{9.4}$$

where

$$c(a) = \begin{cases} -1, & \text{if } a < 0, \\ 0, & \text{if } a = 0, \\ 1, & \text{if } a > 0. \end{cases} \tag{9.5}$$

Thus, for each pair of subscripts (i, j), with $1 \leq i < j \leq n$, score 1 if $D_j - D_i$ is positive, and score -1 if $D_j - D_i$ is negative. The statistic C (9.4) is then just the sum of these 1's and -1s.

a. *One-Sided Upper-Tail Test*. To test the null hypothesis

$$H_0 : \beta = \beta_0$$

versus the alternative that the slope is larger than the specified β_0 corresponding to

$$H_1 : \beta > \beta_0, \tag{9.6}$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } \bar{C} \geq k_\alpha; \quad \text{otherwise do not reject,} \tag{9.7}$$

where the constant k_α is chosen to make the type I error probability equal to α and $\bar{C} = C/(n(n-1)/2)$. (See Comment 2.)

b. *One-Sided Lower-Tail Test*. To test the null hypothesis

$$H_0 : \beta = \beta_0$$

versus the alternative that the slope is smaller than the specified β_0 corresponding to

$$H_2 : \beta < \beta_0, \tag{9.8}$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } \bar{C} \leq -k_\alpha; \quad \text{otherwise do not reject.} \tag{9.9}$$

c. *Two-Sided Test*. To test the null hypothesis

$$H_0 : \beta = \beta_0$$

versus the alternative that the slope is simply not equal to the specified β_0 corresponding to

$$H_3 : \beta \neq \beta_0, \tag{9.10}$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } |\bar{C}| \geq k_{\alpha/2}; \quad \text{otherwise do not reject.} \tag{9.11}$$

This two-sided procedure is the two-sided symmetric test with $\alpha/2$ probability in each tail of the null distribution of $\bar{C}$.

Large-Sample Approximation

The large-sample approximation is based on the asymptotic normality of C, suitably standardized. For this standardization, we need to know the expected value and variance of C when the null hypothesis H_0 (9.2) is true. Under H_0, the expected value and variance of C are

$$E_0(C) = 0 \tag{9.12}$$

and

$$\text{var}_0(C) = \frac{n(n-1)(2n+5)}{18}, \tag{9.13}$$

respectively. (See Comment 2.)

The standardized version of C is

$$C^* = \frac{C - E_0(C)}{\{\text{var}_0(C)\}^{1/2}} = \frac{C}{\{n(n-1)(2n+5)/18\}^{1/2}}. \tag{9.14}$$

When H_0 is true, C^* has, as n tends to infinity, an asymptotic $N(0,1)$ distribution (See Comment 2). The normal theory approximation for procedure (9.7) is

$$\text{Reject } H_0 \text{ if } C^* \geq z_\alpha; \quad \text{otherwise do not reject,} \tag{9.15}$$

the normal theory approximation for procedure (9.9) is

$$\text{Reject } H_0 \text{ if } C^* \leq -z_\alpha; \quad \text{otherwise do not reject,} \tag{9.16}$$

and the normal theory approximation for procedure (9.11) is

$$\text{Reject } H_0 \text{ if } |C^*| \geq z_{\alpha/2}; \quad \text{otherwise do not reject.} \tag{9.17}$$

Ties

If there are ties among the D_i (9.3) differences, C may be computed as described in (9.4), but keep in mind that procedures (9.7), (9.9), and (9.11) are then approximate rather than exact. Sen (1968) suggested a way to deal with ties among the values of the independent variable x.

EXAMPLE 9.1 ***Effect of Cloud Seeding on Rainfall.***

Smith (1967) described experiments performed in Australia to investigate the effects of a particular method of cloud seeding on the amount of rainfall. In one experiment that took place in the Snowy Mountains, two areas served as target and control, respectively, and during any one period, a random process was used to determine whether clouds over the target area should be seeded. The effect of seeding was measured by the *double ratio* [T/Q (seeded)]/[T/Q (unseeded)], where T and Q are the total rainfalls in the target and control areas, respectively. Table 9.1 provides the double ratio calculated for each year of a 5-year experiment.

The slope parameter β represents the rate of change in Y per unit change in x. We apply the one-sided lower-tail test (9.9) with β_0 equal zero. This should be viewed as a

Table 9.1 Double Ratio for 5 Years in the Snowy Mountains of Australia

Years seeded, x_i	Double ratio, Y_i
1	1.26
2	1.27
3	1.12
4	1.16
5	1.03

Source: E. J. Smith (1967).

test of the null hypothesis that the double ratio does not change with time (i.e., the effects of seeding during one year do not overlap into other years) against the alternative that there is a decrease over time, either in the rainfall increases resulting from the seeding or in the ability of the experiments to detect such increases.

From (9.3), with $\beta_0 = 0$, we see that $D_i = Y_i$. We now illustrate the computations required to obtain the value of C (9.4) for these data.

(i,j)	$D_j - D_i$	$c(D_j - D_i)$
(1, 2)	.01	1
(1, 3)	−.14	−1
(1, 4)	−.10	−1
(1, 5)	−.23	−1
(2, 3)	−.15	−1
(2, 4)	−.11	−1
(2, 5)	−.24	−1
(3, 4)	.04	1
(3, 5)	−.09	−1
(4, 5)	−.13	−1

Thus, we find the value of C and $\bar{C}$ to be

$$C = \sum_{i=1}^{4} \sum_{j=i+1}^{5} c(D_j - D_i) = -6, \quad \bar{C} = \frac{C}{n(n-1)/2} = -.6.$$

Using the fact that the null distribution of C is symmetric about zero (See Comment 2), we find that the P-value for these data is $P(\bar{C} \leq -.6) =$ `pKendall(-.6, N = 5, lower.tail = T)` = .117. Thus, there is not much evidence of a decrease over time of the rainfall increases resulting from the seeding.

To illustrate the normal theory approximation (which should not be expected to be highly accurate for a sample size as small as 5), we first find from (9.14) that

$$C^* = \frac{-6}{\{5(4)(15)/18\}^{1/2}} = -1.47.$$

Thus, the smallest significance level at which we can reject $H_0 : \beta = 0$ in favor of $\beta < 0$ using the normal theory approximation is .0708, since $z_{.0708} = -1.47$. As expected for this small sample size ($n = 5$), this is not in especially good agreement with the exact P-value of .117 found previously.

The above analysis may also be carried out in R. The command `theil` requires arguments for data vectors x and y, the null hypothesized value `beta. 0` and `type="t"`, `"1"`, or `"u"` for a two-tailed, lower-tail, or upper-tail test, respectively. If taking `x` and `y` to be the data from Table 9.1, a call of the function `theil(x, y, beta.0=0, type="1")` results in the following output, which reproduces the analysis above:

```
Null:  beta less than 0
C = -6, C.bar = -0.6, P = 0.117.
```

Comments

1. *Motivation for the Test.* From (9.4), we see that C will be large when $D_j > D_i$ for many (i,j) pairs. Now

$$D_j - D_i = [Y_j - \beta_0 x_j - (Y_i - \beta_0 x_i)] = [Y_j - Y_i + \beta_0(x_i - x_j)].$$

   Furthermore, under model (9.1), the median of $Y_j - Y_i = [\beta(x_j - x_i) + (e_j - e_i)]$ is $\beta(x_j - x_i)$. Thus, under model (9.1), the median of $D_j - D_i$ is $[\beta(x_j - x_i) + \beta_0(x_i - x_j)] = (\beta - \beta_0)(x_j - x_i)$. Hence, we tend to obtain positive $D_j - D_i$ differences when $\beta > \beta_0$, and these positive differences lead to large values of C. This serves as partial motivation for procedure (9.7).
2. *Relationship to Kendall's Correlation Statistic K.* The statistic C (9.4) is simply Kendall's correlation statistic K (8.6) computed between the x and $Y - \beta_0 x$ values. In particular, a test of $\beta_0 = 0$ can be interpreted as a test for correlation between the x and Y sequences. Moreover, the null H_0 (9.2) distribution properties of the statistic C (when there are no tied D values) are identical with the corresponding distributional properties of Kendall's statistic K under its null hypothesis of independence (See Section 8.1). This leads immediately to the use of the critical values k_α in procedures (9.7), (9.9), and (9.11). In addition, the symmetry about zero for the null distribution of C follows from Comment 8.8 and the values of $E_0(C)$ and $\text{var}_0(C)$ are direct consequences of the corresponding values of $E_0(K)$ and $\text{var}_0(K)$, respectively, developed in Comment 8.10. Finally, the asymptotic ($n \to \infty$) normality for the standardized statistic C^* under the null hypothesis H_0 (9.2) derives from the similar property for the standardized K^*, as discussed in Comment 8.10.
3. *Testing for Trends over Time.* In the special case when the x-values are the time order (as in Example 9.1), the procedures in (9.7), (9.9), and (9.11) (with β_0 set equal to zero) can be viewed as tests against a time trend and have been suggested for this use by Mann (1945). (See also Comment 8.14.)

Properties

1. *Consistency.* The tests defined by (9.7), (9.9), and (9.11) are consistent against the alternatives 0 $\beta >, <$, and $\neq \beta_0$, respectively.
2. *Asymptotic Normality.* See Comment 2.
3. *Efficiency.* See Sen (1968) and Section 9.8.

Problems

1. Johnson et al. (1970) considered the behavior of a cenosphere-resin composite under hydrostatic pressure. The authors pointed out that most deep submersible vehicles utilize a buoyancy material, known as *syntactic foam*, that is a composite of closely packed hollow glass microspheres embedded in a resin matrix. These microspheres are relatively expensive to manufacture, and the cost of the syntactic foam is principally determined by the cost of the microspheres. The authors also noted that the ash from generating stations burning pulverized coal contains a small proportion of hollow glassy microspheres, known as *cenospheres*, and these have about the right size distribution for use in syntactic foam. The cenospheres can be readily collected from the ash-disposal method used in certain British generating stations. The authors were thus interested in whether the cenospheres would, in some applications, perform as well as the manufactured microspheres.

In attempting to assess the usefulness of cenospheres as a component of syntactic foam, Johnson et al. investigated the effects of hydrostatic pressure (such as exists in the ocean depths) on the density of a cenosphere-resin composite. The results are given in Table 9.2. What is the P-value for a test of $H_0 : \beta = 0$ against the alternative $\beta > 0$ for these data?

2. Explain why the effect of the unknown intercept parameter α (See model (9.1)) is "eliminated" in the application of procedure (9.4) to a set of data.

3. Consider the tapeworm data discussed in Problem 8.1. Using the mean weight of the initial force-fed cysticerci as the independent (predictor) variable, test the hypothesis that there was virtually no change in the mean weight of the cysticerci over the 20-day period following introduction into the dogs against the alternative that the typical tapeworm grew in size during the period of the study.

4. Stitt, Hardy, and Nadel (1971) studied the relationship between the surface area and body weight of squirrel monkeys. The data in Table 9.3 represent the total surface areas (cm^3) and

Table 9.2 The Effects of Hydrostatic Pressure on the Density of a Cenosphere-Resin Composite

Specimen	Pressure (psi)	Density (g/cm^3)
1	0	0.924
2	5,000	0.988
3	10,000	0.992
4	15,000	1.118
5	20,000	1.133
6	25,000	1.145
7	30,000	1.157
8	100,000	1.357

Source: A. A. Johnson, K. Mukherjee, S. Schlosser, and E. Raask (1970).

Table 9.3 Body Weight and Total Surface Area of Squirrel Monkeys

Monkey	Body weight, g	Total surface area, cm^3
1	660	780.6
2	705	887.6
3	994	1122.8
4	1129	1125.2
5	1005	1070.4
6	923	1039.2
7	953	1040.0
8	1018	1133.4
9	1181	1148.0

Source: J. T. Stitt, J. D. Hardy, and E. R. Nadel (1971).

body weights (g) for nine squirrel monkeys. Treating body weight as the independent variable, test for the presence of a linear relationship between these two measurements in squirrel monkeys.

5. Explain the meaning of the intercept parameter α and slope parameter β in model (9.1).
6. Consider the odor periods data of Table 8.8 in Problem 8.19. Test the conjecture that over the period 1950–1964, the number of odor periods for Lake Michigan generally increased at a rate greater than two per year.

9.2 A SLOPE ESTIMATOR ASSOCIATED WITH THE THEIL STATISTIC (THEIL)

Procedure

To estimate the slope parameter β of model (9.1), compute the $N = n(n-1)/2$ individual sample slope values $S_{ij} = (Y_j - Y_i)/(x_j - x_i), 1 \le i < j \le n$. The estimator of β (Theil (1950c)) associated with the Theil statistic, C, is

$$\hat{\beta} = \text{median } \{S_{ij}, 1 \le i < j \le n\}. \tag{9.18}$$

Let $S^{(1)} \le \cdots \le S^{(N)}$ denote the ordered values of the sample slopes $S_{i,j}$. Then if N is odd, say $N = 2k + 1$, we have $k = (N-1)/2$ and

$$\hat{\beta} = S^{(k+1)}, \tag{9.19}$$

the value that occupies position $k + 1$ in the list of the ordered S_{ij} values. If N is even, say $N = 2k$, then $k = N/2$ and

$$\hat{\beta} = [S^{(k)} + S^{(k+1)}]/2. \tag{9.20}$$

That is, when N is even, $\hat{\beta}$ is the average of the two S_{ij} values that occupy positions k and $k + 1$ in the ordered list of all N sample slopes S_{ij}.

EXAMPLE 9.2 ***Effect of Cloud Seeding on Rainfall—Example 9.1 Continued.***

Consider the double-ratio data of Table 9.1. The ordered values of the $N = 5(4)/2 = 10$ sample slopes $S_{ij} = (Y_j - Y_i)/(x_j - x_i)$ are $S^{(1)} \le \cdots \le S^{(10)} : -.150, -.130, -.080, -.070, -.0575, -.055, -.045, -.033, .010,$ and $.040$. As $N = 10$ is even, we use (9.20) with $k = \frac{10}{2} = 5$ to obtain the slope estimate $\hat{\beta} = [S^{(5)} + S^{(6)}]/2 = [-.0575 - .055]/2 = -.0563$.

