Block 11

Variances

Variances

We now consider various extensions to linear model that relax variance assumptions on errors.
This greatly expands range of problems linear regression can be applied.

Weighted least squares

The assumption that the variance function $\mathrm{Var}(Y|X)$ is the same for all values of the terms X can be relaxed as follows.

$$\mathrm{Var}(Y|X = x_i) = \mathrm{Var}(e_i) = \frac{\sigma^2}{w_i},$$

where $w_1, \ldots, w_n$ are known positive numbers. This leads to the use of weighted least squares, or WLS, in place of OLS, to get estimates.
In matrix terms, the model can be written as

$$Y = X\beta + e, \quad \mathrm{Var}(e) = \sigma^2 W^{-1}.$$

The estimator $\hat{\beta}$ is chosen to minimize the weighted residual sum of squares function,

$$RSS(\beta) = (Y - X\beta)'W(Y - X\beta) = \sum_i w_i(y_i - x_i\beta)^2.$$

The WLS estimator is given by

$$\hat{\beta} = (X'WX)^{-1}X'WY.$$

Let $W^{1/2}$ be the $n \times n$ diagonal matrix with ith diagonal element $\sqrt{w_i}$, and so $W^{-1/2}$ is a diagonal matrix with $1/\sqrt{w_i}$ on the diagonal. Define $Z = W^{1/2}Y$,

$M = W^{1/2}X$, and $d = w^{1/2}e$, and (1) is equivalent to

$$Z = M\beta + d.$$

This model can be solved using OLS,

$$\hat{\beta} = (M'M)^{-1}M'Z = (X'WX)^{-1}X'WY,$$

which is identical to the WLS estimator.

$$E(\hat{\beta}|X) = E\left[(X'WX)^{-1}X'WY|X\right]$$

$$= (X'WX)^{-1}X'WE(Y|X)$$

$$= (X'WX)^{-1}X'WX\beta$$

$$= \beta$$

$\therefore$ WLS estimator is unbiased.

$$\mathrm{Var}(\hat{\beta}|X) = \mathrm{Var}\left((X'WX)^{-1}X'WY|X\right)$$

$$= (X'WX)^{-1}X'W\left[\mathrm{Var}(Y|X)\right]WX(X'WX)^{-1}$$

$$= (X'WX)^{-1}X'W\left[\sigma^2W^{-1}\right]WX(X'WX)^{-1}$$

$$= \sigma^2(X'WX)^{-1}$$

When $W = I$, we get back the usual formula as in the unweighted case.

Applications of weighted least squares

Known weight w_i can occur in many ways. If the ith response is an average of n_i equally variable observations, then $\mathrm{Var}(y_i) = \sigma/n_i$, and $w_i = n_i$. If y_i is a total of n_i observations, $\mathrm{Var}(y_i) = n_i\sigma^2$, and $w_i = 1/n_i$. If variance is proportional to some predictor x_i, $\mathrm{Var}(y_i) = x_i\sigma^2$, then $w_i = 1/x_i$.

Strong interaction

In physics, a theoretical model of the strong interaction force (that holds nuclei together) predicts that

$$E(y|s) = \beta_0 + \beta_1 s^{-1/2} + \text{relatively small terms.}$$

s: square of total energy of system

In an experiment, the following data are observed:

At each value of s ($x = s^{-1/2}$), a very large number of particles was counted, and as a result the values of $\text{Var}(y|s = s_i) = \sigma^2/w_i$ are known almost exactly; the square roots of these values are given in the third column of the Table below.

Table 7.1 The Strong Interaction Data

$x = s^{-1/2}$	y (mb)	SD_i
0.345	367	17
0.287	311	9
0.251	295	9
0.225	268	7
0.207	253	7
0.186	239	6
0.161	220	6
0.132	213	6
0.084	193	5
0.060	192	5

Table 7.2 WLS Estimates for the Strong Interaction Data

	Estimate	Std. Error	t-Value	$\Pr(>\|t\|)$
(Intercept)	148.4732	8.0786	18.38	0.0000
x	530.8354	47.5500	11.16	0.0000

