8

Alternate Assumptions for Regression

8.1 Overview

The inference procedures for simple and multiple linear regression we have discussed so far are based on assumptions (A) or (B) given in Chapters 3 and 4. In those chapters we also discussed various diagnostic tools for examining the validity of these assumptions. In situations where we know or suspect that one or more of the required assumptions are not satisfied, it is useful to have alternative valid approaches. In this chapter we discuss alternate sets of assumptions under which valid inferences are possible even if assumptions (A) and/or assumptions (B) do not hold. Section 8.2 introduces procedures that can be used when the assumption of homogeneity of subpopulation standard deviations (variances) is not satisfied. When the Gaussian assumption does not hold for subpopulations, the assumptions discussed in Section 8.3 for the case of straight line regression and the corresponding inference procedures may apply. Our presentation of the topics in this chapter is necessarily limited in scope and hence is only introductory. Sections 8.2 and 8.3 in the laboratory manuals discuss the use of the computer to perform the calculations needed in this chapter.

8.2 Straight Line Regression with Unequal Subpopulation Standard Deviations

The procedures for regression analysis discussed in Chapters 3 and 4 were based on the assumption of *homogeneity of standard deviations*, sometimes referred to as *homogeneity of variances*; i.e., it was assumed that the standard deviations, or equivalently, the variances, of all the subpopulations are the same. This assumption may never be satisfied *exactly* in practical applications, but it is often a reasonable approximation. However, there are situations when the assumption of equal standard deviations is not appropriate. For such situations, alternate assumptions and procedures are needed, and in this section we discuss one of the alternate procedures for straight line regression.

Recall that $\sigma_Y(x)$ represents the standard deviation of the subpopulation of Y values corresponding to $X = x$. When all the subpopulation standard deviations are equal, we have used the symbol $\sigma_{Y|X}$ (σ for short) for their common value. In this section we consider the situation where the $\sigma_Y(x)$ are not all equal, but where *the relative values of the standard deviations of the different subpopulations are known.* This amounts to the assumption that

$$\sigma_Y(x) = \sigma_0 g(x) \tag{8.2.1}$$

where σ_0 is an *unknown* constant and $g(x)$ is a *known* function of x. To illustrate, let us suppose that $\sigma_Y(x) = \sigma_0\sqrt{x}$. Then $\sigma_Y(4) = \sigma_0\sqrt{4} = 2\sigma_0$, and $\sigma_Y(9) = \sigma_0\sqrt{9} = 3\sigma_0$ so that $\sigma_Y(9)/\sigma_Y(4) = 3\sigma_0/2\sigma_0 = 1.5$, which is a known constant. Thus the standard deviation of the subpopulation corresponding to $X = 9$ is 1.5 times as big as the standard deviation of the subpopulation corresponding to $X = 4$.

If the function $g(x)$ is equal to 1 for all allowable values of x, then the subpopulation standard deviations are all the same and are equal to σ_0. In that case σ_0 will be equal to $\sigma_{Y|X}$, the common standard deviation of all the subpopulations.

In this section we discuss inference procedures for straight line regression when the assumptions in Box 8.2.1 hold.

BOX 8.2.1 Weighted Regression Assumptions for Straight Line Regression

Notation A two-variable population $\{(Y, X)\}$ is the study population under investigation.

(Population) Assumption 1 The mean $\mu_Y(x)$ of the subpopulation of Y values for specified x is

$$\mu_Y(x) = \beta_0 + \beta_1 x \tag{8.2.2}$$

where β_0 and β_1 are unknown parameters.

(Population) Assumption 2 The standard deviation of the Y values in the subpopulation determined by $X = x$ is $\sigma_Y(x)$ where

$$\sigma_Y(x) = \sigma_0 g(x) \tag{8.2.3}$$

and σ_0 is an *unknown* positive constant; $g(x)$ is a *known* function of x such that $g(x) > 0$ for all allowable values of x.

(Population) Assumption 3 Each subpopulation of Y values, determined by specified values of X, is Gaussian.

(Sample) Assumption 4 A sample (of size n) is selected either by simple random sampling or by preselecting X values.

(Sample) Assumption 5 All sample values y_i, x_i for $i = 1, \dots, n$ are observed without error.

Note Observe that *weighted regression assumptions* in Box 8.2.1 are identical to assumptions (A) for regression *with the exception of population assumption 2*. Under weighted regression assumptions the standard deviations of subpopulations are allowed to be different, but the ratio of the standard deviation of any one subpopulation, say with $X = x_1$, relative to any other subpopulation, say with $X = x_2$, is assumed to be known and equal to $g(x_1)/g(x_2)$; this is actually population assumption 2 in Box 8.2.1. Under assumptions (A) and (B) of Chapters 3 and 4, all the subpopulation standard deviations are the same, so weighted regression reduces to ordinary regression.

The Method of Weighted Least Squares

In Chapter 3, we estimated the parameters β_0, β_1 by the method of least squares; viz., we found the quantities $\hat{\beta}_0, \hat{\beta}_1$ that minimize the sum of squares of prediction errors given by

$$\sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_i)]^2 \tag{8.2.4}$$

The estimates $\hat{\beta}_0, \hat{\beta}_1$ we obtained were called *least squares estimates*. The quantities y_i, x_i are the data values corresponding to sample item i.

When subpopulation standard deviations are unequal but weighted regression assumptions are satisfied, the estimates of β_0, β_1 given in Chapter 3 are not the best estimates, and the method of least squares needs to be modified. The prediction errors are first *weighted* by dividing each prediction error by a factor proportional to the corresponding subpopulation standard deviation. This ensures that the method of estimation will give more weight to observations from subpopulations with smaller standard deviations because these observations are more reliable, and less weight will be given to observations from subpopulations with larger standard deviations because these observations are less reliable. The weighted prediction error corresponding to sample item i, when $\beta_0 + \beta_1 x_i$ is used to predict y_i, is denoted by $e_i^{(w)}$, and it is given by

$$e_i^{(w)} = \frac{y_i - (\beta_0 + \beta_1 x_i)}{g(x_i)} \tag{8.2.5}$$

Note that the denominator of the right-hand side of (8.2.5) is a quantity that is proportional to the standard deviation of the subpopulation of Y values with $X = x_i$. Thus the weighted prediction error corresponding to the ith sample observation, when $\beta_0 + \beta_1 x_i$ is used to predict y_i, is obtained by weighting the prediction error $y_i - (\beta_0 + \beta_1 x_i)$ by the quantity $1/g(x_i)$, i.e., by a quantity that is inversely proportional to the corresponding subpopulation standard deviation.

