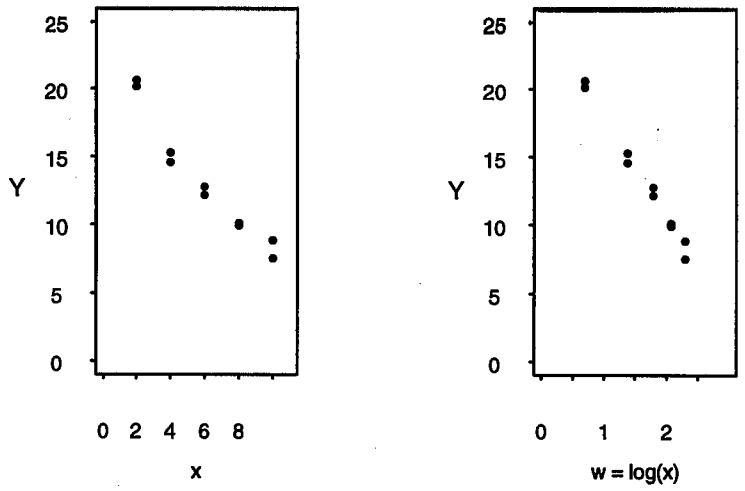

FIGURE 4.6

Problem 4.1.2: Find a 90% confidence interval on γ_1 for the data in Example 4.1.2. What model justifies this interval?

Problem 4.1.3: For the (x_i, Y_i) pairs of Table 4.1.3 and Figure 4.7 define $w_i = \log x_i$, $z_i = \log Y_i$, $u_i = 1/x_i$. Suggest a model, estimate the parameters, and sketch the resulting function $h(x)$.

Problem 4.1.4: Verify the two values of R^2 given by Scott and Wild.

4.2 SPECIFICATION ERROR

It is often difficult to determine which of many possible variables in $\tilde{\mathbf{x}} = (x_1, \ldots, x_k)$ to use in estimating the regression function $g(\tilde{\mathbf{x}}) \equiv E(\mathbf{Y}|\tilde{\mathbf{x}})$ or in predicting $\mathbf{Y}$, particularly in cases for which n is relatively small. The statistican is torn between the wish to keep the model simple and the wish for a good approximation. If a poor choice of a subset $x_{i_1}, \ldots, x_{i_r}$ of possible measurements is made, what will the penalty be?

For example, in trying to determine the regression function $g(\tilde{\mathbf{x}})$, should we use a fifth degree polynomial, or will a quadratic function suffice? Obviously we can fit the data more closely with a fifth degree polynomial, but may pay a price in increased complication, poor extrapolation, and, as we shall see, a loss of precision. On the other hand, if the true regression is cubic (it would be better to say, is *approximately* cubic for x of interest) and we fit a quadratic function some inaccuracy (bias) would seem to result.

To make the discussion precise suppose $\mathbf{Y} = \boldsymbol{\theta} + \boldsymbol{\varepsilon}$ for $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I}_n)$ and