$\hat{\sigma} = 1.6565$ with $8\,df$, $R^2 = 0.9397$

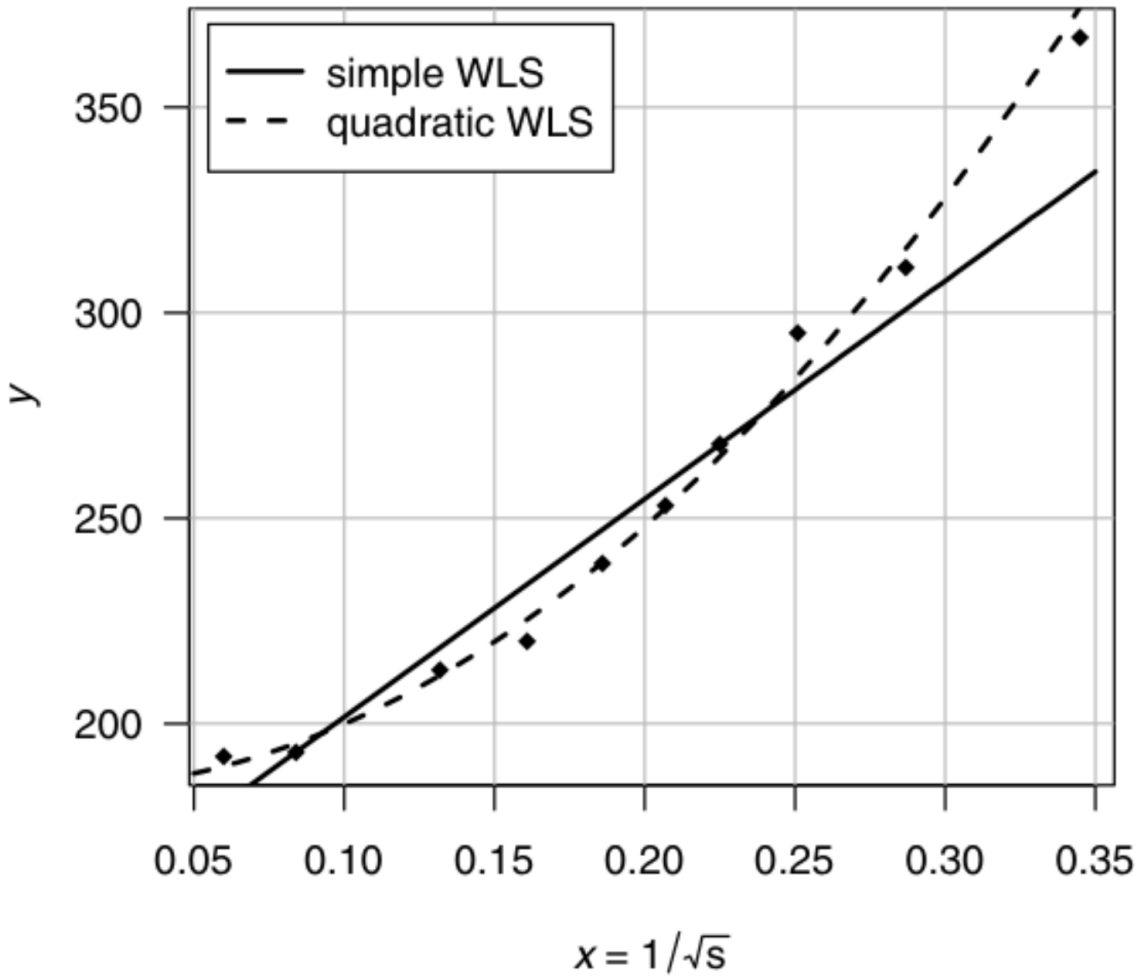


Figure 7.1 Scatterplot for the strong interaction data.

Case Study

Weighted Least Squares Regression

Dose-Response Study for Rosuvastatin in Japanese Patients with High Cholesterol

"Randomized Dose-Response Study of Rosuvastatin in Japanese Patients with Hypercholesterolemia" Saito, et al, Journal of Atherosclerosis, 2003, 10:329-336

Errors are independent

Variance of errors are not all equal (Heteroscedastic)

Variances may be known or estimated

Estimates can be obtained by regression when the variance is a power function of the mean

General case with known variance structure (up to σ^2)

$$\text{Var}(e) = \text{Var}\begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} = \sigma^2 V = \sigma^2 \begin{pmatrix} v_1 & 0 & \dots & 0 \\ 0 & v_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & v_n \end{pmatrix}$$

Weighted Least Squares Procedure

Give higher weights to observations with smaller variances

Create a Weight matrix W, that is diagonal with elements equal to reciprocals of square roots of elements of V Transform the Y vector and X matrix by pre-multiplyin

$$W = V^{-1/2} = \begin{pmatrix} 1/\sqrt{v_1} & 0 & \dots & 0 \\ 0 & 1/\sqrt{v_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1/\sqrt{v_n} \end{pmatrix}$$

$$X^* = WX \quad \text{and} \quad Y^* = WY \quad \Rightarrow \quad \hat{\beta}_W = (X^{*'}X^*)^{-1}X^{*'}Y^*$$

Example – Cholesterol Drug Dose-Response Study

- Study of Drug Rosuvastatin in Japanese Patients with high cholesterol
- 6 Doses – 1, 2.5, 5, 10, 20, 40
- Response - % Change in LDL Cholesterol @ week 12
- Data Reported – Group Means by Dose
- Sample sizes Varied by dose
- Assuming equal variance among individual patients:

$V\left(\bar{Y}_j\right) = \frac{\sigma^2}{n_j} \quad j = 1, \dots, 6$

Data and Y, Y^*, X, X^*, V and W

↓We will use this

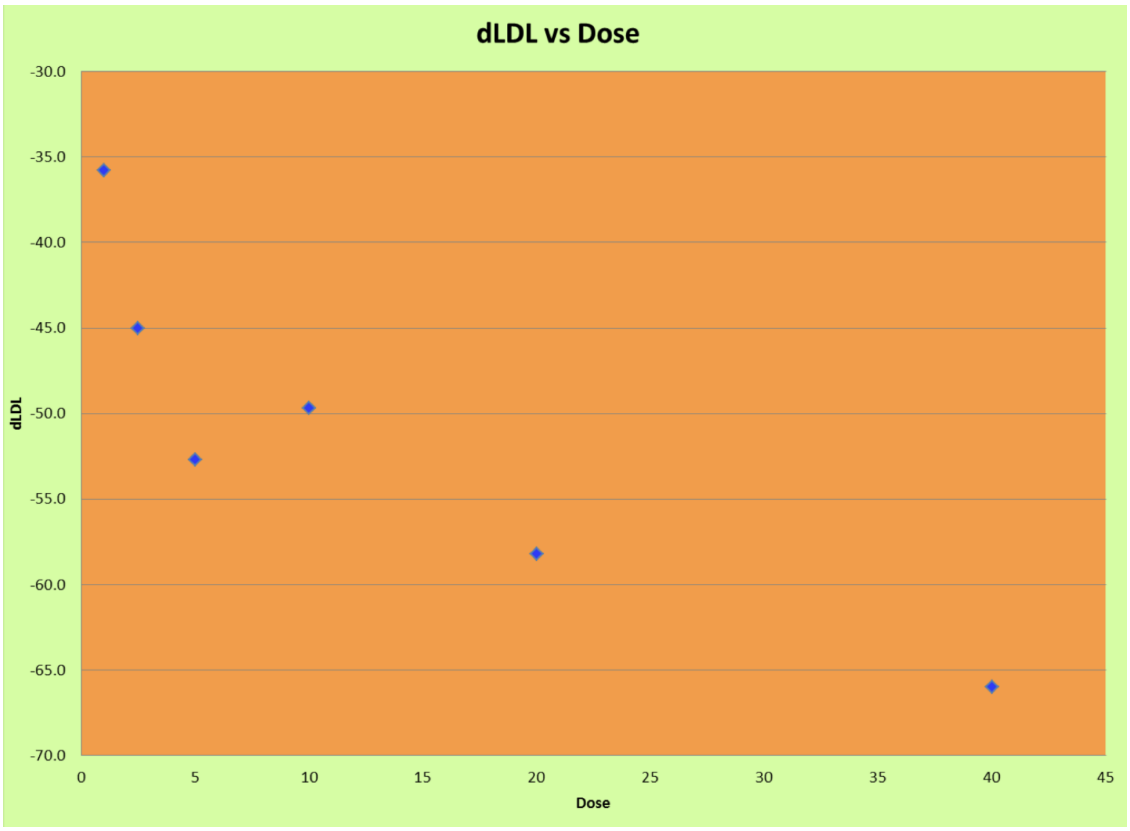
TxGrp	dLDL	DOSE	ln(DOSE)	n
1	-35.8	1	0.0000	15
2	-45.0	2.5	0.9163	17
3	-52.7	5	1.6094	12
4	-49.7	10	2.3026	14
5	-58.2	20	2.9957	18
6	-66.0	40	3.6889	13