Comments

4. *Generalization for Nondistinct x-Values.* Sen (1968) generalized Theil's (1950c) estimator to the case where the x's are not distinct. Let N' denote the number of positive $x_j - x_i$ differences, for $1 \le i < j \le n$. (In the case where the x's are distinct, $N' = N$.) Sen's estimator of β is the median of the N' sample slope values that can be computed from the data. In the special case when

$x_1 = x_2 = \ldots = x_m = 0$ and $x_{m+1} = x_{m+2} = \ldots = x_{m+q} = 1$ (with $n = m + q$ and $m < n$), Sen's estimator reduces to the median of the $mq(Y_j - Y_i)$ differences, where $i = 1, \ldots, m$ and $j = m + 1, \ldots, m + q$. That is, Sen's estimator reduces to the Hodges–Lehmann two-sample estimator of Section 4.2 applied to the two samples $Y_1, \ldots, Y_m$ and $Y_{m+1}, \ldots, Y_{m+q}$.

Dietz (1989) considered various nonparametric estimators of the slope including Theil's estimator. She found that Theil's estimator is robust, easy to compute, and competitive in terms of mean squared error with alternative slope estimators. She also considered various nonparametric estimators of the intercept and of the mean response at a given x-value.

5. *Sensitivity to Gross Errors.* The estimator $\hat{\beta}$ (9.18) is less sensitive to gross errors than is the classical least squares estimator

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

where $\bar{x} = \sum_{i=1}^{n} x_i/n$ and $\bar{Y} = \sum_{j=1}^{n} Y_j/n$.

6. *Median versus Weighted Average.* The estimator $\hat{\beta}$ (9.18) is the median of the N individual slope estimators $S_{ij} = (Y_j - Y_i)/(x_j - x_i)$. The least squares estimator $\bar{\beta}$ (See Comment 5) is a weighted average of the S_{ij}'s.

7. *Sample Slopes*. The command `theil` will output the $n(n-1)/2$ sample slopes if the additional argument `slopes=T` is specified. `If x` and `y` are the data from Table 9.1, then `theil(x, y, slopes=T)` will output the following table:

i	j	S_ij
1	2	0.01000000
1	3	−0.07000000
1	4	−0.03333333
1	5	−0.05750000
2	3	−0.15000000
2	4	−0.05500000
2	5	−0.08000000
3	4	0.04000000
3	5	−0.04500000
4	5	−0.13000000

A different slope estimator is due to Siegel (1982). For a fixed point, the Siegel estimator computes the $n - 1$ slopes with the remaining points and takes the median of these $n - 1$ values. This is done for each point, resulting in n medians. The median of these n medians is the estimate of β.

Properties

1. *Standard Deviation of $\hat{\beta}$ (9.18).* For the asymptotic standard deviation of $\hat{\beta}$ (9.18), see Sen (1968).

2. *Asymptotic Normality.* See Sen (1968).
3. *Efficiency.* See Sen (1968) and Section 9.8.

Problems

7. Estimate β for the cenosphere-resin data of Table 9.2.

8. Compute the least squares estimator $\bar{\beta}$ (See Comment 5) for the cenosphere-resin data of Table 9.2, and compare $\bar{\beta}$ with the $\hat{\beta}$ value obtained in Problem 7. In general, which of $\hat{\beta}$ and $\bar{\beta}$ is easier to compute?

9. Estimate β for the body-weight and surface-area data for squirrel monkeys discussed in Problem 4.

10. Obtain the set of 28 ordered individual sample slopes for the cenosphere-resin data of Table 9.2.

11. Estimate β for the tapeworm data discussed in Problems 3 and 8.1.

12. Obtain the set of 45 ordered individual sample slopes for the tapeworm data discussed in Problems 3 and 8.1.

9.3 A DISTRIBUTION-FREE CONFIDENCE INTERVAL ASSOCIATED WITH THE THEIL TEST (THEIL)

Procedure

For a symmetric two-sided confidence interval for β, with confidence coefficient $1-\alpha$, first obtain the upper $(\alpha/2)$th percentile point $k_{\alpha/2}$ of the null distribution of $\overline{C}$ (9.4). Let $C_\alpha = n(n-1)/2 \cdot k_{\alpha/2} - 2$ and set

$$M = \frac{N - C_\alpha}{2}, \tag{9.21}$$

and

$$Q = \frac{N + C_\alpha}{2} = M + C_\alpha, \tag{9.22}$$

where, once again, $N = n(n-1)/2$. The $100(1-\alpha)\%$ confidence interval (β_L, β_U) for the slope β that is associated with the two-sided Theil test (Section 9.1) is then given by

$$\beta_L = S^{(M)}, \quad \beta_U = S^{(Q+1)}, \tag{9.23}$$

where $S^{(1)} \le \cdots \le S^{(N)}$ are the ordered individual sample slopes $S_{ij} = (Y_j - Y_i)/(x_j - x_i), 1 \le i < j \le n$, used in computing the point estimator $\hat{\beta}$ (9.18). That is, β_L is the sample slope S_{ij} that occupies position M in the list of N ordered sample slopes. The upper end point β_U is the sample slope S_{ij} value that occupies position $Q = M + C_\alpha$ in this ordered list. With β_L and β_U given by display (9.23), we have

$$P_\beta(\beta_L < \beta < \beta_U) = 1 - \alpha \text{ for all } \beta. \tag{9.24}$$

For upper or lower confidence bounds for β associated with appropriate one-sided Theil's test procedures, see Comment 9.

Large-Sample Approximation

For large n, the integer C_α may be approximated by

$$C_\alpha \approx z_{\alpha/2}\left\{\frac{n(n-1)(2n+5)}{18}\right\}^{1/2}. \tag{9.25}$$

In general, the value of the right-hand side of (9.25) is not an integer. To be conservative, take C_α to be the largest integer that is less than or equal to the right-hand side of (9.25) for use in (9.21) and (9.22).

EXAMPLE 9.3 ***Effect of Cloud Seeding on Rainfall—Example 9.1 Continued.***

Consider the double-ratio data of Table 9.1. We illustrate how to obtain the 95% confidence interval for β. With $1-\alpha = .95$ (so that $\alpha = .05$), we see that $k_{\alpha/2} = k_{.025} = .8$. Thus, $C_{.025} = \frac{5.4}{2}k_{.025} - 2 = 8 - 2 = 6$. Since $N = 5(4)/2 = 10$, we see from (9.21) and (9.22) that

$$M = \frac{10-6}{2} = 2$$

and

$$Q = \frac{10+6}{2} = 8.$$

Using these values of $M = 2$ and $Q = 8$ in display (9.23), we see that

$$\beta_L = S^{(2)}, \quad \beta_U = S^{(9)}$$

provide the end points of our 95% confidence interval for β. From the ordered list of sample slope values given in Example 9.2, we obtain $S^{(2)} = -.130$ and $S^{(9)} = .010$, so that our 95% confidence interval for β is

$$(\beta_L, \beta_U) = (-.130, .010).$$

Note that the critical value given by `R` results in a confidence level of $1-\alpha = .917$, not $1-\alpha = .95$. Using the `theil` command with the arguments `alpha=1-.05` and `type="t"` results in the following output:

```
1 - alpha = 0.05 two-sided CI for beta:
-0.15, 0.04
```

The theil command produces an interval whose confidence level is at least $1-\alpha$. The actual confidence level for this interval is $1-\alpha = .983$, compared to $1-\alpha = .917$ for the interval determined by hand.

Comments

8. *Use of* `R` *to Compute the End Points of the Confidence Interval (9.3).* The $n(n-1)/2$ individual sample slope values $S_{ij} = (Y_j - Y_i)/(x_j - x_i), 1 \le i \le$

$j \le n$, can also be obtained from the `theil` command by specifying the argument `slopes=T`. For details, See Comment 7.

9. *Confidence Bounds*. In many settings, we are interested only in making one-sided confidence statements about the parameter β; that is, we wish to assert with specified confidence that β is no larger (or, in other settings, no smaller) than some upper (lower) confidence bound based on the sample data. To obtain such one-sided confidence bounds for β, we proceed as follows. For specified confidence coefficient $1-\alpha$, find the upper αth [not $(\alpha/2)$th, as for the confidence interval] percentile point k_a of the null distribution of C (9.4). Let $C_\alpha^* = \frac{n(n-1)}{2}k_\alpha - 2$ and set

$$M^* = \frac{N - C_\alpha^*}{2} \quad \text{and} \quad Q^* = \frac{N + C_\alpha^*}{2}. \tag{9.26}$$

The $100(1-\alpha)\%$ *lower* confidence bound β_{L}^* for β is then given by

$$(\beta_{\mathrm{L}}^*, \infty) = (S^{(M^*)}, \infty), \tag{9.27}$$

where, as before, $S^{(1)} \le \cdots \le S^{(N)}$ are the ordered individual sample slopes. With β_{L}^* given by display (9.27), we have

$$P_\beta(\beta_{\mathrm{L}}^* < \beta < \infty) = 1 - \alpha \text{ for all } \beta. \tag{9.28}$$

The corresponding $100(1-\alpha)\%$ *upper* confidence bound β_{U}^* is given by

$$(-\infty, \beta_{\mathrm{U}}^*) = (-\infty, S^{(Q^*+1)}). \tag{9.29}$$

It follows that

$$P_\beta(-\infty < \beta < \beta_{\mathrm{U}}^*) = 1 - \alpha \text{ for all } \beta. \tag{9.30}$$

For large n, the integer C_α^* may be approximated by

$$C_\alpha^* \approx z_\alpha \left\{ \frac{n(n-1)(2n+5)}{18} \right\}^{1/2}. \tag{9.31}$$

In general, the value of the right-hand side of (9.31) is not an integer. To be conservative, take C_α^* to be the largest integer that is less than or equal to the right-hand side of (9.31) for use in display (9.26).

10. *Midpoint of the Confidence Interval as an Estimator.* The midpoint of the confidence interval given by (9.23), namely, $[S^{[M)} + S^{(Q+1)}]/2$, suggests itself as a reasonable estimator of β. (Note that this actually yields a class of estimators depending on the value of α.) In general, this midpoint does not give the same value as $\hat{\beta}$ (9.18).

Properties

1. *Distribution-Freeness.* For populations satisfying Assumptions A1 and A2, (9.24) holds (See Theil (1950b, 1950c)). Hence, we can control the coverage probability to be $1-\alpha$ without having more specific knowledge about the form of the common underlying distribution of the e_i's. Thus, (β_L, β_U) is a distribution-free confidence interval for β over a very large class of populations.
2. *Efficiency.* See Sen (1968) and Section 9.8.

Problems

13. Obtain a 90% confidence interval for β for the cenosphere-resin data in Table 9.2.

14. Obtain a 95% confidence interval for β for the body weight and surface area data for squirrel monkeys discussed in Problems 4 and 9.

15. Consider a fixed set of data. Show that for $\alpha_2 > \alpha_1$, the symmetric two-sided $(1-\alpha_1)$ confidence interval for β given by (9.23) is always as long or longer than the symmetric two-sided $(1-\alpha_2)$ confidence interval for β.

16. Obtain a 95% upper confidence bound (See Comment 9) for β for the double-ratio cloud-seeding data in Table 9.1.

17. Obtain a 95% lower confidence bound (See Comment 9) for β for the tapeworm data discussed in Problems 8.1, 3, and 11.

18. Obtain a 90% confidence interval for β for the Lake Michigan odor periods data discussed in Problems 8.19 and 6.

19. Find the midpoint of the 95% confidence interval for β obtained in Problem 14 for the squirrel monkey body-weight and surface-area data. As noted in Comment 10, this midpoint can be used to estimate the value of β. Compare this midpoint estimator with the value of $\hat{\beta}$ (9.18) obtained in Problem 9.

9.4 AN INTERCEPT ESTIMATOR ASSOCIATED WITH THE THEIL STATISTIC AND USE OF THE ESTIMATED LINEAR RELATIONSHIP FOR PREDICTION (HETTMANSPERGER–McKEAN–SHEATHER)

Procedure

To estimate the intercept parameter α of model (9.1), we define

$$A_i = Y_i - \hat{\beta}x_i, \quad i = 1, \ldots, n, \tag{9.32}$$

where $\hat{\beta}$ is the point estimator of β given in (9.18). An estimator associated with the Theil statistic C and suggested by Hettmansperger, McKean, and Sheather (1997) is

$$\hat{\alpha} = \text{median } \{A_1, \ldots, A_n\}. \tag{9.33}$$

Let $A^{(1)} \leq \cdots \leq A^{(n)}$ denote the ordered A_i values (9.32). Then if n is odd, say $n = 2k+1$, we have $k = (n-1)/2$ and

$$\hat{\alpha} = A^{(k+1)}, \tag{9.34}$$

the value that occupies position $k+1$ in the list of ordered A_i values. If n is even, say $n = 2k$, then $k = n/2$ and

$$\hat{\alpha} = \frac{A^{(k)} + A^{(k+1)}}{2}. \tag{9.35}$$

That is, when n is even, $\hat{\alpha}$ is the average of the two values that occupy positions k and $k+1$ in the ordered list of all n A_i's.

Employing both the estimator $\hat{\beta}$ (9.18) for the slope and the estimator $\hat{\alpha}$ (9.33) for the intercept, our estimated linear relationship between the x and Y variables is then given by

$$\overbrace{\text{med } Y_{x=x^*}} = \overbrace{[\text{median } Y \text{ when } x = x^*]} = \hat{\alpha} + \hat{\beta}\, x^*. \tag{9.36}$$

That is, we would predict $\overbrace{\text{med } Y_{x=x^*}}$ to be the typical value of the dependent variable Y for a future setting of the independent variable x at x^*. (See also Comment 12.)

EXAMPLE 9.4 ***Effect of Cloud Seeding on Rainfall—Example 9.1 Continued.***

Once again, consider the double-ratio data of Table 9.1. From Example 9.2, we see that the slope estimate for these data is $\hat{\beta} = -.0563$. Combining this value with the (x_i, Y_j) pairs from Table 9.1, the five ordered A (9.32) values are $A^{(1)} \le \cdots \le A^{(5)}$: 1.2889, 1.3115, 1.3163, 1.3826, and 1.3852. As $n = 5$ is odd, we use (9.34) with $k = (5-1)/2 = 2$ to obtain the intercept estimate $\hat{\alpha} = A^{(3)} = 1.3163$. This estimate is provided by the command `theil`.