The best estimates of β_0, β_1 under weighted regression assumptions are obtained by minimizing the *sum of squares*

$$\sum_{i=1}^{n}\{e_i^{(w)}\}^2$$

of weighted prediction errors. The resulting estimates of β_0, β_1 are called **weighted least squares (WLS) estimates** of β_0, β_1, and they are denoted by $\hat{\beta}_0^{(w)}, \hat{\beta}_1^{(w)}$. The corresponding minimum value of the sum of squares of weighted prediction errors is called **weighted sum of squared errors**, and it is denoted by $WSSE(X)$. We define $\hat{e}_i^{(w)}$ by the equation

$$\hat{e}_i^{(w)} = \frac{y_i - (\hat{\beta}_0^{(w)} + \hat{\beta}_1^{(w)} x_i)}{g(x_i)} \tag{8.2.6}$$

and call this quantity the *weighted residual* for sample item i. We then have

$$WSSE(X) = \sum_{i=1}^{n} \{\hat{e}_i^{(w)}\}^2 \tag{8.2.7}$$

$$= \sum_{i=1}^{n} \left[\frac{y_i - (\hat{\beta}_0^{(w)} + \hat{\beta}_1^{(w)} x_i)}{g(x_i)} \right]^2 \tag{8.2.8}$$

$$= \sum_{i=1}^{n} w_i \left[y_i - (\hat{\beta}_0^{(w)} + \hat{\beta}_1^{(w)} x_i) \right]^2 \tag{8.2.9}$$

where

$$w_i = \left[\frac{1}{g(x_i)} \right]^2 \tag{8.2.10}$$

The quantities w_i are called *weights*.

To distinguish between weighted least squares estimates of β_0, β_1 discussed in this section and the estimates of β_0, β_1 discussed in Chapter 3, we refer to the estimates discussed in Chapter 3 as *unweighted least squares estimates* or *ordinary least squares (OLS) estimates* of β_0, β_1. Note that if the subpopulation standard deviations are all equal (i.e., if $g(x) = 1$), then the WLS estimates of β_0, β_1, given in (8.2.11), are the same as the OLS estimates given in (4.4.8).

Point Estimation and Confidence Intervals

We now discuss point and confidence interval estimation for straight line regression when weighted regression assumptions hold. As usual we let $\beta = [\beta_0, \beta_1]^T$. The weighted least squares estimate of β is denoted by $\hat{\beta}^{(w)} = [\hat{\beta}_0^{(w)}, \hat{\beta}_1^{(w)}]^T$. It can be proved that

$$\hat{\beta}^{(w)} = (X^T W X)^{-1} X^T W y \tag{8.2.11}$$

where

$$X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_i \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \qquad y = \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix} \tag{8.2.12}$$

and

$$W = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{bmatrix} \tag{8.2.13}$$

Note that the diagonal elements of W are the weights $w_1, w_2, \ldots, w_n$, where $w_i = [1/g(x_i)]^2$ and the off-diagonal elements of W are all zero.

An estimate of the standard deviation $\sigma_Y(x)$ of the subpopulation corresponding to $X = x$ is

$$\hat{\sigma}_Y(x) = g(x)\hat{\sigma}_0 \tag{8.2.14}$$

where

$$\hat{\sigma}_0 = \sqrt{WMSE(X)} \tag{8.2.15}$$

is an estimate of σ_0 and

$$WMSE(X) = \frac{WSSE(X)}{(n-2)} \tag{8.2.16}$$

The quantity $WSSE(X)$ given in (8.2.7)–(8.2.9) is called the *weighted sum of squared errors*, and it has $n - 2$ degrees of freedom associated with it. The corresponding quantity $WMSE(X)$ in (8.2.16) is called the weighted mean squared error.

To compute confidence intervals for

$$\beta_0, \beta_1 \tag{8.2.17}$$

$$Y(x), \mu_Y(x) \tag{8.2.18}$$

and

$$a^T\beta = a_0\beta_0 + a_1\beta_1 \tag{8.2.19}$$

we need their point estimates and corresponding standard errors. Point estimates of β_0, β_1 are obtained from (8.2.11). The corresponding standard errors are given by

$$SE(\hat{\beta}_{i-1}^{(w)}) = \sqrt{WMSE(X)c_{ii}^{(w)}} \tag{8.2.20}$$

$$= \hat{\sigma}_0\sqrt{c_{ii}^{(w)}} \qquad \text{for } i = 1, 2 \tag{8.2.21}$$

where $c_{ii}^{(w)}$ is the ith diagonal element of the 2 by 2 matrix $C^{(w)}$, given by

$$C^{(w)} = \begin{bmatrix} c_{11}^{(w)} & c_{12}^{(w)} \\ c_{21}^{(w)} & c_{22}^{(w)} \end{bmatrix} \tag{8.2.22}$$

$$= (X^T W X)^{-1} \tag{8.2.23}$$

Point estimates and standard errors for the quantities in (8.2.18) and (8.2.19) are given in (8.2.24)–(8.2.29).

$$\hat{Y}^{(w)}(x) = \hat{\beta}_0^{(w)} + \hat{\beta}_1^{(w)} x \tag{8.2.24}$$

$$SE(\hat{Y}^{(w)}(x)) = \hat{\sigma}_0 \sqrt{[g(x)]^2 + \boldsymbol{x}^T C^{(w)} \boldsymbol{x}} \tag{8.2.25}$$

$$\hat{\mu}_Y^{(w)}(x) = \hat{\beta}_0^{(w)} + \hat{\beta}_1^{(w)} x \tag{8.2.26}$$

$$SE(\hat{\mu}_Y^{(w)}(x)) = \hat{\sigma}_0 \sqrt{\boldsymbol{x}^T C^{(w)} \boldsymbol{x}} \tag{8.2.27}$$

$$\boldsymbol{a}^T \hat{\boldsymbol{\beta}}^{(w)} = a_0 \hat{\beta}_0^{(w)} + a_1 \hat{\beta}_1^{(w)} \tag{8.2.28}$$

$$SE(\boldsymbol{a}^T \hat{\boldsymbol{\beta}}^{(w)}) = \hat{\sigma}_0 \sqrt{\boldsymbol{a}^T C^{(w)} \boldsymbol{a}} \tag{8.2.29}$$

where $\boldsymbol{x} = [1, x]^T$ and $\boldsymbol{a} = [a_0, a_1]^T$ are specified. As usual, we use $SE(\hat{Y}^{(w)}(x))$ to denote $SE(\hat{Y}^{(w)}(x) - Y(x))$.