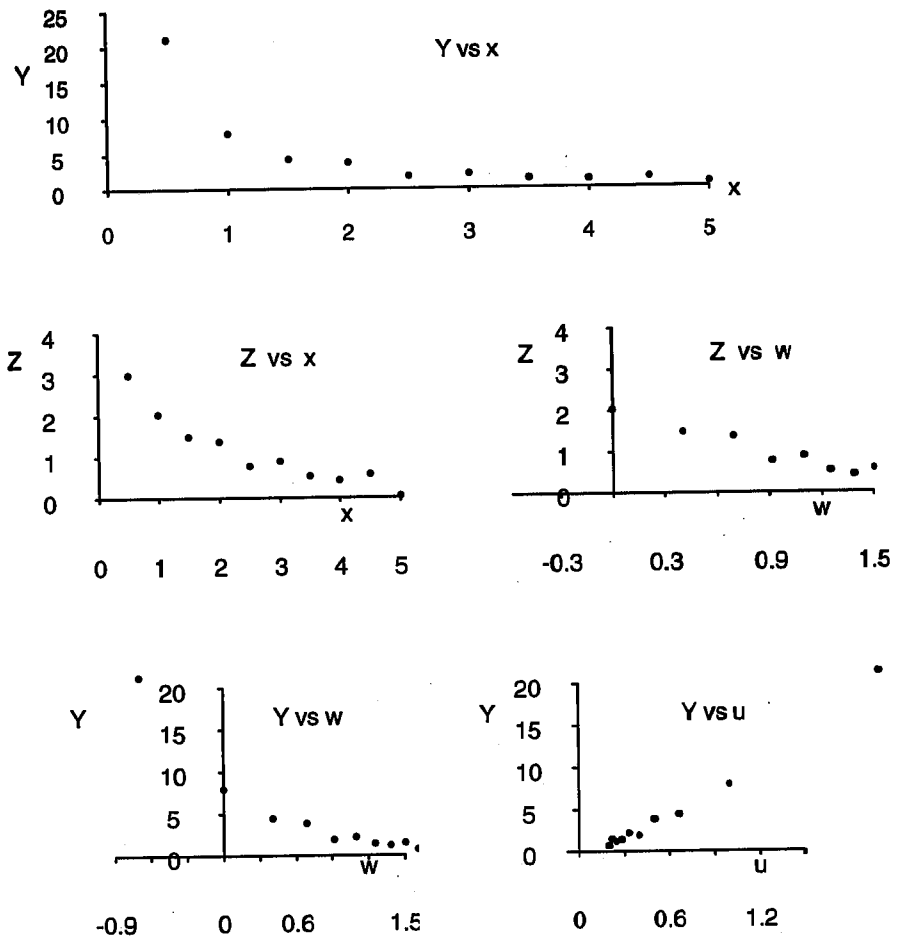

FIGURE 4.7 Scatterplots among x, $w = \log x$, $u = 1/x$, Y, and $z = \log Y$.

our postulated model is $\boldsymbol{\theta} \in V$, a known subspace of R_n of dimension k. As will be seen by the following analysis, if $\boldsymbol{\theta} \notin V$, errors may result. To see this, let $\boldsymbol{\theta} = \boldsymbol{\theta}_V + \boldsymbol{\theta}_\perp$ for $\boldsymbol{\theta}_V = p(\boldsymbol{\theta}|V)$. Let $\boldsymbol{\varepsilon}_V = p(\boldsymbol{\varepsilon}|V)$ and $\boldsymbol{\varepsilon}_\perp = \boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_V$. Then the least squares estimator of $\boldsymbol{\theta}$ is $\hat{\mathbf{Y}} = \boldsymbol{\theta}_V + \boldsymbol{\varepsilon}_V$ and the error in the estimation of $\boldsymbol{\theta}$ is $\mathbf{d} = \hat{\mathbf{Y}} - \boldsymbol{\theta} = -\boldsymbol{\theta}_\perp + \boldsymbol{\varepsilon}_V$. Thus, $\hat{\mathbf{Y}}$ has bias $-\boldsymbol{\theta}_\perp$. We can assess the expected sizes of the errors made in estimating $\boldsymbol{\theta}$ by computing

$$E(\hat{\mathbf{Y}} - \boldsymbol{\theta})(\hat{\mathbf{Y}} - \boldsymbol{\theta})' = \boldsymbol{\theta}_\perp\boldsymbol{\theta}_\perp' + E(\boldsymbol{\varepsilon}_V\boldsymbol{\varepsilon}_V') = \boldsymbol{\theta}_\perp\boldsymbol{\theta}_\perp' + E(\mathbf{P}_V\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{P}_V) = \boldsymbol{\theta}_\perp\boldsymbol{\theta}_\perp' + \sigma^2\mathbf{P}_V.$$

Here we have taken advantage of the orthogonality of $\boldsymbol{\theta}_\perp$ and $\boldsymbol{\varepsilon}_V$ (Figure 4.8). To gauge the size of this we can compute the sum of the expected squared errors.