Combining the slope estimate of $\hat{\beta} = -.0563$ and this intercept estimate of $\alpha = 1.3163$, our final estimated linear relationship between the x and Y variables is then given by (9.36) to be

$$\overbrace{\text{med } Y_{x=x^*}} = \overbrace{[\text{median } Y \text{ when } x = x^*]} = 1.3163 - .0563x^*. \tag{9.37}$$

Thus, for example, we would estimate the median double-ratio value after 4.5 years of the cloud-seeding study to have been

$$\begin{aligned}\overbrace{\text{med } Y_{x=4.5}} &= \overbrace{[\text{median } Y \text{ when } x = 4.5 \text{ years}]} \\ &= 1.3163 - .0563(4.5) = 1.06295.\end{aligned}$$

Using (9.37) once again, the predicted double-ratio value if the study were to continue for a sixth year would be

$$\begin{aligned}\overbrace{\text{med } Y_{x=6}} &= \overbrace{[\text{median } Y \text{ when } x = 6 \text{ years}]} \\ &= 1.3163 - .0563(6) = 0.9785.\end{aligned}$$

One must always exercise caution in using an estimated linear relationship to predict typical values of the dependent variable Y for values of the independent variable x that are too different from the range of x-values used in establishing the estimated linear relationship (See Comment 12). For example, it would make no sense whatsoever to use the relationship (9.37) to estimate the median double-ratio value for negative values of x^*,

because they are not possible. In addition, although $x^* = 20$ years would certainly be a possible value for the independent variable (corresponding to 20 consecutive years of the cloud-seeding study), in order to use (9.37) to predict the typical double-ratio value Y after 20 years of the study would require the assumption that the regression relationship (9.1) remains linear for that extended time period. Although this may be a reasonable assumption to make, it is not one that comes automatically. Careful consideration should be given to its validity before using (9.37) to predict the double-ratio value that far into the future based solely on the 5 years of available data.

Comments

11. *Competing Estimators.* The intercept estimator given by (9.33) is not the only nonparametric estimator of α that has been studied in the statistical literature. In the case of symmetry of the underlying distribution for the error random variables $e_1, \ldots, e_n$ in Assumption A2, Hettmansperger and McKean (1977) proposed the competing estimator $\tilde{\alpha}$ associated with the median of the $n(n+1)/2$ Walsh averages of the n individual A_i (9.32) differences. Adichie (1967) proposed and studied the asymptotic properties of an entire class of estimators for α associated with rank tests.
12. *Appropriate Range of Values of the Independent Variable for Purposes of Prediction.* When we choose to use the estimated linear relationship (9.36) to predict typical values of the dependent variable Y for a particular setting of the independent variable x^*, caution must always be the rule. For prediction purposes, we must be relatively confident that the linear relationship holds at least approximately when x assumes the value x^*. This is seldom of concern when x^* is well situated among the values of the independent variable at which we observed sample-dependent variables in obtaining the estimated relationship (9.36) in the first place. However, when we are interested in predicting the typical value of the dependent variable Y for a setting of the independent variable x^* that is outside the range for which sample data had been collected in obtaining the estimated relationship (9.36), we must not automatically assume that the linear relationship (9.1) is still appropriate. Careful consideration must be given to justification of the reasonableness of this relationship for the particular problem of interest prior to using (9.36) for prediction purposes when considering such extended ranges of the independent variable.

Problems

20. Estimate α for the cenosphere-resin data of Table 9.2.

21. Estimate α for the body-weight and surface-area data for squirrel monkeys discussed in Problem 4.

22. Use the linear relationship (9.36) to estimate the typical density of a cenosphere-resin composite under hydrostatic pressure of 17,500 psi.

23. Use the linear relationship (9.36) to estimate the typical total surface area (cm^3) for a squirrel monkey with a body weight of 1000 g.

24. Estimate α for the tapeworm data discussed in Problems 8.1, 3, 11, and 17. Use the linear relationship (9.36) to estimate the typical weight of worms recovered from a dog that had been force-fed 20 mg cysticerci of *Taenia hydatigena.*

25. Consider the cenosphere-resin data of Table 9.2. Discuss the reasonableness of using the linear relationship (9.36) established in Problem 20 for these data to estimate the typical density of cenosphere-resin composites under hydrostatic pressures of 35,000, 75,000, and 200,000 psi.

26. Consider the body-weight and surface-area data for squirrel monkeys presented in Problem 4. Discuss the reasonableness of using the linear relationship (9.36) established in Problem 21 for these data to estimate the typical total surface area (cm^3) for squirrel monkeys with body weights of 320, 975, and 2500 g.

k(≥2) REGRESSION LINES

9.5 AN ASYMPTOTICALLY DISTRIBUTION-FREE TEST FOR THE PARALLELISM OF SEVERAL REGRESSION LINES (SEN, ADICHIE)

In this section, we discuss an asymptotically distribution-free procedure to test for parallelism of $k \geq 2$ regression lines. Thus, we are concerned with testing equality of the k slope parameters without additional constraints on the corresponding, unspecified intercepts.

Data. For the ith line, $i = 1, \ldots, k$, we observe the value of the ith response random variable Y_i at each of n_i fixed levels, $x_{i1}, \ldots, x_{in_i}$, of the ith independent (predictor) variable x_i. Thus, for the ith line, $i = 1, \ldots, k$, we obtain a set of observations $Y_{i1}, \ldots, Y_{in_i}$, where Y_{ij} is the value of the response variable Y_i when $x_i = x_{ij}$.

Assumptions

Bl. We take as our straight-line model

$$Y_{ij} = \alpha_i + \beta_i x_{ij} + e_{ij}, \quad i = 1, \ldots, k; \quad j = 1, \ldots, n_i, \tag{9.38}$$

where the x_{ij}'s are known constants and $\alpha_1, \ldots, \alpha_k$ and $\beta_1, \ldots, \beta_k$ are the unknown intercept and slope parameters, respectively.

B2. The $N = n_1 + \cdots + n_k$ random variables $e_{11}, \ldots, e_{1n_1}, \ldots, e_{k1}, \ldots, e_{kn_k}$ are mutually independent.

B3. The random variables $\{e_{i1}, \ldots, e_{in_i}\}, i = 1, \ldots, k$, are k random samples from a common continuous population with distribution function $F(.)$.

Hypothesis

The null hypothesis of interest here is that the k regression lines in model (9.38) have a common, but unspecified, slope, β, namely,

$$H_0 : [\beta_1 = \cdots = \beta_k = \beta, \text{ with } \beta \text{ unspecified}]. \tag{9.39}$$

Note that this null hypothesis does not place any conditions whatsoever on the intercept parameters $\alpha_1, \ldots, \alpha_k$. Thus, the assertion in H_0 (9.39) is simply that the k regression lines in model (9.38) are parallel.

Procedure

To construct the Sen–Adichie statistic V, we first align each of the k regression samples. Let $\bar{\beta}$ be the pooled least squares estimator for the common slope β under the null hypothesis H_0 (9.39), as given by

$$\bar{\beta} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i) Y_{ij}}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}, \tag{9.40}$$

where

$$\bar{x}_i = \sum_{j=1}^{n_i} \frac{x_{ij}}{n_i}, \quad \text{for } i = 1, \ldots, k. \tag{9.41}$$

For each of the k regression samples, compute the aligned observations

$$Y_{ij}^* = (Y_{ij} - \bar{\beta}\, x_{ij}), \quad i = 1, \ldots, k; \quad j = 1, \ldots, n_i. \tag{9.42}$$

Order these aligned observations Y_{ij}^* from least to greatest separately within each of the k regression samples. Let r_{ij}^* denote the rank of Y_{ij}^* in the joint ranking of the aligned observations $Y_{i1}^*, \ldots, Y_{in_i}^*$ in the ith regression sample.

Compute

$$T_i^* = \sum_{j=1}^{n_i} [(x_{ij} - \bar{x}_i) r_{ij}^*]/(n_i + 1), \quad i = 1, \ldots, k, \tag{9.43}$$

where $\bar{x}_i$ is given by (9.41). Setting

$$C_i^2 = \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2, \quad i = 1, \ldots, k, \tag{9.44}$$

the Sen–Adichie statistic V is then given by

$$V = 12 \sum_{i=1}^{k} \left[\frac{T_i^*}{C_i} \right]^2. \tag{9.45}$$

To test

$$H_0 : [\beta_1 = \cdots = \beta_k = \beta, \text{ with } \beta \text{ unspecified}]$$

versus the general alternative

$$H_1 : [\beta_1, \ldots, \beta_k \text{ not all equal}] \tag{9.46}$$

at the approximate α level of significance,

$$\text{Reject } H_0 \text{ if } V \geq \chi^2_{k-1,\alpha}; \quad \text{otherwise do not reject,} \tag{9.47}$$

where $\chi^2_{k-1,\alpha}$ is the upper α percentile of a chi-square distribution with $k - 1$ degrees of freedom. Values of $\chi^2_{k-1,\alpha}$ can be obtained from the `R` command `qchisq`.

Ties

If there are ties among the n_i aligned observations Y_{ij}^* (9.42) for the ith regression sample, use average ranks to break the ties and compute the weighted sum T_i^* (9.43) contribution to V (9.45) for that sample.

EXAMPLE 9.5 ***Ammonium Flux in Coastal Sediments.***

Coastal sediments are an important reservoir for organic nitrogen (ON). The degradation and mineralization of ON in coastal sediments is bacterially mediated and is known to involve several distinct steps. Moreover, it is possible to measure the rates of the processes at each of these steps. During the first stage of ON remineralization, ammonium is generated by heterotrophic bacteria during a process called *ammonification*. Ammonium can then be released to the environment or be microbially transformed to other nitrogenous species.

Mortazavi (1997) collected four sediment cores from Apalachicola Bay, Florida, and analyzed them at the Florida State University. The flux of ammonium (μ moles N per square meter of surface area) to the overlying water was measured for each core sample every 90 minutes during a 6-hour incubation period. These data are presented in Table 9.4 for the four core samples.

We are interested in assessing whether the rate of ammonium flux is similar across these four coastal sediments (at least over the 6-hour period of the study). Thus, if we

Table 9.4 Coastal Sediment Ammonium Flux in Apalachicola Bay, Florida

Core sample, i	Time, x_{ij} (h)	Ammonium flux, Y_{ij} (μ moles N/m^2)
Core 1	0	0
	1.5	33.019
	3	111.314
	4.5	196.205
	6	230.658
Core 2	0	0
	1.5	131.831
	3	181.603
	4.5	230.070
	6	258.119
Core 3	0	0
	1.5	33.351
	3	97.463
	4.5	196.615
	6	217.308
Core 4	0	0
	1.5	8.959
	3	105.384
	4.5	211.392
	6	255.105

Source: B. Mortazavi (1997).

let β_i correspond to the rate of ammonium flux for the ith coastal sediment core sample, $i = 1, \ldots, 4$, we are interested in testing the null hypothesis H_0 (9.39) against the general alternative (9.46) that the rates are not the same for the four coastal areas in Apalachicola from which the core samples were drawn.

First, we must obtain the pooled least squares estimator $\bar{\beta}$ (9.40). The set of x_{ij} values is the same for each of the coastal sediment samples, so

$$\bar{x}_1 = \bar{x}_2 = \bar{x}_3 = \bar{x}_4 = \frac{0 + 1.5 + 3 + 4.5 + 6}{5} = 3.$$

Hence, from (9.44), we obtain

$$\begin{aligned} C_1^2 = C_2^2 = C_3^2 = C_4^2 &= (0-3)^2 + (1.5-3)^2 + (3-3)^2 + (4.5-3)^2 + (6-3)^2 \\ &= 9 + 2.25 + 0 + 2.25 + 9 + 22.5, \end{aligned}$$

which, in turn, yields

$$\sum_{i=1}^{4} \sum_{j=1}^{5} (x_{ij} - \bar{x}_i)^2 = \sum_{i=1}^{4} C_i^2 = 4(22.5) = 90.$$

For the numerator of $\bar{\beta}$ (9.40), we see that

$$\begin{aligned} \sum_{i=1}^{4} \sum_{j=1}^{5} (x_{ij} - \bar{x}_i)(Y_{ij}) = [&(0-3)(0+0+0+0) \\ &+ (1.5-3)(33.019 + 131.831 + 33.351 + 8.959) \\ &+ (3-3)(111.314 + 181.603 + 97.463 + 105.384) \\ &+ (4.5-3)(196.205 + 230.070 + 196.615 + 211.392) \\ &+ (6-3)(230.658 + 258.119 + 217.308 + 255.105)] \\ = [&0 - 310.74 + 0 + 1251.423 + 2883.57] = 3824.253. \end{aligned}$$

Combining these two quantities, we obtain the value of the pooled least squares slope estimator (9.40) to be

$$\bar{\beta} = \frac{3824.253}{90} = 42.49.$$

Next, we create the aligned observations Y_{ij}^* (9.42) for each of the core samples:

$$\begin{aligned} \text{Core 1}: \quad Y_{11}^* &= 0 - 42.49(0) = 0 \\ Y_{12}^* &= 33.019 - 42.49(1.5) = -30.716 \\ Y_{13}^* &= 111.314 - 42.49(3) = -16.156 \\ Y_{14}^* &= 196.205 - 42.49(4.5) = 5 \\ Y_{15}^* &= 230.658 - 42.49(6) = -24.282 \end{aligned}$$

$$
\begin{aligned}
\text{Core 2}: \quad Y_{21}^* &= 0 - 42.49(0) = 0 \\
Y_{22}^* &= 131.831 - 42.49(1.5) = 68.096 \\
Y_{23}^* &= 181.603 - 42.49(3) = 54.133 \\
Y_{24}^* &= 230.070 - 42.49(4.5) = 38.865 \\
Y_{25}^* &= 258.119 - 42.49(6) = 3.179
\end{aligned}
$$

$$
\begin{aligned}
\text{Core 3}: \quad Y_{31}^* &= 0 - 42.49(0) = 0 \\
Y_{32}^* &= 33.351 - 42.49(1.5) = -30.384 \\
Y_{33}^* &= 97.463 - 42.49(3) = -30.007 \\
Y_{34}^* &= 196.615 - 42.49(4.5) = 5.41 \\
Y_{35}^* &= 217.308 - 42.49(6) = -37.632
\end{aligned}
$$

$$
\begin{aligned}
\text{Core 4}: \quad Y_{41}^* &= 0 - 42.49(0) = 0 \\
Y_{42}^* &= 8.959 - 42.49(1.5) = -54.776 \\
Y_{43}^* &= 105.384 - 42.49(3) = -22.086 \\
Y_{44}^* &= 211.392 - 42.49(4.5) = 20.187 \\
Y_{45}^* &= 255.105 - 42.49(6) = 0.165.
\end{aligned}
$$