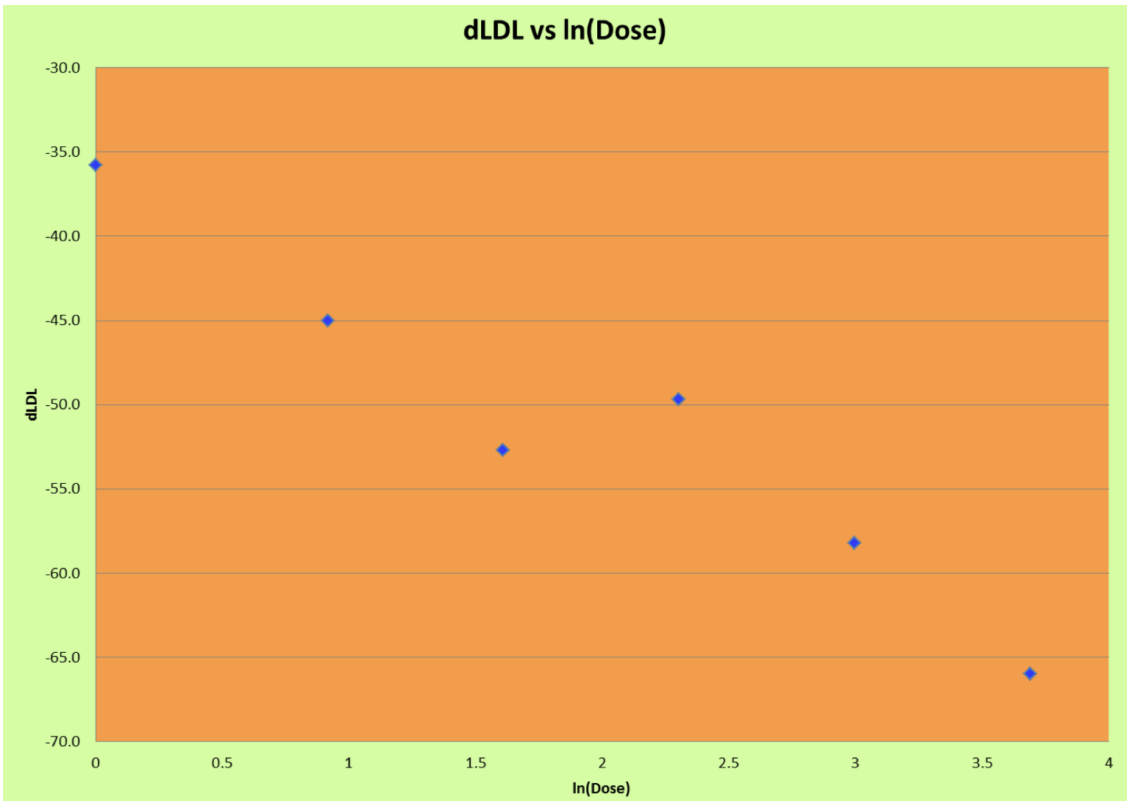
X			Y
1	0.0000		-35.8000
1	0.9163		-45.0000
1	1.6094		-52.7000
1	2.3026		-49.7000
1	2.9957		-58.2000
1	3.6889		-66.0000

V					
0.066667	0	0	0	0	0
0	0.058824	0	0	0	0
0	0	0.083333	0	0	0
0	0	0	0.071429	0	0
0	0	0	0	0.055556	0
0	0	0	0	0	0.076923

W					
3.872983	0	0	0	0	0
0	4.123106	0	0	0	0
0	0	3.464102	0	0	0
0	0	0	3.741657	0	0
0	0	0	0	4.242641	0
0	0	0	0	0	3.605551

X*			Y*
3.872983	0		-138.653
4.123106	3.777963		-185.54
3.464102	5.575256		-182.558
3.741657	8.615485		-185.96
4.242641	12.70982		-246.922
3.605551	13.30044		-237.966


dLDL vs Dose
-30.0
-35.0
-40.0
-45.0
-50.0
-55.0
-60.0
-65.0
-70.0
dLDL
0
5
10
15
20
25
30
35
40
45
Dose


dLDL vs ln(Dose)
-30.0
-35.0
-40.0
-45.0
-50.0
-55.0
-60.0
-65.0
-70.0
dLDL
0
0.5
1
1.5
2
2.5
3
3.5
4
ln(Dose)

WLS Regression Estimates, Fitted Values, Residuals

X*'X*			X*'Y*
89	169.005		-4535.8
169.005	458.0243		-9624.3
INV(X*'X*)			Beta_w
0.0375384	-0.01385		-36.9588
-0.013851	0.007294		-7.37531

X*			Y*	Yhat*	e*
3.8729833	0		-138.653	-143.141	4.488157
4.1231056	3.777963		-185.54	-180.249	-5.29093
3.4641016	5.575256		-182.558	-169.148	-13.4098
3.7416574	8.615485		-185.96	-201.829	15.86876
4.2426407	12.70982		-246.922	-250.542	3.620149
3.6055513	13.30044		-237.966	-231.352	-6.61457

SS(e*)	s2(e*)
536.63499	134.1587

Parameter	Estimate	SE	t	Pr>|t|
Intercept	-36.9588	2.2441	-16.4691	0.0001
ln(DOSE)	-7.3753	0.9892	-7.4556	0.0017

ln(DOSE)	dLDL	Yhat_w	e
0	-35.8	-36.9588	1.1588
0.916291	-45.0	-43.7168	-1.2832
1.609438	-52.7	-48.8289	-3.8711
2.302585	-49.7	-53.9411	4.2411
2.995732	-58.2	-59.0533	0.8533
3.688879	-66.0	-64.1654	-1.8346

Misspecified Variances

Suppose the true regression model is

$$E(Y|X) = X\beta, \quad \text{Var}(Y|X) = \sigma^2 W^{-1},$$

where W has positive weights on the diagonal and zeros elsewhere. We get the weights wrong, and fit the model using OLS with the estimator

$$\hat{\beta}^{\text{OLS}} = (X'X)^{-1}X'Y.$$

Similar to the correct WLS estimate,

$$E(\hat{\beta}^{\text{OLS}}|X) = \beta$$

However,

$$\text{Var}(\hat{\beta}^{\text{OLS}}|X) = \sigma^2(X'X)^{-1}(X'W^{-1}X)(X'X)^{-1}$$

"Sandwich" form

$\Rightarrow$ Confidence statements and tests can be incorrect.