Confidence intervals for the quantities in (8.2.17)–(8.2.19) can be computed using (4.6.1) with the point estimates and standard errors given in (8.2.11), (8.2.21), and (8.2.24)–(8.2.29). Confidence intervals for σ_0 have the same form as those in (4.6.13) and (4.6.14), with $\sigma_{Y|X}$ replaced by σ_0, $\hat{\sigma}_{Y|X}$ by $\hat{\sigma}_0$, and $SSE(X)$ by $WSSE(X)$. Confidence intervals and tests for the subpopulation standard deviation $\sigma_Y(x) = \sigma_0 g(x)$ can be obtained from those for σ_0 because $g(x)$ is a known multiplier. Example 8.2.1 illustrates the computations.

EXAMPLE 8.2.1

A study was conducted to understand the relationship, if any, between Y, the levels of carbon monoxide (CO) in the air (measured in parts per million) and X, the number (in thousands) of automobiles in various U.S. cities that do not have an ongoing clean air program. Thirteen cities were chosen using simple random sampling from the study population (which is also the target population in this problem), which consists of all cities in the United States that have a population of more than 50,000 and do not have an ongoing clean air program. Data for these thirteen cities are given in Table 8.2.1 and are also stored in the file **carbmon.dat** on the data disk. It is known that the subpopulation standard deviations $\sigma_Y(X)$ are not all the same, but the investigator expects the weighted regression assumptions in Box 8.2.1 to hold with $\mu_Y(x) = \beta_0 + \beta_1 x$ and $\sigma_Y(x) = \sigma_0 g(x)$ for $100 \le x \le 1200$, where σ_0 is an unknown constant and $g(x) = \sqrt{x}$. To illustrate the formulas of this section, we compute point estimates and confidence intervals for β_0 and β_1 using the method of weighted least squares.

TABLE 8.2.1
Carbon Monoxide Data

City	CO Y (in ppm)	Number of Automobiles X (in thousands)
1	5817	873
2	1063	109
3	2616	398
4	2018	353
5	3147	506
6	7210	1026
7	4339	862
8	5153	742
9	4450	786
10	5591	896
11	2747	377
12	3712	720
13	2354	655

The matrices y, X, and W are (note $w_i = [1/g(x_i)]^2 = 1/x_i$):

$$y = \begin{bmatrix} 5817 \\ 1063 \\ 2616 \\ 2018 \\ 3147 \\ 7210 \\ 4339 \\ 5153 \\ 4450 \\ 5591 \\ 2747 \\ 3712 \\ 2354 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 873 \\ 1 & 109 \\ 1 & 398 \\ 1 & 353 \\ 1 & 506 \\ 1 & 1026 \\ 1 & 862 \\ 1 & 742 \\ 1 & 786 \\ 1 & 896 \\ 1 & 377 \\ 1 & 720 \\ 1 & 655 \end{bmatrix} \tag{8.2.30}$$

$$W = \begin{bmatrix} \frac{1}{873}, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0 \\ 0, \frac{1}{109}, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0 \\ 0, 0, \frac{1}{398}, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0 \\ 0, 0, 0, \frac{1}{353}, 0, 0, 0, 0, 0, 0, 0, 0, 0 \\ 0, 0, 0, 0, \frac{1}{506}, 0, 0, 0, 0, 0, 0, 0, 0 \\ 0, 0, 0, 0, 0, \frac{1}{1026}, 0, 0, 0, 0, 0, 0, 0 \\ 0, 0, 0, 0, 0, 0, \frac{1}{862}, 0, 0, 0, 0, 0, 0 \\ 0, 0, 0, 0, 0, 0, 0, \frac{1}{742}, 0, 0, 0, 0, 0 \\ 0, 0, 0, 0, 0, 0, 0, 0, \frac{1}{786}, 0, 0, 0, 0 \\ 0, 0, 0, 0, 0, 0, 0, 0, 0, \frac{1}{896}, 0, 0, 0 \\ 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, \frac{1}{377}, 0, 0 \\ 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, \frac{1}{720}, 0 \\ 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, \frac{1}{655} \end{bmatrix} \tag{8.2.31}$$

We get

$$\hat{\beta}^{(w)} = (X^T W X)^{-1} X^T W y = \begin{bmatrix} 371.620 \\ 5.466 \end{bmatrix} \tag{8.2.32}$$

Thus $\hat{\beta}_0^{(w)} = 371.620$ and $\hat{\beta}_1^{(w)} = 5.466$.

To calculate the standard errors of $\hat{\beta}_0^{(w)}$ and $\hat{\beta}_1^{(w)}$ we need $C^{(w)} = (X^T W X)^{-1}$ and $\hat{\sigma}_0$. First we obtain

$$C^{(w)} = (X^T W X)^{-1} = \begin{bmatrix} 114.596 & -0.179 \\ -0.179 & 0.000401 \end{bmatrix} \tag{8.2.33}$$

Next we calculate $WSSE(X)$, the weighted sum of squared errors, using the formula in (8.2.9). We get

$$\begin{aligned} WSSE(X) &= \frac{1}{873}[5817 - (371.620 + 5.466 \times 873)]^2 \\ &\quad + \frac{1}{109}[1063 - (371.620 + 5.466 \times 109)]^2 \\ &\quad + \cdots \\ &\quad + \cdots \\ &\quad + \frac{1}{655}[2354 - (371.620 + 5.466 \times 655)]^2 \\ &= 8498.69 \end{aligned}$$

From this we obtain

$$WMSE(X) = \frac{WSSE(X)}{(n-2)} = \frac{8498.69}{11} = 772.61 \text{ (to two decimal places)} \tag{8.2.34}$$

and consequently

$$\hat{\sigma}_0 = \sqrt{WMSE(X)} = \sqrt{772.61} = 27.8 \text{ (rounded to one decimal place)} \tag{8.2.35}$$

Using (8.2.20) we get $SE(\hat{\beta}_0^{(w)}) = 297.6$ and $SE(\hat{\beta}_1^{(w)}) = 0.5569$. Hence a two-sided 90% confidence interval for β_0 is given by the confidence statement

$$
\begin{aligned}
&C[\hat{\beta}_0^{(w)} - t_{0.95:11}SE(\hat{\beta}_0^{(w)}) \le \beta_0 \le \hat{\beta}_0^{(w)} + t_{0.95:11}SE(\hat{\beta}_0^{(w)})] \\
&\quad = C[371.62 - 1.796 \times 297.6 \le \beta_0 \le 371.62 + 1.796 \times 297.6] \\
&\quad = C[-162.9 \le \beta_0 \le 906.1] = 0.90
\end{aligned}
$$

Likewise, a two-sided 90% confidence interval for β_1 is given by the confidence statement

$$
\begin{aligned}
&C[\hat{\beta}_1^{(w)} - t_{0.95:11}SE(\hat{\beta}_1^{(w)}) \le \beta_1 \le \hat{\beta}_1^{(w)} + t_{0.95:11}SE(\hat{\beta}_1^{(w)})] \\
&\quad = C[5.466 - 1.796 \times 0.5569 \le \beta_1 \le 5.466 + 1.796 \times 0.5569] \\
&\quad = C[4.466 \le \beta_1 \le 6.466] = 0.90
\end{aligned}
$$