Ordering these aligned observations Y_{ij}^* from least to greatest separately within each of the four core samples, we obtain the following within-samples rankings:

$$
\begin{aligned}
&\text{Core 1}: r_{11}^* = 4, \quad r_{12}^* = 1, \quad r_{13}^* = 3, \quad r_{14}^* = 5, \quad \text{and} \quad r_{15}^* = 2. \\
&\text{Core 2}: r_{21}^* = 1, \quad r_{22}^* = 5, \quad r_{23}^* = 4, \quad r_{24}^* = 3, \quad \text{and} \quad r_{25}^* = 2. \\
&\text{Core 3}: r_{31}^* = 4, \quad r_{32}^* = 2, \quad r_{33}^* = 3, \quad r_{34}^* = 5, \quad \text{and} \quad r_{35}^* = 1. \\
&\text{Core 4}: r_{41}^* = 3, \quad r_{42}^* = 1, \quad r_{43}^* = 2, \quad r_{44}^* = 5, \quad \text{and} \quad r_{45}^* = 4.
\end{aligned}
$$

The values of $T_1^*, \ldots, T_4^*$ are then obtained from (9.43) to be

$$
\begin{aligned}
T_1^* &= \frac{[(0-3)(4) + (1.5-3)(1) + (3-3)(3) + (4.5-3)(5) + (6-3)(2)]}{(5+1)} = 0, \\
T_2^* &= \frac{[(0-3)(1) + (1.5-3)(5) + (3-3)(4) + (4.5-3)(3) + (6-3)(2)]}{(5+1)} = 0, \\
T_3^* &= \frac{[(0-3)(4) + (1.5-3)(2) + (3-3)(3) + (4.5-3)(5) + (6-3)(1)]}{(5+1)} = -.75, \\
T_4^* &= \frac{[(0-3)(3) + (1.5-3)(1) + (3-3)(2) + (4.5-3)(5) + (6-3)(4)]}{(5+1)} = 1.5.
\end{aligned}
$$

Combining these T_i^* values with the corresponding values of C_i^2 previously obtained, we see from (9.45) that the Sen–Adichie statistic V for these data is given by

$$V = 12\left\{\frac{(0)^2}{22.5} + \frac{(0)^2}{22.5} + \frac{(-.75)^2}{22.5} + \frac{(1.5)^2}{22.5}\right\}$$
$$= 12\{0 + 0 + 0.25 + .1\} = 1.5.$$

For the Sen–Adichie procedure (9.47), we compare the value of V to the chi-square distribution with $k - 1 = 3$ degrees of freedom. We see that the observed value of $V = 1.5$ is the .318 percentile for the chi-square distribution with 3 degrees of freedom. Thus, the P-value for these data and test procedure (9.47) is .682, indicating that there is virtually no sample evidence in support of significant differences in the rates (slopes) of ammonium flux for the four coastal areas sampled.

The `R` command `sen.adichie` replicates this analysis. The argument is a list `z`. There are k items in the list `z`, one for each set of data corresponding to a specific linear relation. Each of these k items is a matrix with the first column the x values, the second the Y values.

Comments

13. *Motivation for the Test.* The pooled least squares estimator $\bar{\beta}$ (9.40) estimates some weighted combination, say β^*, of the k individual slopes $\beta_1, \ldots, \beta_k$ (9.38). From Assumptions Bl and B3, it follows that the aligned observations Y_{ij}^* (9.42), $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$, will tend to have values near

$$\begin{aligned}\text{med}(Y_{ij}^*) &= \text{med}(Y_{ij} - \bar{\beta}x_{ij}) \\ &\approx \text{med}(Y_{ij}) - \beta^* x_{ij} = \alpha_i + \beta_i x_{ij} - \beta^* x_{ij} + \text{med}(e_{ij}) \\ &= \alpha_i + (\beta_i - \beta^*)x_{ij} + \text{med}(e_{ij}). \end{aligned} \tag{9.48}$$

If the null hypothesis H_0 (9.39) is true, then $\beta_1 = \cdots = \beta_k = \beta^* = \beta$ and we would expect each of $Y_{i1}^*, \ldots, Y_{in_i}^*$ to be near $\alpha_i + \text{med}(e_{ij})$, for each of the regression samples $i = 1, \ldots, k$. As the r_{ij}^* ranks are obtained separately within each of the k samples, it follows that under H_0 (9.39) the ranks $r_{i1}^*, \ldots, r_{in_i}^*$ should behave like a random permutation of the integers $1, \ldots, n_i$ and exhibit no additional relationship with the regression constants $x_{i1}, \ldots, x_{in_i}$, for $i = 1, \ldots, k$. Thus, the null hypothesis setting should lead to values of T_i^* near zero, for $i = 1, \ldots, k$, and subsequently to small values of the Sen–Adichie test statistic V (9.45). On the other hand, if the null hypothesis H_0 (9.39) is not true, then some of the β_i's will be larger than β^* and some of them will be smaller than β^*. For those regression populations for which β_i is larger than β^*, we see from (9.48) that the aligned observations Y_{ij}^* (9.42) will be positively related to the values of the corresponding regression constants x_{ij}. This would tend to produce large positive values for the corresponding T_i^*'s (9.43). For those regression populations for which β_i is smaller than β^*, we see from (9.48) that the aligned observations Y_{ij}^* (9.42) will be negatively related to the values of the corresponding regression constants x_{ij}. This would tend to produce large negative values for the corresponding T_i^*'s (9.43). Each of these T_i^* values

is squared in the calculation of the Sen–Adichie test statistic V (9.45). Therefore, regression populations with either β_i larger or smaller than β^* will tend to produce large contributions to the test statistic V, providing partial motivation for procedure (9.47).

14. *Historical Development.* The general form of the test procedure (9.47), but using a rank estimate for the common value of the slope parameter β under H_0 (9.39) in the construction of the aligned observations Y_{ij}^* (9.42), was first proposed and studied by Sen (1969). The use of the pooled least squares estimator $\bar{\beta}$ (9.40) in the construction of the Y_{ij}^*'s was first suggested by Adichie (l984).

15. *Potthoff's Conservative Test of Parallelism.* For the case $k = 2$, Potthoff (1974) proposed a Wilcoxon-type test of $\beta_1 = \beta_2$. He compared each sample slope that can be computed from line 2 data with each sample slope that can be computed from line 1 data, scoring 1 if the sample 2 slope is larger than the sample 1 slope and 0 otherwise. His statistic was the average of the $n_1(n_1 - 1)(n_2)(n_2 - 1)/4$ such indicators. (To avoid complications, he assumed no two x_{1j}'s are equal and no two x_{2j}'s are equal.) The test associated with his statistic was neither distribution-free nor asymptotically distribution-free. Instead, he used an upper bound for the null variance of the statistic to produce a conservative test procedure.

16. *Competitor Based on Joint Rankings When the Intercept Is Common.* The Sen–Adichie procedure (9.47) is based on the individual rankings of the aligned observations Y_{ij}^* (9.42) *separately* within each of the k samples. This requires a good deal more computational time than if we could use a single simultaneous ranking of all $N = n_1 + \cdots + n_k$ aligned observations. Although such a joint ranking is not appropriate for the general model (9.38), Adichie (1974) proposed a procedure based on the joint ranking of all N of the aligned observations for settings where it is also reasonable to assume equality of the k intercepts $\alpha_1, \ldots, \alpha_k$ in model (9.38). Thus, Adichie's procedure is appropriate for testing H_0 (9.39) under Assumptions B2, B3 and the following more restrictive Assumption B1′ replacing Assumption B1:
Bl′. We take as our straight-line model

$$Y_{ij} = \alpha + \beta_i x_{ij} + e_{ij}, \quad i = 1, \ldots, k; j = 1, \ldots, n_i, \tag{9.49}$$

where the x_{ij}'s are known constants, α is the common (unknown) intercept and $\beta_1, \ldots, \beta_k$ are the unknown slope parameters, respectively.

Adichie's (1974) test statistic for this more restrictive setting is quite similar in form to the Sen–Adichie test statistic V (9.45). The major difference is the use of the single simultaneous ranking of all N of the aligned observations, rather than the k separate rankings utilized in constructing V. (We note, in passing, that the assumption of a common intercept α would be quite reasonable for the ammonium flux data considered in Example 9.5.)

17. *Test Procedures for Restricted Alternatives.* The Sen–Adichie procedure (9.47) is designed to test H_0 (9.39) against the class of general alternatives H_1 (9.46). Other authors have proposed nonparametric procedures designed to test H_0 against more restricted classes of alternatives. Adichie (1976) and Rao and Gore (1984) studied asymptotically distribution-free test procedures designed to reach a decision between H_0 and the class of ordered alternatives $H_2 : [\beta_1 \leq \cdots \leq \beta_k,$

with at least one strict inequality]. Finally, Kim and Lim (1995) considered asymptotically distribution-free procedures for testing H_0 against umbrella alternatives of the form $H_3 : [\beta_1 \leq \cdots \leq \beta_p \geq \beta_{p+1} \geq \cdots \geq \beta_k$, with at least one strict inequality].

18. *Comparing Several Regression Lines with a Control.* The Sen–Adichie procedure (9.47) is designed to test H_0 (9.39) against the class of general alternatives H_1 (9.46). In this context, the test involves a comparison of each regression line with every other regression line. For settings where one of the regression lines corresponds to a standard line for a control population, we might want to make only the $k - 1$ comparisons between the noncontrol regression lines and this control line. Lim and Wolfe (1997) proposed and studied an asymptotically distribution-free procedure for testing the null hypothesis H_0 (9.39) against the "treatments" versus control alternative $H_4 : [\beta_1 \leq \beta_i,\ i = 2, \ldots, k$, with at least one strict inequality], where, without loss of generality, the first regression line plays the role of the control line.

Properties

1. *Asymptotic Chi-Squareness.* See Sen (1969).
2. *Efficiency.* See Sen (1969) and Section 9.8.

Problems

27. Wells and Wells (1967) discussed Project SCUD, an attempt to study the effects of cloud seeding on cyclones. The basic hypothesis of interest was that cloud seeding in areas of cyclogenesis on the east coast of the United States had no measurable effect on the development of storms there. Table 9.5, based on a subset (Experiment 1 of Table I of Wells and Wells (1967)) of the observational data from Project SCUD, gives "*RI*" and "*M*" values for 11 seeded and 10 control units. The quantity *RI* is a measure of precipitation and the quantity *M*, the geostrophic meridional circulation index, was used in predicting cyclogenesis. Cyclones

Table 9.5 Precipitation Amounts *RI* and Circulation Index *M* for Seeded and Control Units

	Seeded		Control	
Unit, j	$x_{1j}(M)$	$Y_{1j}(RI)$	$x_{2j}(M)$	$Y_{2j}(RI)$
1	24	.180	−7	.138
2	28	.175	10	.081
3	30	.178	17	.072
4	37	.021	25	.188
5	43	.260	44	.075
6	47	.715	51	.435
7	52	.441	53	.423
8	57	.205	63	.339
9	71	.417	75	.519
10	87	.498	90	.738
11	115	.603		

Source: J. M. Wells and M. A. Wells (1967).

were expected to develop only when M was predicted positive. Test that the regression lines of RI on M for seeded and control units are parallel.

28. Consider the aligned observations Y_{ij}^* (9.42), $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$. Discuss why additional knowledge about the intercept parameters $\alpha_1, \ldots, \alpha_k$ is not necessary in order to use the *separate* within-samples ranks of the Y_{ij}^*'s in the construction of the test statistic V (9.45).

29. Consider the aligned observations Y_{ij}^* (9.42), $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$.

(a) Discuss why it would *not* be appropriate, in general, to use a single simultaneous ranking of all $N = n_1 + \cdots + n_k$ aligned observations in the construction of a statistic for testing H_0 (9.39).

(b) Under what conditions on the intercept parameters $\alpha_1, \ldots, \alpha_k$ might such a single simultaneous ranking be appropriate for developing a statistic to test H_0 (9.39)? (See Comment 16.)

30. Wardlaw and van Belle (1964) discussed the mouse hemidiaphragm method for assaying insulin. This procedure depends on the ability of the hormone to stimulate glycogen synthesis by the diaphragm tissue, in vitro. Hemidiaphragms are dissected from mice of uniform weight that have been starved for 18 hours. The tissues are incubated in tubes, and after incubation, the hemidiaphragms are washed with water and analyzed for glycogen content using anthrone reagent. The content is measured in terms of optical density. The procedure makes use of the fact that increasing the concentration of insulin in the incubation medium tends to increase glycogen synthesis by the hemidiaphragms. Specifically, for levels of insulin between .1 and 1.0 μ/ml, there is an approximate linear relationship between glycogen content and log concentration of insulin (See Wardlaw and Moloney (1961)).

The data in Table 9.6 are the log concentrations of insulin and the glycogen contents for 12 observations each from two varieties of insulin, namely, standard insulin and sample 1 insulin. For both standard and sample 1 lines, there are six observations at an insulin volume of .3 ml and six observations at a volume of 1.5 ml. In this insulin assay, and in many bioassays, the question of parallelism is extremely important, because the concept of relative potency (of a test preparation with respect to a standard) depends on the assumption that the dose–response lines are parallel. Using the data in Table 9.6, test the hypothesis that the dose–response lines for standard insulin and sample 1 insulin are parallel.

31. Experimental geneticists use survival under stressful conditions to compare the relative fitness of different species. Dowdy and Wearden (1991) considered data relating to the survival of

Table 9.6 Glycogen Content of Hemidiaphragms Measured by Optical Density in the Anthrone Test × 1000

	Standard insulin		Sample I insulin	
j	x_{1j} (log dose)	Y_{1j} (glycogen)	x_{2j} (log dose)	Y_{2j} (glycogen)
1	log (0.3)	230	log (0.3)	310
2	log (0.3)	290	log (0.3)	265
3	log (0.3)	265	log (0.3)	300
4	log (0.3)	225	log (0.3)	295
5	log (0.3)	285	log (0.3)	255
6	log (0.3)	280	log (0.3)	280
7	log (1.5)	365	log (1.5)	415
8	log (1.5)	325	log (1.5)	375
9	log (1.5)	360	log (1.5)	375
10	log (1.5)	300	log (1.5)	275
11	log (1.5)	360	log (1.5)	380
12	log (1.5)	385	log (1.5)	380

Source: A. C. Wardlaw and G. van Belle (1964).