Accommodating Incorrect Variances

To estimate $\text{Var}(\hat{\beta}^{\text{OLS}}|X)$ requires estimating $\sigma^2 W^{-1}$.
Let $\hat{e}_i = y_i - \hat{\beta}'^{\text{OLS}} x_i$ (i-th residual from misspecified model).
Can estimate σ^2/w_i by $\hat{e}_i^2$.
If we replace $\sigma^2 W^{-1}$ by a diagonal matrix with $\hat{e}_i^2$ on the diagonal, this produces a consistent estimate of $\text{Var}(\hat{\beta}^{\text{OLS}}|X)$.
An alternative version of this is called the sandwich estimator.
$\text{Var}(\hat{\beta}^{\text{OLS}}|X) = (X'X)^{-1}[X' \,\text{diag}\left(\frac{\hat{e}_i^2}{(1-h_{ii})^2}\right) X](X'X)^{-1}$
i-th leverage (to be discussed soon).

Sniffer Data

When gasoline is pumped into a tank, hydrocarbon vapors are forced out of the tank and into the atmosphere. To reduce this significant source of air pollution, devices are installed to capture the vapor. In testing these vapor recovery systems, a "sniffer" measures the amount recovered. To estimate the efficiency of the system, some method of estimating the total amount given off must be used. To this end, a laboratory experiment was conducted in which the amount of vapor given off was measured under controlled conditions. Four predictors are relevant for modeling:

Predictors:

TankTemp = initial tank temperature (degrees F)

GasTemp = temperature of the dispensed gasoline (degrees F)

TankPres = initial vapor pressure in the tank (psi)

GasPres = vapor pressure of the dispensed gasoline (psi)

Response: the hydrocarbons Y emitted in grams.