For the purpose of illustration we calculate $\hat{Y}(x)$, $\hat{\sigma}_Y(x)$, and $SE(\hat{Y}(x))$ for $x = 300$.

$$\hat{Y}(300) = \hat{\beta}_0^{(w)} + \hat{\beta}_1^{(w)}300 = 2011.48 \tag{8.2.36}$$

$$\hat{\sigma}_Y(300) = \hat{\sigma}_0 g(300) = (27.8)(\sqrt{300}) = 481.51 \tag{8.2.37}$$

$$SE(\hat{Y}(300)) = \hat{\sigma}_0\sqrt{[g(300)]^2 + x^T C^{(w)} x} \tag{8.2.38}$$

$$= 27.8\sqrt{300 + [1 \quad 300]\,C^{(w)}\begin{bmatrix} 1 \\ 300 \end{bmatrix}} = 514.9$$

With this information we can use (4.6.1) and calculate confidence intervals for $Y(300)$.

Finally we illustrate how to obtain a two-sided 80% confidence interval for $\sigma_Y(300)$. This is done by first computing a two-sided 80% confidence interval for σ_0 using (4.6.13) with $WSSE(X)$ in place of $SSE(X)$. We get

$$
\begin{aligned}
&C\left[\sqrt{\frac{WSSE(X)}{\chi^2_{1-\alpha/2:n-2}}} \le \sigma_0 \le \sqrt{\frac{WSSE(X)}{\chi^2_{\alpha/2:n-2}}}\right] \\
&= C\left[\sqrt{\frac{8498.69}{\chi^2_{0.9:11}}} \le \sigma_0 \le \sqrt{\frac{8498.69}{\chi^2_{0.1:11}}}\right] \\
&= C\left[\sqrt{\frac{8498.69}{17.275}} \le \sigma_0 \le \sqrt{\frac{8498.69}{5.578}}\right] \\
&= C\left[22.18 \le \sigma_0 \le 39.03\right] = 0.80
\end{aligned}
$$

Hence, by multiplying each term in the preceding confidence statement by $g(300) = \sqrt{300}$, we get

$$C[22.18\ g(300) \le \sigma_0 g(300) \le 39.03\ g(300)] = 0.80$$

i.e.,

$$C[384.17 \le \sigma_Y(300) \le 676.02] = 0.80 \quad \blacksquare$$

Most computer packages will perform a weighted least squares regression analysis if the user supplies the weights. In Section 8.2 of the laboratory manuals we show how to use the computer to perform the calculations needed for weighted regression discussed in this section.

Exhibit 8.2.1, which is obtained using MINITAB, is a typical output from a weighted regression program. The output is very similar to the output from an ordinary (unweighted) regression analysis. The data used are from Example 8.2.1.

Note the weights w_i for performing a weighted least squares regression are $w_i = [1/g(x_i)]^2 = 1/x_i$ for this problem.

The values of $\hat{\beta}_0^{(w)}$ and $\hat{\beta}_1^{(w)}$ are given in (8.2.39) and (8.2.40), respectively. Compare these with the values in (8.2.32). $WSSE(X) = 8499$ and $WMSE(X) = 773$ are given in (8.2.41) under the headings `SS` and `MS`, respectively. Thus, $\hat{\sigma}_0$ may be obtained as $\hat{\sigma}_0 = \sqrt{WMSE(X)} = \sqrt{773} = 27.8$, the same as in (8.2.35), to within rounding error. The matrix $C^{(w)}$ is given in (8.2.42).

EXHIBIT 8.2.1
MINITAB Output for Example 8.2.1

Row	CO	cars	weights
1	5817	873	0.0011455
2	1063	109	0.0091743
3	2616	398	0.0025126
4	2018	353	0.0028329
5	3147	506	0.0019763
6	7210	1026	0.0009747
7	4339	862	0.0011601
8	5153	742	0.0013477
9	4450	786	0.0012723
10	5591	896	0.0011161
11	2747	377	0.0026525
12	3712	720	0.0013889
13	2354	655	0.0015267

```
The regression equation is
CO = 372 + 5.47 cars
```

Predictor	Coef	Stdev	t-ratio	p	
Constant	371.6	297.6	1.25	0.238	(8.2.39)
cars	5.4662	0.5569	9.82	0.000	(8.2.40)

EXHIBIT 8.2.1
(Continued)

```
Analysis of Variance

SOURCE        DF        SS        MS       F        p
Regression     1     74445     74445   96.36    0.000
Error         11      8499       773                          (8.2.41)
Total         12     82944

The weighted C matrix is

  114.5957    -0.1794
   -0.1794     0.000401                                       (8.4.42)
```

Problems 8.2

8.2.1 Consider Problem 3.5.1 where an investigator is studying the association between sulfur dioxide (SO_2) concentrations in a national park and the rate of emission of SO_2 by a coal burning power plant 25 miles away. A certain fraction of the emitted SO_2 will be transported by winds to the national park. At the national park, there is always a certain amount of background SO_2 that is not emitted by the power plant. The SO_2 emissions (X, in tons/hour) by the power plant and the SO_2 concentrations at the national park (Y, in micrograms/cubic meter, or mg/m^3) were recorded at various randomly selected times during a particular year. The data are given in Table 8.2.2 and are also stored in the file named **so2.dat** on the data disk. Suppose that weighted regression assumptions in Box 8.2.1 are valid with

$$\mu_Y(x) = \beta_0 + \beta_1 x \tag{8.2.43}$$

and $\sigma_Y(x) = \sigma_0 g(x) = \sigma_0 x$ where β_0, β_1, and σ_0 are unknown constants. So $g(x) = x$ and the weights are $w_i = [1/g(x_i)]^2 = 1/x_i^2$. We use weighted regression to obtain point estimates and confidence intervals for the unknown parameters.

The computer output in Exhibit 8.2.2 lists the data along with the weights and the results from a weighted regression analysis. Two weights, those for sample items 2 and 5, have not been computed.

a Compute the weights for items 2 and 5.

b Verify that the weights are all correct.

c What are the weighted least squares estimates for β_0 and β_1?

d Compare the estimates in (c) with the unweighted estimates obtained in Problem 3.5.2.

e State what an appropriate target population might be for this problem. What is the study population?