Table 9.7 Numbers of *Drosophila* Flies (Three Different Species) That Survive to Adulthood after Exposure to Various Levels (ppm) of an Organic Phosphorus Insecticide

Species	Level of insecticide (ppm)	Number survived to adulthood
Drosophila melanogaster	0.0	91
	0.3	71
	0.6	23
	0.9	5
Drosophila pseudoobscura	0.0	89
	0.3	77
	0.6	12
	0.9	2
Drosophila serrata	0.0	87
	0.3	43
	0.6	22
	0.9	8

Source: S. Dowdy and S. Wearden (1991).

three species of *Drosophila* under increasing levels of organic phosphorus insecticide. Four batches of medium, identical except for the levels of insecticide they contained, were prepared. One hundred eggs from each of three *Drosophila* species were deposited on each of the four medium preparations and the level of insecticide (x) in parts per million (ppm) and number of *Drosophila* flies that survived to adulthood (y) for each combination are recorded in Table 9.7. Test the hypothesis that the three species of *Drosophila* exhibit the same response to increasing levels of insecticide in the medium studied.

32. Among the pieces of information used to assess the age of primates are measurements of skull, muzzle, and long-bone development. Reed (1973) collected such measurements for *Papio cynocephalus* baboons over a period of 5 years and developed a regression relationship between these attributes and age. A portion of Reed's data (from African colonies existing at the Southwest Foundation for Research and Education in San Antonio, Texas) is presented in Table 9.8 for male and female *Papio cynocephalus* baboons. The recorded data are age, in months, and the sum of skull, muzzle, and long-bone measurements, in millimeters.

Use these data to decide whether there is any difference in the slopes defining the linear relationships between age and the sum of skull, muzzle, and long-bone measurements for male and female *Papio cynocephalus* baboons.

GENERAL MULTIPLE LINEAR REGRESSION

9.6 ASYMPTOTICALLY DISTRIBUTION-FREE RANK-BASED TESTS FOR GENERAL MULTIPLE LINEAR REGRESSION (JAECKEL, HETTMANSPERGER–McKEAN)

The statistical procedures discussed in Sections 9.1–9.4 are concerned with the case of a straight-line relationship between a single independent (predictor) variable x and a response random variable Y. In Section 9.5, we presented a test procedure for assessing parallelism of two such straight-line relationships, each with a single independent (predictor) variable. However, in many settings where a regression relationship is of interest there are several independent (predictor) variables that potentially influence the value of a single response random variable. In this section, we present an asymptotically

Table 9.8 Age, in Months, and Sum of Skull, Muzzle, and Long-Bone Measurements, in Millimeters, for Male and Female *Papio cynocephalus* Baboons

	Male		Female	
j	x_{1j} (sum)	Y_{1j} (age)	x_{2j} (sum)	Y_{2j} (age)
1	175.0	1.36	175.0	1.58
2	183.0	2.20	183.0	2.48
3	190.0	3.05	190.0	3.40
4	200.0	4.45	200.0	4.92
5	211.0	6.19	211.0	6.87
6	220.0	7.78	220.0	8.66
7	230.0	9.70	230.0	10.86
8	239.5	11.66	239.5	13.14
9	245.5	12.96	245.5	14.67
10	260.0	16.33	260.0	18.68
11	271.5	19.21	271.5	22.14
12	284.0	22.52	284.0	26.18
13	291.0	24.46	291.0	28.57
14	302.5	27.78	302.5	32.68
15	314.0	31.25	314.0	37.03
16	318.5	32.65	318.5	38.80
17	327.0	35.36	327.0	42.22
18	337.0	38.65		
19	345.5	41.52		
20	360.0	46.61		
21	375.0	52.10		
22	384.5	55.69		
23	397.0	60.55		
24	411.0	66.18		
25	419.5	69.68		
26	428.5	73.47		
27	440.0	78.41		
28	454.5	84.81		

Source: O. M. Reed (1973).

distribution-free rank-based procedure for testing appropriate hypotheses in such a setting, commonly known as *multiple linear regression.*

Data. Let $\mathbf{x}' = (x_1, \ldots, x_p)$ be a row vector of p independent (predictor) variables and let $\mathbf{x}_1' = (x_{11}, \ldots, x_{p1}), \ldots, \mathbf{x}_n' = (x_{1n}, \ldots, x_{pn})$ denote n fixed values of this vector. At each of these fixed vectors $\mathbf{x}_1', \cdots \mathbf{x}_n'$, we observe the value of the single response random variable Y. Thus, we obtain a set of observations $Y_1, \ldots, Y_n$, where Y_i is the value of the response variable when $\mathbf{x}' = \mathbf{x}_i'$.

Assumptions

C1. Our model for multiple linear regression is

$$Y_i = \xi + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + e_i = \xi + \mathbf{x}_i' \boldsymbol{\beta}, \quad i = 1, \ldots, n, \quad (9.50)$$

where $\mathbf{x}_1' = (x_{11}, \ldots, x_{p1}), \ldots, \mathbf{x}_n' = (x_{1n}, \ldots, x_{pn})$ are vectors of known constants, ξ is the unknown "intercept" parameter, and $\boldsymbol{\beta}' = (\beta_1, \ldots, \beta_p)$ is a row

vector of unknown parameters, commonly referred to as the *set of regression coefficients*. For convenience later, we also write expression (9.50) in matrix notation. Let $\mathbf{Y}' = (Y_1, \ldots, Y_n), \boldsymbol{\xi}' = (\xi, \ldots, \xi)$, and set

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{21} & \ldots & x_{p1} \\ x_{12} & x_{22} & \ldots & x_{p2} \\ \vdots & \vdots & & \vdots \\ x_{1,n-1} & x_{2,n-1} & \ldots & x_{p,n-1} \\ x_{1n} & x_{2n} & \ldots & x_{pn} \end{bmatrix}. \tag{9.51}$$

Then, using matrix notation, the multiple linear regression model (9.50) can also be written as

$$\mathbf{Y} = \boldsymbol{\xi} + \mathbf{X}\boldsymbol{\beta}. \tag{9.52}$$

C2. The error random variables $e_1, \ldots, e_n$ are a random sample from a continuous distribution that is symmetric about its median 0, has cumulative distribution function $F(\cdot)$, and probability density function $f(\cdot)$ satisfying the mild condition that $\int_{-\infty}^{\infty} f^2(t)dt < \infty$.

Hypothesis

We are interested in testing the null hypothesis that a specific subset $\boldsymbol{\beta}_q$ of the regression parameters $\boldsymbol{\beta}$ are zero. Without loss of generality (because the ordering of the $(x_1, \beta_1), \ldots, (x_p, \beta_p)$ pairs in model (9.50) is arbitrary), we take this subset $\boldsymbol{\beta}_q$ to be the first q components of $\boldsymbol{\beta}$; that is, we take $\boldsymbol{\beta}'_q = (\beta_1, \ldots, \beta_q)$. Thus, we wish to test the null hypothesis

$$H_0 : [\boldsymbol{\beta}'_q = \mathbf{0}; \boldsymbol{\beta}'_{p-q} = (\beta_{q+1}, \ldots, \beta_p) \text{ and } \xi \text{ unspecified}]. \tag{9.53}$$

Thus, the null hypothesis asserts that the independent variables $x_1, \ldots, x_q$ do not play significant roles in determining the value of the dependent variable Y. (In many settings, we are interested in assessing the effect of *all* the independent variables simultaneously, which corresponds to taking $q = p$ in H_0 (9.53). Also see Problem 35.)

Procedure

To compute the Jaeckel–Hettmansperger–McKean test statistic HM, we proceed in several distinct steps. First, we obtain an unrestricted estimator for the vector of regression parameters $\boldsymbol{\beta}$. Let $R_i(\boldsymbol{\beta})$ denote the rank of $Y_i - \mathbf{x}'_i\boldsymbol{\beta}$ among $Y_1 - \mathbf{x}'_1\boldsymbol{\beta}, \ldots, Y_n - \mathbf{x}'_n\boldsymbol{\beta}$, as a function of $\boldsymbol{\beta}$, for $i = 1, \ldots, n$ (See Comment 20). The unrestricted estimator for $\boldsymbol{\beta}$, corresponding to a special case of a class of such estimators proposed by Jaeckel (1972), is then that the value of $\boldsymbol{\beta}$, say, $\hat{\boldsymbol{\beta}}$, that minimizes the measure of dispersion (once again, see Comment 20)

$$D_j(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \sum a(R_i(\boldsymbol{\beta}))(Y_i - \mathbf{x}'_i\boldsymbol{\beta}),$$

where a is a nondecreasing function on the integers $1, 2, \ldots, n$ such that $\Sigma_k a(k) = 0$. Typically, a is written as a score function ϕ on [0,1] by the relation $a(k) = \phi(k/(n+1))$.

The function ϕ is standardized so that $\int \phi = 0$ and $\int \phi^2 = 1$. One such ϕ is the Wilcoxon score function:

$$\phi(x) = \sqrt{12}\left(x - \frac{1}{2}\right), x \in [0, 1].$$

Using this ϕ, the dispersion is

$$D_J(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = (12)^{1/2}(n+1)^{-1}\sum_{i=1}^{n}\left[R_i(\boldsymbol{\beta}) - \frac{n+1}{2}\right](Y_i - \mathbf{x}_i'\boldsymbol{\beta}). \tag{9.54}$$

The estimator $\hat{\boldsymbol{\beta}}$ does not, in general, have a closed-form expression (See Comment 21 for a special case where such a closed-form expression is available), and iterative computer methods are generally necessary to obtain numerical solutions.

The second step in the computation of the Jaeckel–Hettmansperger–McKean test statistic HM involves repeating the steps leading to $\hat{\boldsymbol{\beta}}$ except now the minimization of the Jaeckel dispersion measure $D_J(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ is obtained under the condition imposed by the null hypothesis H_0 (9.53), namely, that $\boldsymbol{\beta}_q = \mathbf{0}$, with $\boldsymbol{\beta}_{p-q}$ unspecified. Let $\hat{\boldsymbol{\beta}}_0$ denote the value of $\boldsymbol{\beta}$ that minimizes $D_J(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ in (9.54) under the null constraint that $\boldsymbol{\beta}_q = \mathbf{0}$.

Let $D_J(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ and $D_J(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_0)$ denote the overall minimum and the minimum under the null constraint that $\boldsymbol{\beta}_q = \mathbf{0}$, respectively, of the Jaeckel dispersion measure $D_J(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ in (9.54) and set

$$D_J^* = D_J(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_0) - D_J(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \tag{9.55}$$

We note that D_J^* represents the drop or reduction in Jaeckel dispersion from fitting the full model as opposed to the reduced model corresponding to the null hypothesis H_0 (9.53) constraint that $\boldsymbol{\beta}_q = \mathbf{0}$.

The third and final step in the construction of the Jaeckel–Hettmansperger–McKean test statistic HM is the computation of a consistent estimator (See Comment 23) of a scale parameter τ. For the Wilcoxon score,

$$\tau = [12]^{-1/2}\left[\int_{-\infty}^{\infty} f^2(t)dt\right]^{-1}. \tag{9.56}$$

Combining the results of these three construction steps, the Jaeckel–Hettmansperger–McKean test statistic HM is given by

$$HM = \frac{2D_J^*}{q\hat{\tau}}. \tag{9.57}$$

When H_0 (9.53) is true, the statistic HM has, as n tends to infinity, an asymptotic F distribution with degrees of freedom q and $n - p - 1$, corresponding to the q constraints placed on β under H_0 and the total number p of predictors.

To test

$$H_0 : [\boldsymbol{\beta}_q' = \mathbf{0}; \boldsymbol{\beta}_{p-q}' = (\boldsymbol{\beta}_{q+1}, \ldots, \beta_p) \text{ and } \xi \text{ unspecified}]$$

against the general alternative

$$H_0 : [\boldsymbol{\beta}_q' \neq \mathbf{0}; \boldsymbol{\beta}_{p-q}' = (\beta_{q+1}, \ldots, \beta_p) \text{ and } \xi \text{ unspecified}] \tag{9.58}$$

at the approximate α level of significance,

$$\text{Reject } H_0 \text{ if } HM \geq F_{q,n-p-1,\alpha}; \quad \text{otherwise do not reject,} \tag{9.59}$$

where $F_{q,n-p-1,\alpha}$ is the upper α percentile of an F distribution with q and $n-p-1$ degrees of freedom. The values of $F_{q,n-p-1,\alpha}$ may be obtained from the R command `qf`.

Instead of an $F_{q,n-p-1,\alpha}$ critical value, one may remove the q from the denominator of the statistics HM and use a $\chi^2_{q,\alpha}$, the upper α percentile of a chi-square distribution. However, Hettmansperger and McKean (1977) and McKean and Sheather (1991) pointed out that the chi-square distribution is often too light tailed for use with small or moderate size samples.

Ties

If there are ties among $Y_1 - \mathbf{x}_1'\boldsymbol{\beta}, \ldots, Y_n - \mathbf{x}_n'\boldsymbol{\beta}$, use average ranks to break the ties in the computation of the minimum $D_J(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$. Similarly, if there are ties among $Y_1 - \mathbf{x}_1'\boldsymbol{\beta}_0 \ldots, Y_n - \mathbf{x}_n'\boldsymbol{\beta}_0$, use average ranks to break the ties in the computation of the minimum $D_J(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_0)$.

EXAMPLE 9.6 ***Snow Goose Departure Times.***

Wildlife science involves the study of how environmental conditions affect wildlife habits. Freund et al. (2010) report data on such a study to assess how a variety of environmental conditions affect the time that lesser snow geese leave their overnight roost sites to fly to their feeding areas. The data in Table 9.9 represent the following observations collected at a refuge near the Texas coast for 36 days of the 1987–1988 winter season:

TIME(Y) :	minutes before (−) or after (+) sunrise,
TEMP(x_1) :	air temperature in degrees Celsius,
HUM(x_2) :	relative humidity,
LIGHT(x_3) :	light intensity,
CLOUD(x_4) :	percent cloud cover.