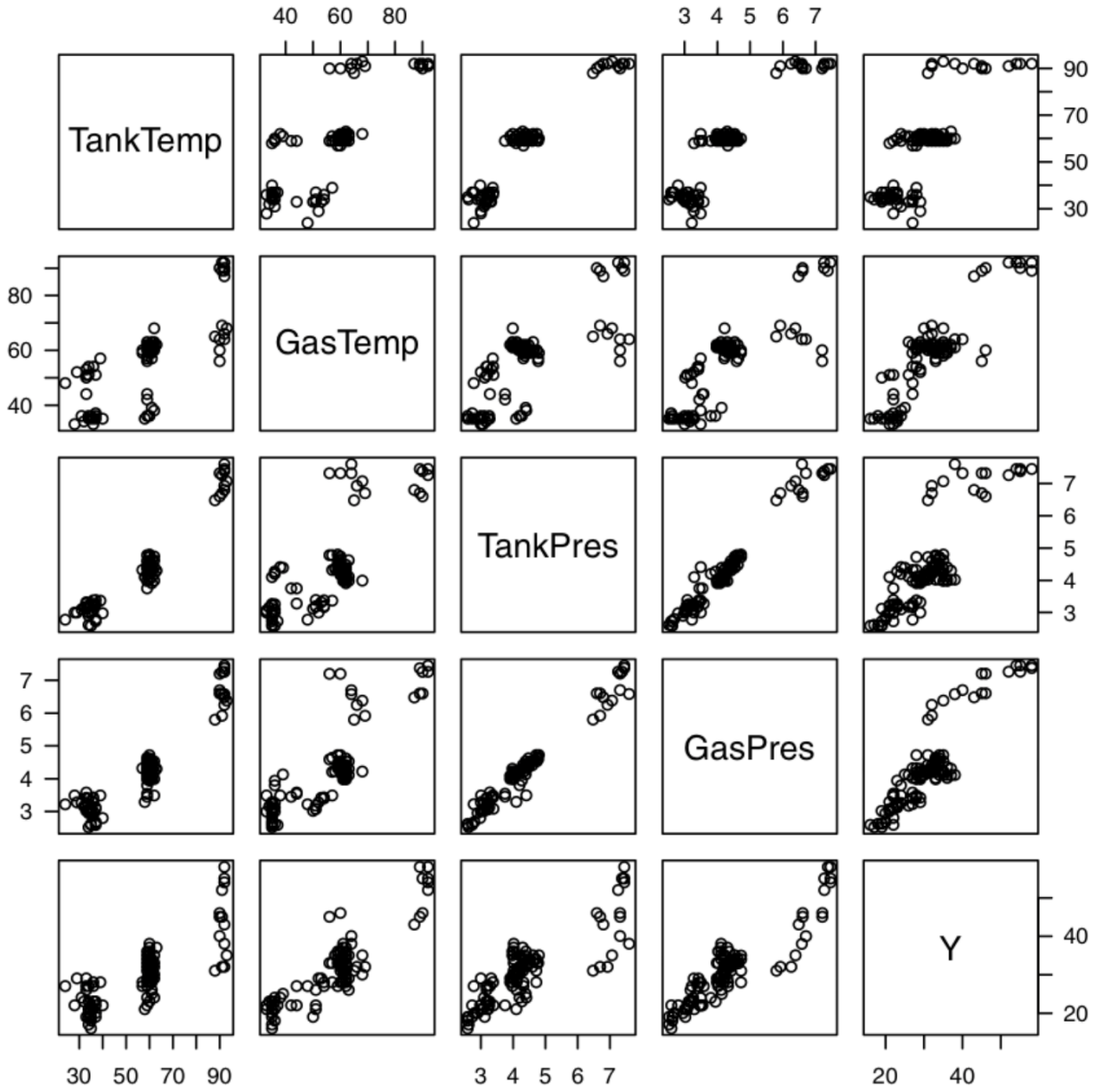


Figure 7.2 Scatterplot matrix for the sniffer data.

sandwich estimator

Table 7.3 Sniffer Data Estimates and Standard Errors

	OLS Est	OLS SE	HC3 SE
(Intercept)	0.154	1.035	1.047
TankTemp	–0.083	0.049	0.044
GasTemp	0.190	0.041	0.034
TankPres	–4.060	1.580	1.972
GasPres	9.857	1.625	2.056

A Test for Constant Variance

Suppose that for some parameter vector λ and some vector of regressors Z

$$\text{Var}(Y|X, Z = z) = \sigma^2 \exp(\lambda' z),$$

where $w = \exp(-\lambda' z)$.
If $\lambda = 0 \Rightarrow$ constant variance
Therefore to test
$H_0 : \lambda = 0$
$H_1 : \lambda \neq 0$
is a test for nonconstant variance.
Great latitude in specifying Z.
e.g., if $Z = Y$, then variance depends on response.
e.g., Z could be same as X or a subset of X.
With normal errors assumed, can derive a score test that has approximate $\chi^2(g)$ under H_0. g = # regressors in Z
$\Rightarrow$ p-value