TABLE 8.2.2
SO_2 Data

Time	Y (micrograms/cubic meter)	X (tons/hour)
1	5.21	1.92
2	7.36	3.92
3	16.26	6.80
4	10.10	6.32
5	5.80	2.00
6	8.06	4.32
7	4.76	2.40
8	6.93	2.96
9	9.36	3.52
10	10.90	4.24
11	12.48	5.12
12	11.70	5.84
13	7.44	3.60
14	6.99	2.80

EXHIBIT 8.2.2
MINITAB Output for Problem 8.2.1

```
Row      Y       X      weights

  1    5.21    1.92    0.271267
  2    7.36    3.92    ********
  3   16.26    6.80    0.021626
  4   10.10    6.32    0.025036
  5    5.80    2.00    ********
  6    8.06    4.32    0.053584
  7    4.76    2.40    0.173611
  8    6.93    2.96    0.114134
  9    9.36    3.52    0.080708
 10   10.90    4.24    0.055625
 11   12.48    5.12    0.038147
 12   11.70    5.84    0.029321
 13    7.44    3.60    0.077160
 14    6.99    2.80    0.127551
```

EXHIBIT 8.2.2
(Continued)

```
The regression equation is
Y = 1.72 + 1.78 X

Predictor        Coef       Stdev    t-ratio        p
Constant       1.7214      0.7683       2.24    0.045
X              1.7762      0.2415       7.36    0.000

Analysis of Variance

SOURCE        DF          SS          MS        F        p
Regression     1      6.0298      6.0298    54.10    0.000
Error         12      1.3375      0.1115
Total         13      7.3672

The weighted C matrix is

  5.29681   -1.54690
 -1.54690    0.52319
```

8.2.2 In Problem 8.2.1 estimate the mean and the standard deviation of the SO_2 concentrations at the park associated with an emission rate of 3.0 tons/hour at the power plant.

8.2.3 In Problem 8.2.1 what is the population parameter that represents the difference between the average SO_2 concentration at the park associated with a power plant emission rate of 5.0 tons/hour, and that associated with a power plant emission rate of 2.5 tons/hour?

8.2.4 Estimate the difference in Problem 8.2.3 and compute a two-sided 95% confidence interval for the difference.

8.2.5 Discuss whether or not claims can be made to the effect that the SO_2 emissions at the power plant cause the SO_2 concentrations at the national park to increase. In particular, can we conclude, on the basis of these data, that the SO_2 concentrations at the national park will decrease if the power plant is shut down?

8.2.6 Compute a 90% two-sided confidence interval for σ_0.

8.2.7 Compute a 90% two-sided confidence interval for $\sigma_Y(4.00)$.

8.3 Straight Line Regression—Theil's Method

Assumptions (A), (B), and the weighted regression assumptions all require that each subpopulation of Y values is Gaussian. In some problems the investigator may know that the subpopulations are not Guassian (not even approximately), or a residual analysis of the sample data may cast doubt on this assumption. In such cases it is useful to have alternative, valid inference procedures available. In this section we discuss one such alternative for straight line regression, called *Theil's method* because it was first proposed by H. Theil [34].

We call the assumptions underlying Theil's method for straight line regression non-Gaussian assumptions, and they are given in Box 8.3.1.

BOX 8.3.1 Non-Gaussian Assumptions for Straight Line Regression

Notation Let $\{(Y, X)\}$ be a two-variable study population.

(Population) Assumption 1 For each distinct value x of X in the population, the mean $\mu_Y(x)$ of the corresponding subpopulation of Y values is given by

$$\mu_Y(x) = \beta_0 + \beta_1 x \qquad a \le x \le b \tag{8.3.1}$$

where β_0 and β_1 are unknown parameters.

(Population) Assumption 2 Each subpopulation of Y values, determined by the distinct values of X, is symmetric and continuous.

(Sample) Assumption 3 The sample of size n is selected either by simple random sampling or by preselecting the X values.

(Sample) Assumption 4 The values of y_i and x_i for $i = 1, 2, \ldots, n$ are observed without error.

We make a few comments about these assumptions.

1 The term *continuous subpopulation* in population assumption 2 means that the response variable Y is a continuous variable such as weight, height, time, etc. and not a discrete variable such as counts of numbers of people, homes, days, etc. In particular, when Y is a continuous variable, no two of its values will be the same if they are measured sufficiently precisely.

2 In population assumption 2, the requirement that the subpopulation of Y values be symmetric for each X value means that for every value of Y in the subpopulation that is d units below the mean $\mu_Y(x) = \beta_0 + \beta_1 x$, there is a corresponding Y value d units above the mean. Subpopulations of Y values for different values of X may be different with respect to their mean values, or standard deviations, or other characteristics, but they must all be symmetric. For example they can all be Gaussian (which is symmetric) with different means and different standard deviations.

3 The subpopulation of Y values determined by the X values need not be Gaussian.

4 *If the subpopulations of Y values happen to be Gaussian for each X, their standard deviations need not all be the same, nor do their relative magnitudes need to be known as in weighted regression assumptions in Box 8.2.1.*

Point Estimation

We now explain the procedure for estimating θ, a linear combination of β_0 and β_1, given by

$$\theta = a_0\beta_0 + a_1\beta_1 \tag{8.3.2}$$

where a_0 and a_1 are specified constants. Note that β_0 is obtained from θ by setting $a_0 = 1$ and $a_1 = 0$ in (8.3.2), whereas β_1 is obtained from θ by setting $a_0 = 0$ and $a_1 = 1$. Also $\mu_Y(x)$ is obtained from θ by letting $a_0 = 1$ and $a_1 = x$, and $\mu_Y(x_1) - \mu_Y(x_2)$ is obtained from θ by setting $a_0 = 0$ and $a_1 = x_1 - x_2$.

Let $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$ be a sample of size n arranged according to increasing values of x; i.e., $x_1 < x_2 < \cdots < x_n$. If several y values are available for a given x value, we let y_i be their mean so we can assume that the x_i's are distinct. If n is an odd number, say $n = 2m + 1$, then discard the middle observation so there are now $2m$ observations (y_i, x_i) for $i = 1, \ldots, 2m$. Of course if n is even, no observation is discarded and $n = 2m$. The observations are arranged as in Table 8.3.1. Remember that $x_1 < x_2 < \cdots < x_{2m}$ and that no two x values are the same. From Table 8.3.1 we compute the quantities z, w, u, v, and t, which are exhibited in Table 8.3.2.