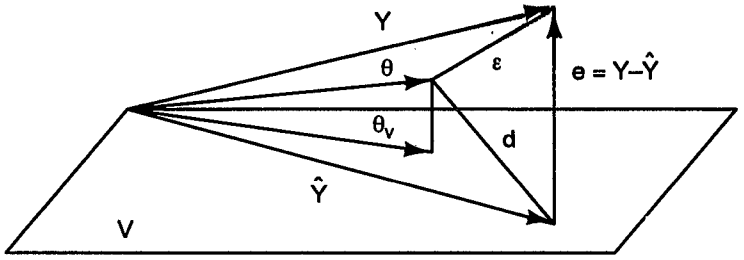

FIGURE 4.8

$$E\|\mathbf{d}\|^2 = E[\|\boldsymbol{\theta}_\perp\|^2 + \|\boldsymbol{\varepsilon}_V\|^2] = \|\boldsymbol{\theta}_\perp\|^2 + k\sigma^2,$$

since V has dimension k.

We might also study the random variable $Q = \|\mathbf{d}\|^2 = \|\boldsymbol{\theta}_\perp\|^2 + \|\boldsymbol{\varepsilon}_V\|^2$ in order to understand the sizes of these errors. Q is a constant plus σ^2 multiplied by a central χ^2 random variable (not noncentral χ^2), with expectation given above.

Error sum of squares is $\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 = \|\boldsymbol{\theta}_\perp + \boldsymbol{\varepsilon}_\perp\|^2$, so that $\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2/\sigma^2 \sim \chi^2_{n-k}(\delta)$ for $\delta = \|\boldsymbol{\theta}_\perp\|^2/\sigma^2$. Thus, $E(S^2) = \sigma^2 + \|\boldsymbol{\theta}_\perp\|^2/(n-k)$.

In searching for a good model we might try to choose a subspace V so that $H_V = E\|\mathbf{d}\|^2/\sigma^2 = \|\boldsymbol{\theta}_\perp\|^2/\sigma^2 + k$ is small. Of course, H_V depends on unknown parameters. It can be estimated if we can find an estimator of *pure error variance* σ^2. We might, for example, use a particularly large subspace V_L in which we are quite sure $\boldsymbol{\theta}$ lies, and use error sum of squares for this subspace to estimate σ^2. Or we might use past data from another experiment with repeated observations on $\mathbf{Y}$ for the same $\tilde{\mathbf{x}}$ to estimate σ^2. Let $\hat{\sigma}^2$ be this estimator of pure error variance. Let S^2 be the estimator for the subspace V.

Then $E(S^2 - \hat{\sigma}^2)(n-k) = \|\boldsymbol{\theta}_\perp\|^2$, so that $C_V = \dfrac{(S^2 - \hat{\sigma}^2)}{\hat{\sigma}^2}(n-k) + k$ can be used as an estimator of H_V. C_V is called *Mallows* C_p for the case that $\dim(V) = p$ (Mallows 1964). Since $H_V = \dim(V)$ for $\boldsymbol{\theta} \in V$, we should hope to find a subspace V such that C_V is close to or smaller than $\dim(V)$.

Consider, for example, a sequence of regression vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots$, with order chosen by the statistician. $\mathbf{x}_j$ might, for example, be the vector of jth powers. Then for $V_k = \mathscr{L}(\mathbf{x}_1, \ldots, \mathbf{x}_k)$ and $C_k = C_{V_k}$, we can compute the sequence $C_1, C_2, \ldots$ and, as recommended by Mallows, plot the points (k, C_k), choosing the subspace V_k for the smallest k for which C_k is close to k.