Here, we consider a multiple regression analysis to assess the influence that the environmental conditions temperature (TEMP), relative humidity (HUM), light intensity (LIGHT), and percent cloud cover (CLOUD) have on the departure times (TIME) of lesser snow geese in this region of the country. For illustrative purposes, we consider the following three distinct null hypotheses:

$$H_{01} : [\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0; \xi \text{ unspecified}], \tag{9.60}$$

$$H_{02} : [\beta_1 = \beta_2 = 0; \beta_3, \beta_4, \text{ and } \xi \text{ unspecified}], \tag{9.61}$$

Table 9.9 Environmental Conditions Related to Snow Goose Departure Times

DATE	TIME	TEMP	HUM	LIGHT	CLOUD
11/10/87	11	11	78	12.6	100
11/13/87	2	11	88	10.8	80
11/14/87	−2	11	100	9.7	30
11/15/87	−11	20	83	12.2	50
11/17/87	−5	8	100	14.2	0
11/18/87	2	12	90	10.5	90
11/21/87	−6	6	87	12.5	30
11/22/87	22	18	82	12.9	20
11/23/87	22	19	91	12.3	80
11/25/87	21	21	92	9.4	100
11/30/87	8	10	90	11.7	60
12/05/87	25	18	85	11.8	40
12/14/87	9	20	93	11.1	95
12/18/87	7	14	92	8.3	90
12/24/87	8	19	96	12.0	40
12/26/87	18	13	100	11.3	100
12/27/87	−14	3	96	4.8	100
12/28/87	−21	4	86	6.9	100
12/30/87	−26	3	89	7.1	40
12/31/87	−7	15	93	8.1	95
01/02/88	−15	15	43	6.9	100
01/03/88	−6	6	60	7.6	100
01/05/88	−14	2	92	9.0	60
01/07/88	−8	2	96	7.1	100
01/08/88	−19	0	83	3.9	100
01/10/88	−23	−4	88	8.1	20
01/11/88	−11	−2	80	10.3	10
01/12/88	5	5	80	9.0	95
01/14/88	−23	5	61	5.1	95
01/15/88	−7	8	81	7.4	100
01/16/88	9	15	100	7.9	100
01/20/88	−27	5	51	3.8	0
01/21/88	−24	−1	74	6.3	0
01/22/88	−29	−2	69	6.3	0
01/23/88	−19	3	65	7.8	30
01/24/88	−9	6	73	9.5	30

Source: R. J. Freund, W. J. Wilson and D. Mohr (2010).

and

$$H_{03} : [\beta_2 = 0; \beta_1, \beta_3, \beta_4, \text{ and } \xi \text{ unspecified}]. \tag{9.62}$$

Consider the null hypothesis H_{01} (9.60). To test if all four parameters are 0, the function `rfit` from package `Rfit` (Kloke and McKean, 2011) is used. The data from Table 9.9 is in the `R` data frame `goose`. This data set includes a column of l's labeled `INT` to represent the intercept term ξ if desired. The calls to perform the rank regression are

```
rfit(TIME ~ TEMP + HUM + LIGHT + CLOUD, data=goose)
```

or

```
rfit(TIME ~ INT + TEMP + HUM + LIGHT + CLOUD, data=goose,
    intercept=F)
```

These produce identical results. The default score function is the Wilcoxon score. The estimates for the intercepts ξ and β_1 through β_4 are output as a result of either of the above calls:

```
Coefficients :

                    TEMP         HUM       LIGHT       CLOUD
-51.41229030 1.03912341 0.12628642 2.53480480 0.08951666
```

The estimates for the parameters β_i and ξ may be used to predict the time the geese leave for their feeding area given values x_i of the environmental variables `TEMP`, `HUM`, `LIGHT`, and `CLOUD`. This estimated regression relation is

$$\hat{Y} = -51.41 + 1.04x_1 + 0.13x_2 + 2.53x_3 + 0.09x_4.$$

More details on these parameter estimates may be obtained by the summary command on the rank fit object generated by `rfit`.

```
Coefficients;
       Estimate Std.Error t.value    p.value
     -51.412290  9.159212  -5.6132  4.128e-06
TEMP   1.039123  0.271468   3.8278  0.0006116
HUM    0.126286  0.116753   1.0817  0.2880251
LIGHT  2.534805  0.770792   3.2886  0.0025748
CLOUD  0.089517  0.045066   1.9863  0.0561974.
```

This provides the same parameter estimates for ξ and β_i. In addition, there are individual tests on whether a parameter is zero or not based on asymptotic normality.

It is possible that not all four of the predictor variables, x_i, are needed to model the response variable Y well in this example. Hypotheses tests on subsets of the β_i parameters can be used to choose a suitable model.

To test the hypotheses (9.60), (9.61), and (9.62), we will compare two models to each other: a full model and a reduced model. For hypothesis H_{01} (9.60), we are interested in comparing a full model with all $q = 4$ parameters β_i and the intercept ξ in it to a reduced model with no β_i parameters, only the intercept. We use D_J^*, the difference of the Jaeckel dispersions D_J for each of these models, to perform the comparison. In (9.55) the first term on the right is the Jaeckel dispersion for the reduced model and the second term is the dispersion for the full model. This difference in dispersions is standardized to become the test statistic *HM*.

The reduced model goes with the null hypothesis, the full model is paired with the alternative. For H_{01}, a reduced model with none of the four β_i is fit with a call of

```
r.01 <- rfit(TIME ~ INT, data=goose, intercept=F)
```

and the alternative hypothesis full model with all parameters is modeled by

```
f.01 <- rfit(TIME ~ TEMP + HUM + LIGHT + CLOUD,
        data=goose)
```

The `R` command `drop.test` will perform the test of H_{01} versus the alternative using the statistic *HM*. This command has two arguments: the first is the rank fit for the full model and the second is the rank fit for the reduced model. We compare the rank fits `f.01` and `r.01` with

```
drop.test(f.01, r.01)
```

The output of this call is

```
Drop in Dispersion Test
 F-Statistic     p-value
  1.7708e+01  1.1619e-07
```

The F-statistic is the value of *HM*. The individual components of this statistic given in (9.57) may be viewed directly. If the full versus reduced model comparison is saved to an R object by using the command `h.01 <- drop.test(f.01, r.01)`, then displaying the components of this analysis with `names(h.01)` shows that there are six pieces of information available: `F`, `p.value`, `RD` ("reduction in dispersion"), `tauhat`, `df1`, and `df2`. These values refer, respectively, to the statistic *HM*, the associated upper-tail *P*-value, $D_J^*, \hat{\tau}$, and the numerator and denominator degrees of freedom for the statistic. The *P*-value and *HM* are automatically displayed with `drop.test`. The others may be printed with the `$` indexing convention of R:

```
h.01$RD=294.0261
h.01$tauhat=8.30223
h.01$df1=4
h.01$df2=31
```

Note that the numerator degrees of freedom is $q = 4$ and the denominator degrees of freedom is $n - p - 1 = 36 - 4 - 1$, as expected. The estimate of $\hat{\tau}$ is found using the method of Koul et al. (1987). From (9.57),

$$HM = \frac{2D_J^*}{q\hat{\tau}} = \frac{2 \cdot 294.0261}{4 \cdot 8.30223} = 17.70766$$

agreeing with the output of `drop.test` above. For a particular α, a critical value could be obtained using the R command `qf(alpha, df 1=4, df 2=31, lower.tail=F)`. The *P*-value is obtained with `pf(17.70766, df 1=4, df2=31, lower.tail=F)` or taken from the output of `drop.test`. Given the low *P*-value for this data and hypothesis, we reject H_{01} in favor of the alternative hypothesis that not all of β_1 through β_4 are 0. Due to complexity of minimizing the Jaeckel dispersion measure, the values of the statistics found above may differ slightly when running R under various hardware and software configurations.

To get additional information about potential contributions of some of the individual independent (predictor) variables x_i, we make use of additional hypotheses. First, consider H_{02} (9.61). This null tests if $\beta_1 = \beta_2 = 0$ versus not both are 0. Under both H_{02} and the alternative H_{12}, ξ, β_3, and β_4 are unspecified. The full model is the same in this case as when testing H_{01} so the correct rank fit is again `f.01`. The reduced model is fit with

```
r.02 <- rfit(TIME ~ LIGHT + CLOUD, data=goose)
```

and the test is performed by

```
drop.test(f.01, r.02)
```

The output of this call is

```
Drop in Dispersion Test
 F-Statistic   p-value
   6.6681478 0.0039024
```

This null is rejected, implying that the variables `TEMP` and/or `HUM` contribute significantly (over and above the contributions of `LIGHT` and `CLOUD`) to the determination of

the time that lesser snow geese leave their overnight roost sites to fly to their feeding areas.

Finally, consider the third null hypothesis H_{03} (9.62). This hypothesis states that $\beta_2 = 0$, so the appropriate reduced model is fit by

```
r.03 <- rfit(TIME ~ TEMP + LIGHT + CLOUD, data=goose)
```

and the full model rank fit is again `f.01`. Using `drop. test (f.01, r.03)`, the result is

```
Drop in Dispersion Test
  F-Statistic   p-value
      1.72126  0.19916
```

Based on the large P-value, there is no evidence that the relative humidity `HUM` contributes significantly (over and above the contributions of `TEMP. LIGHT` and `CLOUD`) to the determination of the time that lesser snow geese leave their overnight roost sites to fly to their feeding areas. This does not conflict with the test of the hypothesis H_{02} because the alternative in that case is that not *both* β_1 `(TEMP)` and β_2 `(HUM)` are 0.

Comments

19. *Motivation for the Test.* Use of a measure of dispersion to assess the effectiveness of a model fit to a set of data is common in regression analysis. The estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{\mathbf{0}}$ are chosen to minimize the Jaeckel dispersion associated with the differences $\mathbf{Y}_i - \mathbf{x}_j'\boldsymbol{\beta}, i = 1, \ldots, n$, under no restrictions on $\boldsymbol{\beta}$ and under the null hypothesis restriction that $\boldsymbol{\beta} = (\boldsymbol{\beta}_q, \boldsymbol{\beta}_{p-q}) = (\mathbf{0}_q, \boldsymbol{\beta}_{p-q})$, respectively. Thus, the numerator of the Jaeckel–Hettmansperger–McKean test statistic $2D_J^* = 2[D_J(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathbf{0}}) - D_J(\mathbf{Y} - \mathbf{X}\hat{\beta})]$ is twice the drop or reduction in the Jaeckel dispersion from fitting the full model as opposed to the reduced null hypothesis model (9.53) with $\boldsymbol{\beta}_q = \mathbf{0}_q$. Large values of this drop in dispersion will lead to large values of HM (9.57) and are indicative of lack of agreement between the collected data and the null hypothesis. This serves as partial motivation for procedure (9.59).

20. *Translation lnvariance—"Effect" of the "Intercept" Parameter* ξ. The Jaeckel dispersion measure $D_J(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ is translation invariant in the sense that it is not affected by the unknown value of the "intercept" parameter, ξ. We note that the rank $R_i(\boldsymbol{\beta})$ of $Y_i - \mathbf{x}_i'\boldsymbol{\beta}$ among $Y_1 - \mathbf{x}_1'\boldsymbol{\beta}$, as a function of $\boldsymbol{\beta}$, is exactly the same as the rank of $Y_i - \xi - \mathbf{x}_i'\boldsymbol{\beta}$ among $Y_1 - \xi - \mathbf{x}_1'\boldsymbol{\beta}, \ldots Y_n - \xi - \mathbf{x}_n'\boldsymbol{\beta}$, for $i = 1, \ldots, n$. It follows that the Jaeckel measure of dispersion is independent of the value of the intercept parameter ξ, because

$$D_J(\mathbf{Y} - \xi - \mathbf{X}\boldsymbol{\beta}) = (12)^{1/2}(n+1)^{-1} \sum_{i=1}^{n} \left[R_i(\boldsymbol{\beta}) - \frac{n+1}{2} \right] (Y_i - \xi - \mathbf{x}_i'\boldsymbol{\beta})$$

$$= D_J(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) - \xi[(12)^{1/2}(n+1)^{-1}] \sum_{i=1}^{n} \left[R_i(\boldsymbol{\beta}) - \frac{n+1}{2} \right]$$

$$= D_J(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}), \text{because} \sum_{i=1}^{n} \left[R_i(\boldsymbol{\beta}) - \frac{n+1}{2} \right] = 0.$$

21. *Closed-Form Expression for $\hat{\boldsymbol{\beta}}$-Special Case of Straight-Line Regression.* One situation where the unrestricted estimator $\hat{\boldsymbol{\beta}}$ minimizing $D_J(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$ in (9.54) has a closed-form expression is when we have only a single independent (predictor) variable so that model (9.50) corresponds to a straight-line regression $Y_i = \xi + \beta_1 x_i + e_i, i = 1, \ldots, n$. For this setting, the Jaeckel (1972) estimator of the slope β_1 is a weighted median of the set of all pairwise slopes $S_{ij} = (Y_j - Y_i)/(x_j - x_i)$, for (i, j) such that $x_i \neq x_j$. This particular estimator was first derived using different criteria and studied by Adichie (1967).

22. *Estimation of the "Intercept" ξ.* Because the Jaeckel dispersion measure $D_J(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ is independent of the unknown value of the "intercept" ξ (See Comment 20), the estimator $\hat{\boldsymbol{\beta}}$ obtained by minimizing $D_J(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ does not provide any information relative to ξ. Hettmansperger and McKean (1977) suggested using the full-model residuals $e_i^* = Y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}$, for $i = 1, \ldots, n$, to estimate ξ. In particular, they proposed the estimator

$$\hat{\xi} = \text{median} \left\{ \frac{e_i^* + e_j^*}{2}, 1 \leq i \leq j \leq n \right\}. \tag{9.63}$$

(We note, in passing, that $\hat{\xi}$ is simply the Hodges–Lehmann one-sample estimator $\hat{\theta}$ (3.23) applied to the full-model residuals $e_1^*, \ldots, e_n^*$.)

23. *Estimation of the Parameter τ (9.56).* Part of the construction of the Jaeckel–Hettmansperger–McKean test statistic, *HM* (9.57), is the computation of a consistent estimator of the parameter τ (9.56). A variety of approaches leading to a number of competing consistent estimators have been considered in the literature. Hettmansperger and McKean (1977) suggested a consistent estimator for τ based on the length of a Wilcoxon signed rank confidence interval (See Section 3.3) applied to the full-model residuals $e_1^*, \ldots, e_n^*$ discussed in Comment 22. Koul, Sievers, and McKean (1987) recommended an estimator of τ based on the empirical distribution function of the absolute differences of the full-model residuals $e_1^*, \ldots, e_n^*$. An approach to the estimation of τ based on window- or kernel-type estimation of the probability density function $f(\cdot)$ has been considered by Schuster (1974) and Schweder (1975, 1981).