TABLE 8.3.1

Column 1	Column 2	Column 3	Column 4
y_1	x_1	y_{m+1}	x_{m+1}
y_2	x_2	y_{m+2}	x_{m+2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_m	x_m	y_{2m}	x_{2m}

TABLE 8.3.2

z	w	u	v	t
$z_1 = y_{m+1} - y_1$	$w_1 = x_{m+1} - x_1$	$u_1 = y_1x_{m+1}$	$v_1 = y_{m+1}x_1$	$t_1 = u_1 - v_1$
$z_2 = y_{m+2} - y_2$	$w_2 = x_{m+2} - x_2$	$u_2 = y_2x_{m+2}$	$v_2 = y_{m+2}x_2$	$t_2 = u_2 - v_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$z_m = y_{2m} - y_m$	$w_m = x_{2m} - x_m$	$u_m = y_mx_{2m}$	$v_m = y_{2m}x_m$	$t_m = u_m - v_m$

Compute q_i^* where

$$q_i^* = (a_0 t_i + a_1 z_i)/w_i \tag{8.3.3}$$

for $i = 1, \ldots, m$. Arrange the q_i^* in increasing order, and denote them by $q_1, q_2, \ldots, q_m$ where $q_1 < q_2 < \cdots < q_m$. If m is an odd number (i.e., $m = 2k + 1$), then the middle number q_{k+1} is the estimate of $a_0\beta_0 + a_1\beta_1$. If m is an even number (i.e., $m = 2k$), then $(q_k + q_{k+1})/2$, the average of the two middle numbers, is the estimate of $a_0\beta_0 + a_1\beta_1$.

Confidence Intervals

A confidence interval for $a_0\beta_0 + a_1\beta_1$ may not be available with confidence coefficient exactly equal to a specified value $1 - \alpha$, so we find confidence intervals with confidence coefficients as close to $1 - \alpha$ as the procedure allows. To do this we follow the instructions in Box 8.3.2.

BOX 8.3.2

1 Let $m = \frac{n}{2}$ if n is even and $m = \frac{n-1}{2}$ if n is odd.

2 For this value of m, examine the numbers given in row m of Table T-6 in Appendix T. These are the confidence coefficients (which are ≥ 0.50) for which a two-sided confidence interval is available. Choose one of these confidence coefficients, say $1 - \alpha$, and proceed to step 3.

3 Go across row m in Table T-6 and select the value of r corresponding to the confidence coefficient $1 - \alpha$ chosen in step 2.

4 Using this value of r, compute $m - r + 1$.

5 For the values of r and $m - r + 1$, select q_r and q_{m-r+1} where $q_1 < q_2 < \ldots < q_m$ are obtained by ordering the q_i^* in (8.3.3).

6 A $1 - \alpha$ two-sided confidence interval for $a_0\beta_0 + a_1\beta_1$ is given by the confidence statement

$$C[q_r \leq a_0\beta_0 + a_1\beta_1 \leq q_{m-r+1}] = 1 - \alpha \tag{8.3.4}$$

Example 8.3.1 illustrates the relevant computations.

EXAMPLE 8.3.1

A random sample of 20 college professors was selected, and their annual salaries (Y) in thousands of dollars and number of years (X) of experience were recorded. The data are given in Table 8.3.3 and are stored in the file **profsal.dat** on the data disk. The investigator believes that neither assumptions (A) nor (B) hold, but the assumptions in Box 8.3.1 are appropriate. We compute a confidence interval for $\mu_Y(10)$, the average annual salary of all college professors with 10 years'

TABLE 8.3.3
Professors' Salary Data

Observation Number	Annual Salary Y (in thousands of dollars)	Experience X (in years)
1	63	19
2	48	14
3	50	14
4	47	9
5	41	7
6	44	10
7	43	7
8	66	20
9	78	28
10	59	16
11	49	12
12	65	21
13	67	21
14	58	13
15	40	8
16	69	22
17	58	15
18	71	20
19	51	12
20	49	13

experience, with confidence coefficient as close to 90% as possible. We exhibit the computations to obtain the q_i.

The observations are rearranged so that X values occur in increasing order. The ordered data along with the original data are given in Table 8.3.4. Note that some of the X values are repeated. For instance, the value $X = 7$ occurs twice with the corresponding Y values being 41 and 43. In each such case we compute the mean of the Y values corresponding to the same X value, which results in the condensed data set shown in Table 8.3.5.

TABLE 8.3.4
Professors' Salary Data

Original Data			Data Rearranged According to Increasing X		
Observation Number	Salary Y	Years X	Observation Number	Salary Y	Years X
1	63	19	5	41	7
2	48	14	7	43	7
3	50	14	15	40	8
4	47	9	4	47	9
5	41	7	6	44	10
6	44	10	11	49	12
7	43	7	19	51	12
8	66	20	14	58	13
9	78	28	20	49	13
10	59	16	2	48	14
11	49	12	3	50	14
12	65	21	17	58	15
13	67	21	10	59	16
14	58	13	1	63	19
15	40	8	8	66	20
16	69	22	18	71	20
17	58	15	12	65	21
18	71	20	13	67	21
19	51	12	16	69	22
20	49	13	9	78	28

TABLE 8.3.5
Condensed Data from Table 8.3.4

Y mean	X
42	7
40	8
47	9
44	10
50	12
53.5	13
49	14
58	15
59	16
63	19
68.5	20
66	21
69	22
78	28

Since $n = 14$, which is an even number, $m = n/2 = 7$. Thus we divide the data in Table 8.3.5 into four columns as in Table 8.3.1. We get

Column 1	Column 2	Column 3	Column 4
42	7	58	15
40	8	59	16
47	9	63	19
44	10	68.5	20
50	12	66	21
53.5	13	69	22
49	14	78	28

First we compute z and w in Table 8.3.2 and get

z	w
16	8
19	8
16	10
24.5	10
16	9
15.5	9
29	14

Next we compute u, v, and t in Table 8.3.2 and get

u	v	t
630	406	224
640	472	168
893	567	326
880	685	195
1050	792	258
1177	897	280
1372	1092	280

Next we compute the q_i^* in (8.3.3). Note that we have taken $a_0 = 1$ and $a_1 = 10$ in (8.3.3). We get

q^*
48.0000
44.7500
48.6000
44.0000
46.4444
48.3333
40.7143

Next we order the q_i^* from smallest to largest and get the q_i as follows:

40.7143 44.0000 44.7500 46.4444 48.0000 48.3333 48.6000

Because $m = 7$ is an odd number, the middle number, namely q_4, is 46.4444. Hence the estimated value of $\mu_Y(10)$ is 46 (rounded to the nearest thousand).