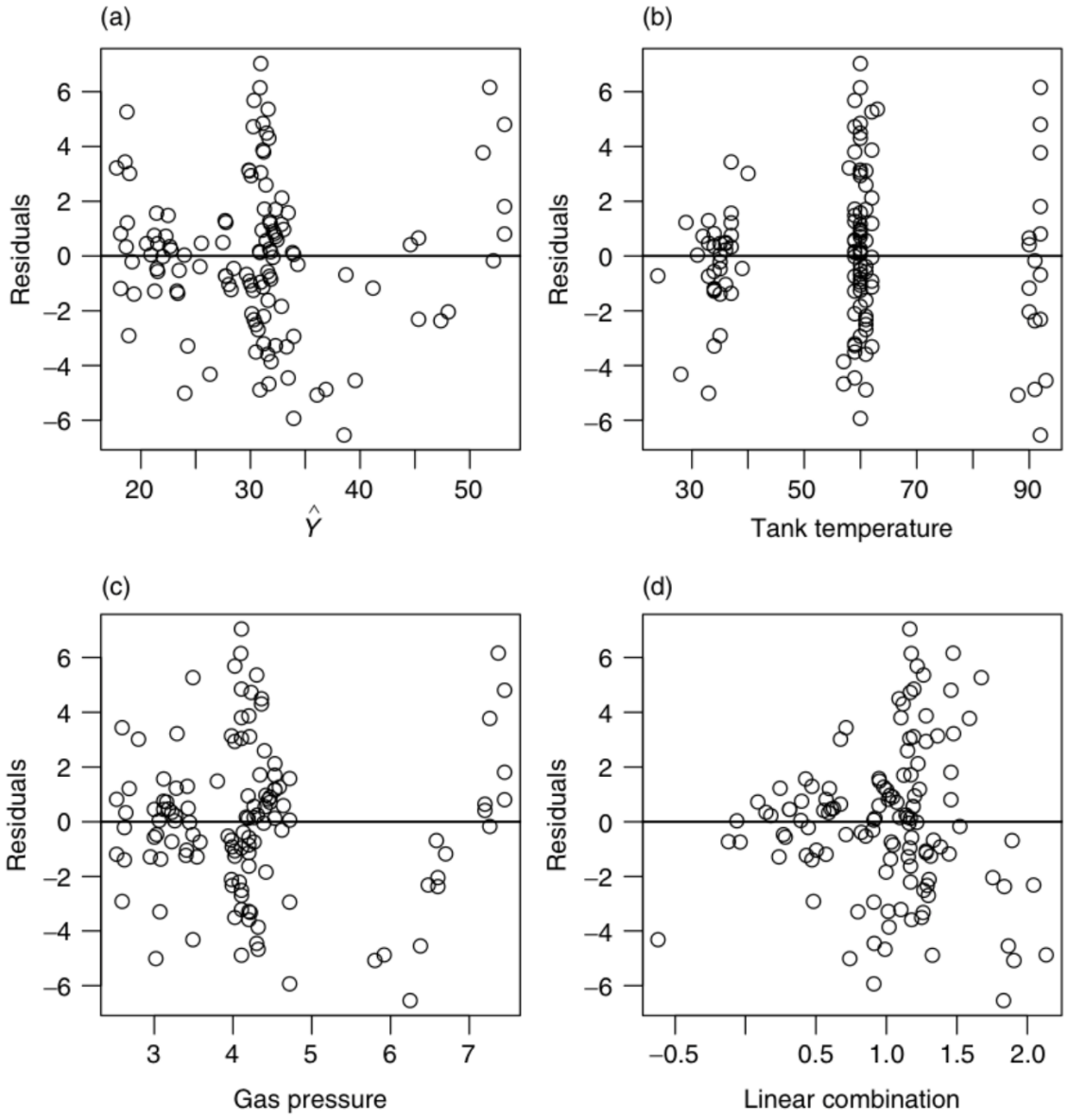


Figure 7.3 Residuals plots for the sniffer data with variance assumed to be constant.

Table 7.4 Score Tests for the Sniffer Data

Choice for Z	*df*	Test stat.	*p*-Value
`GasPres`	1	5.50	.019
`TankTemp`	1	9.71	.002
`TankTemp, GasPres`	2	11.78	.003
`TankTemp, GasTempTankPres, GasPres`	4	13.76	.008
Fitted values	1	4.80	.028

General Correlation Structures

The generalized least squares or GLS model extends WLS one step further, and starts with

$$E(Y|X) = X\beta, \quad \text{Var}(Y|X) = \Sigma,$$

where Σ is an $n \times n$ positive definite symmetric matrix. The WLS model uses $\Sigma = \sigma^2 W^{-1}$ for a diagonal matrix W, and the OLS model uses $\Sigma = \sigma^2 I$. If we have n observations and Σ is completely unknown, then Σ contains $n(n-1)/2$ parameters, which is much larger than the number of observations n. The only hope is to introduce some structure in Σ. Here are some examples.

Compound Symmetry

If all the observations are equally correlated, then

$$\Sigma_{CS} = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix},$$

which has only two parameters ρ and σ^2. Generalized least squares software, such as the `gls` function in the `nlme` package, can be used for estimation.

Autoregressive

This form is generally associated with time series. If data are time ordered and equally spaced, the lag-1 autoregressive covariance structure is

$$\Sigma_{AR} = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho^{n-1} \\ \rho & 1 & \cdots & \rho^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \cdots & 1 \end{pmatrix},$$

which also contains only two parameters.

Block Diagonal

A block diagonal form for Σ can arise if observations are sampled clusters. For example, a study of school performance might sample m children from each of k classrooms. The m children within a classroom may be correlated because they all have the same teacher, but children in different classrooms are independent.

Random Coefficient Models

The random coefficient model, as a special case of mixed models, allows for appropriate inferences. Consider a population regression mean function

$$E(y|\text{loudness} = x) = \beta_0 + \beta_1 x,$$

where subject effects are not allowed. To add them, we hypothesize that each of the subjects may have his or her own slope and intercept. Let $y_{ij}, i = 1, \ldots, 10, j = 1, \ldots, 5$ be the log-response for subject i measured at the jth level of loudness. For the ith subject,

$$E(y_{ij}|\text{loudness} = x, b_{0i}, b_{1i}) = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})\,\text{loudness}_{ij},$$

where b_{0i} and b_{1i} are the deviations from the population intercept and slope for the ith subject, and they are treated as random variables,

$$\begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_0^2 & \tau_{01} \\ \tau_{01} & \tau_1^2 \end{pmatrix} \right).$$

Inferences about (β_0, β_1) concern population behavior. Inferences about (τ_0^2, τ_1^2) concern the variation of the intercepts and slopes between individuals in the population.