24. *Generalized Score Functions.* The rank regression procedure discussed in this section is based on the use of the Wilcoxon-type scoring function of the ranks in the construction of the Jaeckel dispersion function $D_J(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ in (9.54). Other scoring functions, such as $\boldsymbol{\Phi}^{-1}(\cdot)$ associated with the van der Waerden test discussed in Comment 4.12, were also considered by Jaeckel (1972) in the construction and study of an entire class of rank-based dispersion measures for the multiple linear regression setting.

25. *Test for a More General Null Hypothesis.* Hettmansperger, McKean, and Sheather (1997) described a generalization of the null hypothesis presented in H_0 (9.53). They discussed statistical procedures for the more inclusive problem of testing $H_0^* : \mathbf{M}\boldsymbol{\beta} = 0$ versus the general alternative $H_A^* : \mathbf{M}\boldsymbol{\beta} \neq \mathbf{0}$, where $\mathbf{M}$ is an arbitrary full row rank $q \times p$ matrix, for some $q \leq p$.

    As an example, consider the null hypothesis $H_{04} : [\beta_1 = \beta_2; \beta_3, \beta_4$, and ξ unspecified] against the alternative $H_{14} : [\beta_1 \neq \beta_2; \beta_3, \beta_4$, and ξ unspecified] for the snow geese data from Example 9.6. This is equivalent to setting $\boldsymbol{M}$ to the

single row matrix [1 −1 0 0] so that $\mathbf{M}\boldsymbol{\beta} = \mathbf{0}$ is the same as H_{04}. In R, we again use `drop.test` with the full model `f.01` from Example 9.6 and the reduced model from

```
r.04 <- rfit(TIME ~ I(TEMP + HUM) + LIGHT + CLOUD,
        data=goose)
```

where the `I` command uses the + symbol to force `TEMP` and `HUM` to be considered as a sum, rather than using the + symbol to represent additional terms in the rank regression model. The estimated rank regression equation for the reduced model is

```
Coefficients :
               I (TEMP + HUM)     LIGHT      CLOUD
  -65.6185750       0.2366911 3.3860983 0.1346024
```

and the test of the hypotheses H_{04} against H_{14} is given by

```
Drop in Dispersion Test
 F-Statistic   p-value
   8.4534141 0.0066766
```

Thus, there is good evidence that the independent variables temperature and relative humidity do not contribute in the same degree to the determination of the time that lesser snow geese leave their overnight roost sites to fly to their feeding areas. This finding is in good agreement with what had been established previously in Example 9.6.

26. *Extension to the General Linear Model.* Our discussion of rank-based regression in this section has only touched upon a small portion of a much more extensive rank-based approach to the large class of linear models. Although a discussion of this more general setting is beyond the scope of this text, we recommend that the interested reader take advantage of two excellent survey articles on this topic by Draper (1988) and Hettmansperger, McKean, and Sheather (1997).

Properties

1. *Consistency.* Under certain regularity conditions (see, for example, Hettmansperger, McKean, and Sheather (1997)), the test defined by (9.59) is consistent against the alternatives H_1 (9.58).
2. *Asymptotic Chi-Squareness.* See McKean and Hettmansperger (1976).
3. *Efficiency.* See McKean and Hettmansperger (1976) and Section 9.8.

Problems

33. In heart catheterization, a 3-mm-diameter Teflon catheter (tube) is inserted into a major vein or artery at the femoral region and maneuvered up into the heart itself to assess the heart's physiology and functional ability. Heart catheterizations are sometimes performed on children with congenital heart defects. In such cases, the length of the catheter is often determined by a physician's educated guess. Rice (2007) considered a data set obtained by Weindling (1977) in a preliminary study involving 12 children. For each child, the exact catheter length required was determined by using a fluoroscope to check that the tip of the catheter had reached the pulmonary artery. The 12 catheter lengths (cm) and the heights (in) and weights (lb) for the 12 children in the study are given in Table 9.10.

Table 9.10 Required Length of Heart Catheter as a Function of Height and Weight

Child	Height, in	Weight, lb	Length of heart catheter, cm
1	42.8	40.0	37.0
2	63.5	93.5	49.5
3	37.5	35.5	34.5
4	39.5	30.0	36.0
5	45.5	52.0	43.0
6	38.5	17.0	28.0
7	43.0	38.5	37.0
8	22.5	8.5	20.0
9	37.0	33.0	33.5
10	23.5	9.5	30.5
11	33.0	21.0	38.5
12	58.0	79.0	47.0

Source: J. A. Rice (2007).

Treating length of heart catheter as the independent variable, test for the importance of height and weight in determining the required catheter length.

34. Iman (1994) considered data obtained by Leaf et al. (1989) in a study of options for reducing concentrations of total plasma triglycerides. Leaf et al. obtained measurements of the following variables on each of 13 patients:

Y: Total triglyceride level, mmol/l,
x_1: Sex of the patient (coded as female = 0, male = 1),
x_2: Whether patient is obese (coded as no = 0, yes = 1),
x_3: Chylo-microns,
x_4: Very low density lipoprotein (VLDL),
x_5: Low density lipoprotein (LDL),
x_6: High density lipoprotein (HDL),
x_7: Age of the patient.

These data are presented in Table 9.11.

(a) Including all the measured variables, find the approximate P-value for a test of the null hypothesis that obesity does not play a significant role in determination of the total triglyceride level.
(b) Including all the measured variables, find the approximate P-value for a test of the null hypothesis that none of the lipoproteins play significant roles in determination of the total triglyceride level.
(c) Does age play a significant role in the determination of the total triglyceride level, when all the measured variables are taken into account? Justify your answer.
(d) Find an approximate P-value for a test of the null hypothesis that none of the measured variables contribute significantly to the determination of the total triglyceride level.

35. Consider the multiple linear regression model in (9.50). Often it is the case that we are interested in testing whether *any* of the independent variables $x_1, \ldots, x_p$ have significant effects on the determination of the value of the dependent random variable Y. This corresponds to taking $q = p$ in the statement of the null hypothesis (9.53). For this setting, what would be the form of the null constraint estimator $\hat{\boldsymbol{\beta}}_{\mathbf{0}}$? Provide a closed-form expression for $D_J(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathbf{0}})$ for this setting in terms of the ordered Y observations, $Y^{(1)} \leq \cdots \leq Y^{(n)}$.

36. Freund et al. (2010) presented a set of data relating survival times (TIME) of liver transplant patients to the following information collected from the patients prior to their transplant operations:

Table 9.11 Blood Plasma Measurements Related to Total Triglyceride Level

Patient	Total triglyceride level	Sex/Obese	Chylo-microns	VLDL	LDL	HDL	Age
1	20.19	1/1	3.11	4.51	2.05	0.67	53
2	27.00	0/1	4.90	6.03	0.67	0.65	51
3	51.75	0/0	5.72	7.98	0.96	0.60	54
4	51.36	0/1	7.82	9.58	1.06	0.42	56
5	28.98	1/1	2.62	7.54	1.42	0.36	66
6	21.70	0/1	1.48	3.96	1.09	0.23	37
7	14.40	1/1	0.57	8.60	2.16	0.83	41
8	15.14	1/1	0.60	5.46	1.58	0.85	55
9	50.00	1/1	6.29	13.03	1.48	0.28	43
10	23.73	1/1	1.94	7.12	0.91	0.57	58
11	29.33	0/1	0.52	8.94	1.58	0.88	39
12	19.98	0/1	1.11	5.85	1.19	0.62	41
13	13.28	1/0	1.61	3.73	1.58	0.62	54

Source: D. A. Leaf, W. E. Connor, R. Illingworth, S. P. Bacon, and G. Sexton (1989) and R. L. Iman (1994).

CLOT: a measure of the clotting potential of the patient's blood

PROG: a subjective index of the patient's prospect of recovery

ENZ: a measure of a protein present in the body

LIV: a measure relating to white blood cell count.

These data for 54 liver transplant patients are presented in Table 9.12. Examine the relationship of survival time (TIME) to the four measured preoperation variables. Which of them provide significant input into the determination of survival time for liver transplant patients?

37. In Section 9.1, we discussed a procedure designed to test the effect of a single independent (predictor) variable x on a dependent random variable Y when the anticipated relationship between x and Y is linear. Sometimes, the anticipated relationship between x and Y is better described by a higher order polynomial in x, rather than a simple linear relationship. Discuss how the general procedure presented in this section can be used to test for a relationship between x and Y that is best described by a polynomial of degree $p > 1$.

38. Consider the cenosphere-resin composite data of Problem 1. In that problem, you were asked to assess the significance of a possible linear relationship between hydrostatic pressure, x, and the density of the cenosphere-resin composite, Y. Suppose that someone suggested that the relationship between x and Y might be better represented by a cubic polynomial through the expression

$$E[Y|x] = \xi + \beta_1 x + \beta_2 x^2 + \beta_3 x^3,$$

where ξ, β_1, β_2, and β_3 are unknown parameters. (See Problem 37.)

(a) Find the approximate P-value for an appropriate test of the null hypothesis that there is no (cubic, quadratic, or linear) significant relationship between x and Y.

(b) Find the approximate P-value for an appropriate test of the null hypothesis that the relationship between x and Y is actually quadratic, as opposed to cubic.

(c) Find the approximate P-value for an appropriate test of the null hypothesis that the relationship between x and Y is actually linear, as opposed to either quadratic or cubic.

39. Hettmansperger, McKean, and Sheather (1997) described the following generalization of the null hypothesis presented in H_0 (9.53). They discussed statistical procedures for the more general problem of testing $H_0^* : \mathbf{M}\boldsymbol{\beta} = \mathbf{0}$ versus the general alternative $H_A^* : \mathbf{M}\boldsymbol{\beta} \neq \mathbf{0}$, where

Table 9.12 Survival Times of Liver Transplant Patients and Related Biological Measurements

Patient	TIME	CLOT	PROG	ENZ	LIV
1	34	3.7	51	41	1.55
2	58	8.7	45	23	2.52
3	65	6.7	51	43	1.86
4	70	6.7	26	68	2.10
5	71	3.2	64	65	0.74
6	72	5.2	54	56	2.71
7	75	3.6	28	99	1.30
8	80	5.8	38	72	1.42
9	80	5.7	46	63	1.91
10	87	6.0	85	28	2.98
11	95	5.2	49	72	1.84
12	101	5.1	59	66	1.70
13	101	6.5	73	41	2.01
14	109	5.2	52	76	2.85
15	115	5.4	58	70	2.64
16	116	5.0	59	73	3.50
17	118	2.6	74	86	2.05
18	120	4.3	8	119	2.85
19	123	6.5	40	84	3.00
20	124	6.6	77	46	1.95
21	125	6.4	85	40	1.21
22	127	3.7	68	81	2.57
23	136	3.4	83	53	1.12
24	144	5.8	61	73	3.50
25	148	5.4	52	88	1.81
26	151	4.8	61	76	2.45
27	153	6.5	56	77	2.85
28	158	5.1	67	77	2.86
29	168	7.7	62	67	3.40
30	172	5.6	57	87	3.02
31	178	5.8	76	59	2.58
32	181	5.2	52	86	2.45
33	184	5.3	51	99	2.60
34	191	3.4	77	93	1.48
35	198	6.4	59	85	2.33
36	200	6.7	62	81	2.59
37	202	6.0	67	93	2.50
38	203	3.7	76	94	2.40
39	204	7.4	57	83	2.16
40	215	7.3	68	74	3.56
41	217	7.4	74	68	2.40
42	220	5.8	67	86	3.40
43	276	6.3	59	100	2.95
44	295	5.8	72	93	3.30
45	310	3.9	82	103	4.55
46	311	4.5	73	106	3.05
47	313	8.8	78	72	3.20
48	329	6.3	84	83	4.13
49	330	5.8	83	88	3.95
50	398	4.8	86	101	4.10
51	483	8.8	86	88	6.40
52	509	7.8	65	115	4.30
53	574	11.2	76	90	5.59
54	830	5.8	96	114	3.95

Source: R. J. Freund, W. J. Wilson and D. Mohr (2010).

M is an arbitrary full row rank $q \times p$ matrix, for some $q \leq p$ (See Comment 25). Within this more general setting, what matrix **M** corresponds to the special case of the null hypothesis H_0 in (9.53)?

40. In an attempt to gain a better understanding of the complexities of air pollution in general and to predict pollutant levels in particular, the Los Angeles Pollution Control District routinely records the levels of pollutants and several meteorological conditions at various sites around the city. As reported by Rice (2007), the data in Table 9.13 represent the maximum level of an oxidant (a photochemical pollutant) and the morning averages of the four meteorological variables: wind speed, temperature, humidity, and insolation (measure of amount of sunlight) over a 30-day period in a single summer.

Ignoring the distinct possibility that there is some degree of correlation between maximum oxidant levels collected on adjacent days (which would violate Assumption C3 regarding the independence of the observations of the dependent variable), examine the relationship of oxidant level to the four meteorological variables. Which of them contribute significantly to the maximum oxidant level on a given day for the Los Angeles Pollution Control District?

41. For the data discussed in Problem 40, consider a multiple linear regression of the maximum oxidant level on the four meteorological measurements. Find the approximate P-value for

Table 9.13 Maximum Oxidant Level, Wind Speed, Temperature, Humidity, and Insolation for a 30-Day Summer Period in the Los Angeles Pollution Control District

Day	Oxidant level	Wind speed	Temperature	Humidity	Insolation
1	15	50	77	67	78
2	20	47	80	66	77
3	13	57	75	77	73
4	21	38	72	73	69
5	12	52	71	75	78
6	12	57	74	75	80
7	12	53	78	64	75
8	11	62	82	59	78
9	12	52	82	60	75
10	20	42	82	62	58
11	11	47	82	59	76
12	17	40	80	66	76
13	20	42	81	68	71
14	23	40	85	62	74
15	17	48	82	70	73
16	16	50	79	66	72
17	10	55	72	63	69
18	11	52	72	61	57
19	11	48	76	60	74
20	9	52	77	59	72
21	5	52	73	58	67
22	5	48	68	63	30
23	4	65	67	65	23
24	7	53	71	53	72
25	18	36	75	54	78
26	17	45	81	44	81
27	23	43	84	46	78
28	23	42	83	43	78
29	24	35	87	44	77
30	25	43	92	35	79

Source: J. A. Rice (2007).