Look in Table T-6 in Appendix T across row $m = 7$ and find that the confidence coefficient that is nearest to 0.90 is 0.88 and the corresponding value of r is 2. Hence $m - r + 1 = 6$. Thus the 88% two-sided confidence bounds for $\mu_Y(10)$ are $q_2 = 44.0000$ and $q_6 = 48.3333$. We obtain the confidence statement (rounding to the nearest thousand)

$$C[44 \le \mu_Y(10) \le 48] = 0.88 \quad \blacksquare \tag{8.3.5}$$

Problems 8.3

8.3.1 Consider Problem 8.2.1 where an investigator is studying the relationship of sulfur dioxide concentrations, Y, in a national park, and the sulfur dioxide emission rate, X, by a coal burning power plant 25 miles away. Suppose that the regression function of Y on X is of the form

$$\mu_Y(x) = \beta_0 + \beta_1 x$$

and that assumptions in Box 8.3.1 are satisfied. Some of the computations required for this problem are in Tables 8.3.6–8.3.8. The asterisks (****) indicate that some values have not been computed, and you will be asked to supply them.

a What is the value of n?

b What is the value of m?

c Compute the missing values for w, v, and t in Table 8.3.7.

d Compute the missing values for q_i^* and q_i in Table 8.3.8.

TABLE 8.3.6
Computations for Table 8.3.1

Row	Column 1	Column 2	Column 3	Column 4
1	5.21	1.92	7.36	3.92
2	5.80	2.00	10.90	4.24
3	4.76	2.40	8.06	4.32
4	6.99	2.80	12.48	5.12
5	6.93	2.96	11.70	5.84
6	9.36	3.52	10.10	6.32
7	7.44	3.60	16.26	6.80

TABLE 8.3.7
Computations for Table 8.3.2

Row	z	w	u	v	t
1	2.15	2.00	20.4232	14.1312	6.2920
2	5.10	****	24.5920	****	****
3	3.30	1.92	20.5632	19.3440	1.2192
4	5.49	2.32	35.7888	34.9440	0.8448
5	4.77	2.88	40.4712	34.6320	5.8392
6	0.74	2.80	59.1552	****	23.6032
7	8.82	3.20	50.5920	58.5360	−7.9440

TABLE 8.3.8
Computations for Point Estimate and Confidence Interval for β_0

Row	q_i^*	q_i
1	3.14600	−2.48250
2	****	****
3	0.63500	0.63500
4	****	****
5	2.02750	2.02750
6	8.42971	3.14600
7	−2.48250	8.42971

e Use Theil's method and estimate β_0.

f Use Theil's method and estimate β_1.

g Use Theil's method and obtain a confidence interval for β_0 with a confidence coefficient as close to 90% as possible.

h Use Theil's method and obtain a confidence interval for β_1 with a confidence coefficient as close to 90% as possible.

i Compare the estimates in parts (e) and (f) with the estimates obtained using weighted regression in Problem 8.2.1.

j Use Theil's method and estimate the mean SO_2 concentration at the park associated with an emission rate of 5.0 tons/hour at the power plant.

k Write out the population parameters that represent the difference between the average SO_2 concentration at the park corresponding to a power plant SO_2 emission rate of 5.0 tons/hour and that corresponding to a power plant SO_2 emission rate of 2.5 tons/hour.

l Use Theil's method and estimate the difference in part (k) and compute a two-sided confidence interval for it with a confidence coefficient as close to 90% as possible.

8.3.2 Can the assumptions in Box 8.3.1 be satisfied for straight line regression if assumptions (A) are satisfied? Discuss.

8.3.3 Can the assumptions in Box 8.3.1 be satisfied for straight line regression if assumptions (B) are satisfied? Discuss.

8.3.4 Can assumptions (B) be satisfied for straight line regression if assumptions in Box 8.3.1 are satisfied? Discuss.

8.3.5 Can assumptions (A) be satisfied for straight line regression if assumptions in Box 8.3.1 are satisfied? Discuss.

8.3.6 Can the assumptions in Box 8.2.1 be satisfied for straight line regression if assumptions (B) are satisfied? Discuss.

8.3.7 Can the assumptions in Box 8.2.1 be satisfied for straight line regression if assumptions (A) are satisfied? Discuss.

8.3.8 Can the assumptions in Box 8.3.1 be satisfied for straight line regression if assumptions in Box 8.2.1 are satisfied? Discuss.

8.3.9 Can the assumptions in Box 8.2.1 be satisfied for straight line regression if assumptions in Box 8.3.1 are satisfied? Discuss.

8.4 Exercises

8.4.1 The texture score Y of a soybean product (soyburger, a meat substitute) depends to some extent on the percent X of a filler material used. Typically the texture score for a batch of this product is obtained by asking a trained panel of food experts to assign scores (from 0 to 10) and then taking the average of these individual rating scores. Texture scores greater than 7 indicate an acceptable product. The ultimate objective is to find the smallest amount of the filler material that will result in an acceptable texture score for the final product. Consequently a food engineer is interested in studying the relationship between the texture score and the percent of filler used. To do this, he makes several batches of soyburger with different amounts of filler material and obtains the texture scores for each batch. The data are given in Table 8.4.1 and are also stored in the file **soyburgr.dat** on the data disk.

TABLE 8.4.1
Soyburger Data

Batch Number	Texture Score Y	Filler Material X (percent)
1	2.5	0.5
2	2.9	1.0
3	3.4	1.5
4	3.7	2.0
5	4.3	2.5
6	4.5	3.0
7	4.9	3.5
8	5.8	4.0
9	6.4	4.5
10	6.8	5.0
11	6.5	5.5
12	8.0	6.0
13	8.4	6.5
14	8.5	7.0
15	7.4	7.5
16	9.9	8.0

Suppose the regression function of Y on X is of the form

$$\mu_Y(x) = \beta_0 + \beta_1 x \tag{8.4.1}$$

and that the weighted regression assumptions in Box 8.2.1 hold with $g(x) = x^2$ so that the weights are given by $w_i = 1/x_i^4$. A computer output for a weighted regression analysis of Y on X obtained using MINITAB is given in Exhibit 8.4.1. Additionally we also give the quantities SSX, SSY, SXY, $\bar{x}$, and $\bar{y}$ required to compute the ordinary least squares estimates of β_0 and β_1.

a Perform an ordinary regression of Y on X and calculate the residuals and standardized residuals. Examine appropriate plots. Based on these plots, do you think any of assumptions (A) appear to be violated? In particular, does it appear that the homogeneity of variance assumption holds?

Answer parts (b)–(f) using weighted regression.

b Estimate the regression function of Y on X given in (8.4.1).

c Obtain a two-sided 99% confidence interval for the mean texture score of all batches of soyburger made with 6% filler material.

d Obtain a two-sided 90% confidence interval for the texture score of a single batch of soyburger to be made with 6% filler material.

e Obtain a two-sided 80% confidence interval for the standard deviation of the texture scores of all batches of soyburger made with 6% filler material.

f Estimate the proportion of batches of soyburger made with 6% filler material that will have texture scores in the acceptable range (a score of 7 or greater).