One possible criterion for the choice of a subspace V_0 rather than a subspace V in which $\boldsymbol{\theta}$ is known to lie may be developed as follows. Since the bias of $\hat{\mathbf{Y}}_0 = p(\mathbf{Y}|V_0)$ is $p(\boldsymbol{\theta}|V_0) - \boldsymbol{\theta} = -\boldsymbol{\theta}_\perp$ and the sum of the squared errors is $Q = \|\boldsymbol{\theta}_\perp\|^2 + \|\boldsymbol{\varepsilon}_{V_0}\|^2$, we have $E(Q) = \|\boldsymbol{\theta}_\perp\|^2 + k_0\sigma^2$ (see Theorem 2.2.2). The

sum of squared errors for $\hat{\mathbf{Y}} = p(\mathbf{Y}|V)$ is $\|\boldsymbol{\varepsilon}_V\|^2$, which has expectation $k\sigma^2$. Thus we should choose V_0 if

$$\|\boldsymbol{\theta}_\perp\|^2 + k_0\sigma^2 < k\sigma^2, \quad \text{or} \quad \|\boldsymbol{\theta}_\perp\|^2 < (k-k_0)\sigma^2, \tag{4.2.1}$$

equivalently if the noncentrality parameter $\delta = \|\boldsymbol{\theta}_\perp\|^2/\sigma^2$ in the F-test of $H_0: \boldsymbol{\theta} \in V_0$ is less than $k - k_0$.

Let $Q = \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2$ and $Q_0 = \|\mathbf{Y} - \hat{\mathbf{Y}}_0\|^2$. Then $Q - Q_0 = \|\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_0\|^2$ has expectation $\|\boldsymbol{\theta}_\perp\|^2 + (k-k_0)\sigma^2$ and $E(Q) = (n-k)\sigma^2$, so that

$$Q_0 - Q\left(\frac{k-k_0}{n-k}\right)$$

is an unbiased estimator of $\|\boldsymbol{\theta}_\perp\|^2$. Thus, if we replace the parameters in (4.2.1) by unbiased estimators we get $Q_0 < 2Q[(k-k_0)/(n-k)]$, equivalently $F = \dfrac{(Q - Q_0)/(k-k_0)}{Q/(n-k)} < 2$. This is equivalent to $C_k < k$ (see Problem 4.25).

Example 4.2.1: The data of Table 4.2.1 were generated using the regression function

$$g(\tilde{\mathbf{x}}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1^2$$

Table 4.2.1

i	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{Y}_1$	$\mathbf{Y}_2$	$\hat{\mathbf{Y}}_1$	$\hat{\mathbf{Y}}_2$
1	1	1	1	8.04	11.46	9.81	14.33
2	1	2	2	17.57	20.17	16.93	19.20
3	2	3	4	27.43	34.60	28.20	27.11
4	2	4	3	28.96	20.35	27.78	25.23
5	3	1	1	15.78	18.72	11.48	14.69
6	3	2	5	30.91	30.04	29.92	29.70
7	4	3	4	28.95	18.27	31.72	29.55
8	4	4	3	28.79	33.89	31.30	27.67
9	5	1	5	40.74	34.71	31.94	32.72
10	5	2	4	23.04	26.81	31.52	30.84
11	6	3	3	29.60	34.78	35.18	32.77
12	6	4	1	33.98	24.67	30.98	27.50
13	7	1	1	24.65	26.36	25.93	27.88
14	7	2	1	31.42	41.04	29.28	29.37
15	8	3	1	39.82	37.68	38.56	36.76
16	8	4	5	62.77	54.81	57.00	51.76
17	9	1	5	49.23	53.78	53.79	54.21
18	9	2	3	47.81	50.30	49.60	48.95
19	10	3	3	65.14	55.03	60.73	58.41
20	10	4	1	53.56	54.30	56.53	53.15