Table 9.14 Number of *Chaoborus* Larvae and Water Quality of Samples

Sample	Number of larvae	Depth	Brackishness	Dissolved oxygen
1	35	8.4	8.0	1.0
2	10	2.0	6.5	8.5
3	9	3.5	6.2	6.5
4	30	10.4	5.0	1.5
5	20	6.5	6.5	7.5
6	23	6.2	7.3	4.5
7	28	12.4	6.4	4.0
8	8	7.0	6.0	10.0
9	29	5.8	6.1	3.0
10	4	3.0	5.4	11.0
11	18	6.0	7.3	4.5
12	14	5.5	6.6	5.5
13	32	9.0	6.5	2.5
14	6	1.1	5.8	7.0

Source: S. Dowdy and S. Wearden (1991).

an appropriate test of the null hypothesis that the regression coefficients for wind speed and humidity are the same, as are the regression coefficients for temperature and insolation. (See Comment 25.)

42. Dowdy and Wearden (1991) considered the relationship between several environmental factors and the number of larvae of the phantom midge, genus *Chaoborus*, which is similar to a mosquito in appearance, but is not bloodsucking. The larva burrows into the sediment at the bottom of a body of water and remains there during the daylight hours. At night, it migrates to the surface of the water to feed. The larva is itself eaten by larger animals and therefore plays an important role in the food chain for freshwater fish. A team of biologists studied a recreational lake created by damming a small stream and recorded the following measurements at each of 14 sampling points in the lake:

Y: number of larvae of *Chaoborus* collected in a grab sample of the sediment from an area of approximately 225 cm^2 of lake bottom.

X_1: depth (meters) of the lake at the sampling point.

x_2: brackishness (conductivity) of the water at the lake bottom (recorded in mhos per decimeter).

x_3: dissolved oxygen (milligrams per liter) in the water sampled from the lake bottom.

The data from these 14 sampling points are presented in Table 9.14. Examine the relationship of the number of *Chaoborus* larvae to the three measured water quality variables. Which of them provide significant input into the determination of the number of *Chaoborus* larvae in a lake environment?

NONPARAMETRIC REGRESSION ANALYSIS

9.7 AN INTRODUCTION TO NON-RANK-BASED APPROACHES TO NONPARAMETRIC REGRESSION ANALYSIS

In all the previous sections of this chapter, the *modus operandi* has been to consider a specific regression model with associated parameters (e.g., straight line, two straight

lines, multiple linear regression) and then to discuss appropriate rank-based procedures for making statistical inferences about the unknown parameters. They have all been nonparametric in nature, in that the inferential procedures were not dependent upon the assumption of a particular underlying distribution for the error terms. Recently, however, there has been considerable research activity in the literature in an arena that has become known generally as *nonparametric regression.* Although it maintains an indifference to the form of the underlying distribution for the error terms, the distinction in this new area of endeavor is that even a specific regression model is no longer stipulated a priori. The data are asked to provide not only the eventual statistical inference but also aid with the development of an appropriate regression relationship between the dependent random variable and the independent predictor variables(s). Thus, the intent of these nonparametric regression procedures is to permit the data to aid in both the selection of an appropriate model for the regression relationship and the inferences eventually drawn from this model.

All the procedures previously discussed in this chapter have also been rank-based, in the sense that some form of ranking was used in arriving at the appropriate inferences. When the model itself is open for data input, however, ranks are no longer sufficient to provide both model selection and inferential procedures. Hence, the procedures associated with this field known as *nonparametric regression* do not generally utilize rankings in reaching their conclusions. As a result, they are often more complicated and computationally intensive than the level assumed throughout the rest of this text. Consequently, our approach in this section will be to discuss briefly some of the statistical techniques that are commonly used in developing such nonparametric regression procedures rather than to provide details of specific procedures and their applications to appropriate data sets. More detailed summaries of various aspects of this general area of nonparametric regression are provided in Chapters 13 and 14 and, for example, in Cleveland (1994), Eubank (1999), Hastie and Tibshirani (1990), Wasserman (2006), and Ryan (2009).

We concentrate here on the setting where we are interested in obtaining information about the relationship between a single dependent random variable Y and a single independent (predictor) variable x. For available procedures in the area of nonparametric regression when there are multiple independent variables, the reader is referred to work by Friedman (1991) and Stone (1994), for example.

Data. At each of n fixed values, $x_1, \ldots, x_n$, of the independent (predictor) variable x, we observe the value of the response random variable Y. Thus, we obtain a set of observations $Y_1, \ldots, Y_n$, where Y_i is the value of the response variable when $x = x_i$.

Assumptions

The most general nonparametric regression relationship between Y_i and x_i is given by

$$Y_i = \mu(x_i) + e_i, \quad \text{for} \quad i = 1, \ldots, n, \tag{9.64}$$

where the random variables $e_1, \ldots, e_n$ are a random sample from a continuous population that has median 0.

The goal, of course, is to make valid statistical inferences about the form of the regression function $\mu(\cdot)$. Depending on the specific approach to nonparametric regression under consideration, a variety of additional *regularity* conditions are often imposed on the form of $\mu(\cdot)$ to enable development of appropriate statistical methodology.

As there is likely to be a good deal of fuzziness (variability) in the response data $Y_1, \ldots, Y_n$, it is often difficult to describe the relationship between x and Y, as expressed in the median $\mu(x)$, without the aid of a more formal model. Therefore, we search for ways to dampen, or "smooth," the fluctuations present in the Y observations as we move along the various x values. In this section, we discuss a variety of ways to approach this smoothing of the data. Ryan (2009) referred to each of these smoothing techniques as a "smoother" and to the associated estimates $\hat{\mu}(x_i)$ at the nx_i values as a "smooth." The first four smoothers discussed in this section are linear smoothers, in the sense that the estimates $\hat{\mu}(x_1), \ldots, \hat{\mu}(x_n)$ in a particular smooth are always linear combinations of the observations $Y_1, \ldots, Y_n$. The fifth smoother is nonlinear.

Running Line Smoother. One of the earliest attempts at nonparametric regression is associated with the running line smoother proposed and studied by Cleveland (1979). For this smoother, a moving window of points is utilized and a simple least squares linear regression line is computed each time a point is deleted and another added as the window moves along the x values. The plot of these running lines as a function of the independent variable x is referred to as the *running line smoother estimator* for $\mu(x)$.

A number of issues are important relative to the construction of running line smoothers. First, one must decide on how many points are to be used in each window (i.e., the window *size* or *size of the neighborhood)* for which the least squares regression line is to be computed. Clearly, a window size that is too small will result in very little smoothing of the data, whereas a window size that is too large will virtually force a single straight-line relationship on the data, regardless of its validity. This choice of window size is discussed in Hastie and Tibshirani (1987), where they indicate that a window size of roughly 10–15% of the data is reasonable. Another matter of concern with running line smoothers is how to deal with the extremes of the data, where symmetric windows are not possible. Statistical inference about $\hat{\mu}(x)$ associated with running line smoothers is addressed in Hastie and Tibshirani (1990).

Kernel Regression Smoother. As with the running line smoother, the kernel regression smoother utilizes *neighborhood data* to provide its estimate of the regression function $\mu(x)$. In this setting, the neighborhoods are often referred to as *strips* and the size of a strip is called the *bandwidth*. One of the clear distinctions between running line smoothers and kernel regression smoothers is in how they weight the observations in a given window. For a running line smoother, the points in a neighborhood are equally weighted, although, of course, they could have differing influences on the estimation process. On the other hand, for a kernel regression smoother, the distance of the points in a neighborhood from the center of a neighborhood, say, x_0, is used to differentially weight their contributions. Basically, in the process of estimating $\mu(x_0)$, no weight is given to those observations outside of the neighborhood centered at x_0 and the greatest weight in the neighborhood is given to those observations Y_i for which the corresponding x_i are closest to x_0. A *kernel function* is utilized to assign these differential weights to the observations across the various neighborhoods.

Altman (1992) addressed the question of how many strips to use in constructing a kernel regression smoother, as well as some related procedures for statistical inference. The selection of a kernel function and its relationship to the stipulation of both the number of strips and the bandwidth is discussed in Hastie and Tibshirani (1990) and Härdle (1992). One particular shortcoming of kernel regression smoothers is that their performance at the boundaries of the predictor region can be rather poor, as documented by Hastie and Loader (1993) and Fan and Marron (1993).

Local Regression Smoother. Local regression smoothers were first introduced by Cleveland (1979), where he referred to the process as *locally weighted regression*. Local regression smoothers once again use overlapping neighborhoods and, as with the kernel regression smoothers, weight the contributions of points to the estimation of $\mu(x_0)$ in an inverse relationship to their distances from x_0. The estimation in a particular neighborhood is thus like a local weighted least squares fit.

Robust versions of local regression smoothers, which downweight large residuals, have also been proposed (see, e.g., Cleveland (1994) and Cleveland, Grosse, and Shyu (1992)) for the setting where the random errors have a symmetric distribution. Computational methods for local regression smoothers are presented in Cleveland and Grosse (1991). Approaches to statistical inferences for $\mu(x_0)$, as well as diagnostic checks associated with local regression smoothers, are discussed in Cleveland, Grosse, and Shyu (1992).

Running line, kernel regression, and local regression smoothers are discussed in more detail in Chapter 14.

Spline Regression Smoothers. A *spline* is a curve pieced together from a number of individually constructed curve/line segments; that is, a spline is simply a piecewise polynomial. (Smith (1979) and Eubank (1999) provided nice discussions of this general concept.) When each segment of a spline contains only linear terms, it is called a *linear spline*.

The application of splines to regression problems in a general sense is discussed in Wegman and Wright (1983), where they refer to splines associated with a regression model as *regression splines*. The simplest of these regression spline smoothers are those associated with linear splines. The junctures where these lines are pieced together are known as *knots.* When the positions of these knots are known a priori, the use of linear regression spline smoothers is relatively straightforward. However, when the positions of the knots are also unknown, the problem becomes a good deal more complicated. Applications of higher order polynomial splines (in particular, quadratic and cubic splines) are discussed in Eubank (1999).

A different approach to the use of splines in regression problems is associated with the development of *smoothing splines.* For these procedures, the regression smooth results from minimization of a sum of squares augmented by a smoothing term related to the order of the desired smoothing spline. For further information on smoothing splines, the reader is referred to Eubank (1999) and Wahba (1990).

Wavelet Smoother. This smoother represents the observed data with a set of basis functions and their corresponding coefficients. Wavelet functions are commonly used as a basis because they are able to model data sampled from very general relations between Y and x. The wavelet estimate is not linear. The coefficients are modified nonlinearly by thresholding rules which generally set some of the coefficients to 0 and shrink the remaining toward 0. The estimate of μ is found with these modified coefficients, not the original values.

Donoho and Johnstone (1994, 1995) created popular threshold methods for wavelets called *VisuShrink and SureShrink*. Under the assumptions of normal errors, they showed that the asymptotic rates of convergence for these wavelet smoothers are optimal or near optimal. New methods of thresholding with improved estimation properties have been derived. For example, Cai (1999) collapsed groups of coefficients together in order to attain optimal convergence rates with improved visual smoothness of the estimate. Others have extended wavelet smoothers to variable designs for the x_i (see, e.g., Kovac

and Silverman (2000)) and non-normality of errors (e.g., Nason (1996)). Chapter 13 discusses wavelet smoothing in greater detail.

General Discussion. As mentioned previously, all the nonparametric regression procedures discussed in this section are considerably more computationally intensive than the material presented elsewhere in the text. As a result, computer software is essential for the implementation of these nonparametric regression smoothers. Such software is available in R.

Finally, we note that in determining which of these approaches to nonparametric regression is most appropriate for a given problem, the decision invariably comes down to the relative importance of minimizing bias versus minimizing variance. All these smoothers produce biased estimators for the regression function $\mu(x)$ so that the desired trade-off between the sizes of the variance and bias (along with computational capabilities, of course) often strongly influences the choice of a particular nonparametric regression smoother.

9.8 EFFICIENCIES OF REGRESSION PROCEDURES

The asymptotic relative efficiencies of the Theil procedures of Sections 9.1–9.3 with respect to their normal theory counterparts based on the least squares estimator of β have been found by Sen (1968) to be given by the expression

$$e_F = \varepsilon^2 \left[12\sigma_F^2 \left\{ \int_{-\infty}^{\infty} f^2(u)du \right\}^2 \right] = \varepsilon^2 e_F^*, \tag{9.65}$$

where σ_F^2 is the variance of the common underlying (continuous) distribution $F(\cdot)$ for the random variables $e_1, \ldots, e_n$ in (9.1), $f(\cdot)$ is the probability density function corresponding to F, and ϵ^2 is the limiting value (n tending to infinity) of ϵ_n^2, where ϵ_n is the product moment correlation coefficient between $(x_1, \ldots, x_n)$ and $(1, \ldots, n)$ as given in expression (6.2) of Sen (1968). The parameter $\int_{-\infty}^{\infty} f^2(u)du$ is the area under the curve associated with $f^2(\cdot)$, the square of the common probability density function. We note that the expression e_F (9.65) is simply ϵ^2 times the corresponding Pitman efficiencies (e_F^*) in the one-sample, two-sample, and k-sample location settings (See Sections 3.11, 4.5, and 6.10).

We note that ϵ_n clearly depends on the design configuration for the values $(x_1, \ldots, x_n)$ of the independent variable. An important special case where $\epsilon_n = 1$ is the equally spaced, no-replications design, where $x_i = x_1 + (i-1)a$, for some $a > 0$ and $i = 1, \ldots, n$. When $\epsilon^2 = 1$, values of e_F (9.65) correspond to e_F^* and can be obtained from display (3.116).

The asymptotic relative efficiency under a sequence of near alternatives of the Sen–Adichie parallelism test based on V (9.45) with respect to the corresponding normal theory procedure based on least squares estimators was found to be e_F^* (9.65) by Sen (1969). The asymptotic relative efficiency under a sequence of contiguous alternatives of the Jaeckel-Hettmansperger-McKean rank-based multiple linear regression test based on HM (9.57) with respect to the corresponding least squares competitor test procedure was found by McKean and Hettmansperger (1976) to be e_F^* (9.65) as well. Once again, the values of e_F^* can be found in display (3.116).