EXHIBIT 8.4.1

MINITAB Output for Exercise 8.4.1

ROW	texture	filler	weights
1	2.5	0.5	16.0000000000
2	2.9	1.0	1.0000000000
3	3.4	1.5	0.1975308657
4	3.7	2.0	0.0625000000
5	4.3	2.5	0.0255999994
6	4.5	3.0	0.0123456791
7	4.9	3.5	0.0066638901
8	5.8	4.0	0.0039062500
9	6.4	4.5	0.0024386526
10	6.8	5.0	0.0016000000
11	6.5	5.5	0.0010928215
12	8.0	6.0	0.0007716049
13	8.4	6.5	0.0005602045
14	8.5	7.0	0.0004164931
15	7.4	7.5	0.0003160494
16	9.9	8.0	0.0002441406

```
The regression equation is
texture = 2.07 + 0.858 filler

Predictor        Coef       Stdev   t-ratio        p
Constant      2.06979     0.01139    181.71    0.000
filler        0.85776     0.01883     45.56    0.000

Analysis of Variance

SOURCE       DF         SS          MS          F        p
Regression    1    0.74544     0.74544    2075.50    0.000
Error        14    0.00503     0.00036
Total        15    0.75047

The weighted C matrix is

  0.361228  -0.547297
 -0.547297   0.987004

     SSY = 74.2544;   SSX = 85.00;   SXY = 77.525;
     mean of y = 5.86875;       mean of x = 4.25;
```

g What might be an appropriate target population (of items and of numbers) for this problem? What is the study population?

8.4.2 The final exam scores (Y) in a particular statistics course are related to the midterm test scores (X) according to a straight line regression model. A random sample of 24 students during a particular semester yielded the data in Table 8.4.2, which are also in the file **exam.dat** on the data disk. Suppose that an investigator believes assumptions (A) for straight line regression are satisfied. A computer output containing the results of an ordinary regression analysis of Y on X is given in Exhibit 8.4.2.

TABLE 8.4.2
Exam Scores Data

Student	Final Exam Score (Y)	Midterm Exam Score (X)
1	40	44
2	47	48
3	41	49
4	41	50
5	43	52
6	42	53
7	50	54
8	87	58
9	61	61
10	74	66
11	75	75
12	89	76
13	72	77
14	69	78
15	78	80
16	78	83
17	92	84
18	84	85
19	85	86
20	99	87
21	89	90
22	83	91
23	96	95
24	100	99

EXHIBIT 8.4.2

MINITAB Output for Exercise 8.4.2

```
The regression equation is
final = -6.13 + 1.08 midterm


Predictor        Coef       Stdev   t-ratio        p
Constant       -6.132       8.147     -0.75    0.460
midterm        1.0820      0.1106      9.78    0.000

s = 9.093      R-sq = 81.3%     R-sq(adj) = 80.5%

Analysis of Variance

SOURCE        DF        SS          MS          F        p
Regression     1    7910.9      7910.9      95.68    0.000
Error         22    1819.0        82.7
Total         23    9730.0

      ROW  final   midterm       fits    stdresid     nscores
        1     40        44     41.477    -0.17675     0.15498
        2     47        48     45.805     0.14045     0.73025
        3     41        49     46.887    -0.68940    -0.73025
        4     41        50     47.969    -0.81308    -0.87320
        5     43        52     50.133    -0.82653    -1.03703
        6     42        53     51.215    -1.06442    -1.49944
        7     50        54     52.297    -0.26458    -0.26024
        8     87        58     56.626     3.46287     1.95007
        9     61        61     59.872     0.12790     0.60110
       10     74        66     65.282     0.98187     1.03703
       11     75        75     75.020    -0.00225     0.48148
       12     89        76     76.102     1.45102     1.49944
       13     72        77     77.184    -0.58364    -0.48148
       14     69        78     78.266    -1.04414    -1.23521
       15     78        80     80.430    -0.27446    -0.36854
       16     78        83     83.676    -0.64404    -0.60110
       17     92        84     84.758     0.82320     0.87320
       18     84        85     85.840    -0.20961     0.05147
       19     85        86     86.922    -0.21944    -0.05147
       20     99        87     88.004     1.25817     1.23521
       21     89        90     91.250    -0.25961    -0.15498
       22     83        91     92.332    -1.07989    -1.95007
       23     96        95     96.661    -0.07752     0.36854
       24    100        99    100.989    -0.11806     0.26024
```

a Estimate the regression function $\mu_Y(x) = \beta_0 + \beta_1 x$.

b Plot the standardized residuals r_i against the fitted values $\hat{\mu}_Y(x_i)$. Does this plot suggest that any of the assumptions are violated?

c Obtain a Gaussian rankit-plot of r_i. Do the standardized residuals appear to be a simple random sample from a Gaussian population with zero mean and unit standard deviation?

d Suppose that from the plot in part (b) and other considerations, an investigator believes that assumptions (A) are not valid, and she decides to use Theil's method to estimate β_0, β_1, and $\mu_Y(x)$. Some results to help you to do the computations for Theil's method for regression are given in Exhibit 8.4.3.

EXHIBIT 8.4.3

Some Calculations for Theil's Method of Regression

Results for Table 8.3.1

ROW	column1	column2	column3	column4
1	40	44	72	77
2	47	48	69	78
3	41	49	78	80
4	41	50	78	83
5	43	52	92	84
6	42	53	84	85
7	50	54	85	86
8	87	58	99	87
9	61	61	89	90
10	74	66	83	91
11	75	75	96	95
12	89	76	100	99

EXHIBIT 8.4.3
(Continued)

Results for Table 8.3.2

ROW	z	w	u	v	t
1	32	33	3080	3168	-88
2	22	30	3666	3312	354
3	37	31	3280	3822	-542
4	37	33	3403	3900	-497
5	49	32	3612	4784	-1172
6	42	32	3570	4452	-882
7	35	32	4300	4590	-290
8	12	29	7569	5742	1827
9	28	29	5490	5429	61
10	9	25	6734	5478	1256
11	21	20	7125	7200	-75
12	11	23	8811	7600	1211

i Estimate the regression function $\mu_Y(x) = \beta_0 + \beta_1 x$ using Theil's method.

ii Using Theil's method obtain two-sided confidence intervals for β_0 and β_1 with confidence coefficient as close to 85% as possible.

iii Using Theil's method predict the final exam score of a student who obtained 75 points on the midterm exam.

iv Using Theil's method obtain a two-sided confidence interval for $\mu_Y(75)$, the average final exam score of all students who obtain 75 point on the midterm examination. Use a confidence coefficient as close to 85% as possible.

e Explain what might be an appropriate target population for this problem. What is the study population?