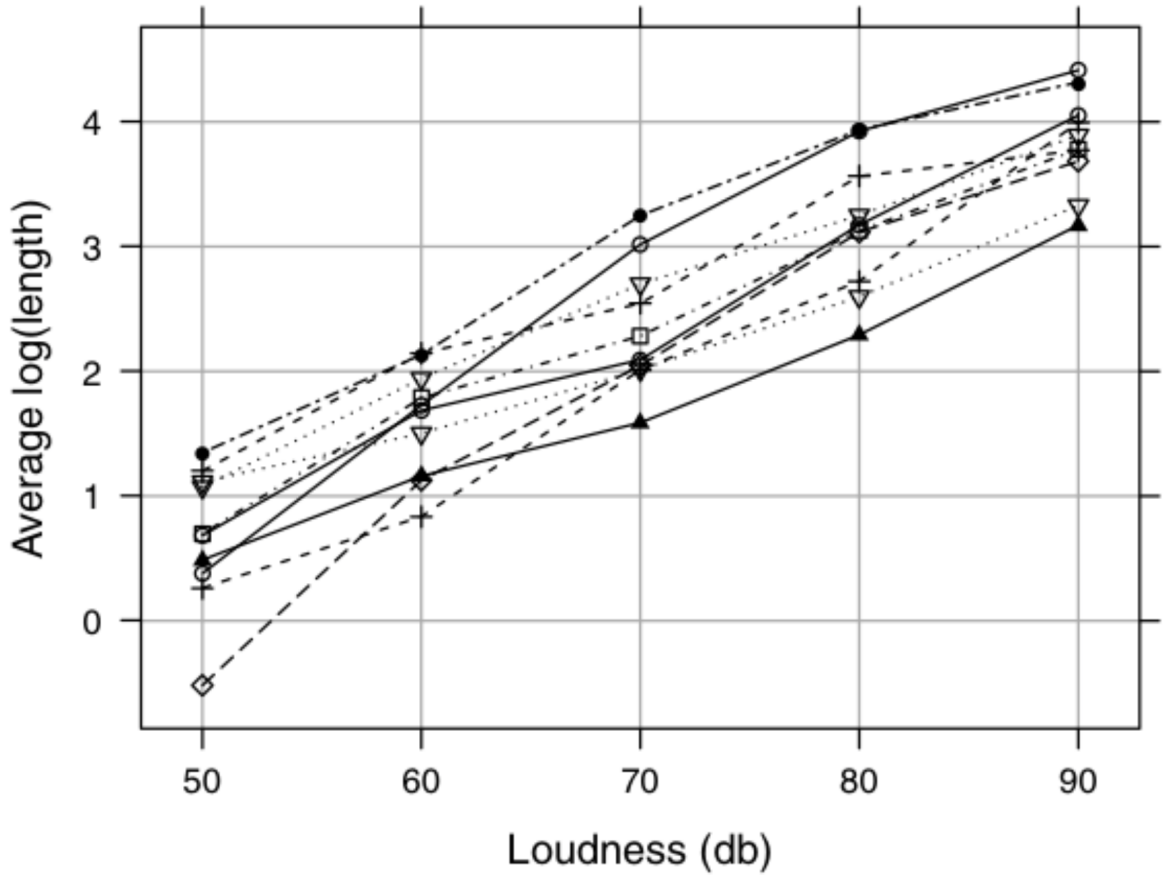


Figure 7.4 Psychophysics example.

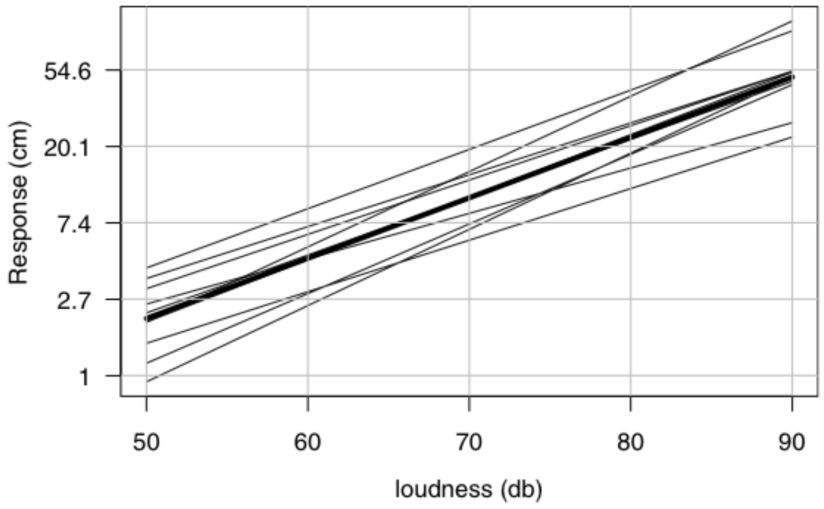


Figure 7.5 Fitted mixed model for the psychophysics example. The thick line is the population estimate, while the thinner lines are predicted lines for individuals.

Variance Stabilizing Transformation

Suppose that the response is strictly positive, and the variance function before transformation is

$$\text{Var}(Y|X = x) = \sigma^2 g(E(Y|X = x)),$$

where $g(E(Y|X = x))$ is a function that is increasing with the value of its argument. For example, if the distribution of $Y|X$ has a Poisson distribution, then $g(E(Y|X = x)) = E(Y|X = x)$.
For distributions in which the mean and variance are functionally related, Scheffe (1959) provides a general theory for determining transformations that can stabilize variance. Table # lists the common variance stabilizing transformations.

Y_T	Comments
$\sqrt{Y}$	Used when Var$(Y\|X) \propto E(Y\|X)$, as for Poisson distributed data. $Y_T = \sqrt{Y} + \sqrt{Y+1}$ can be used if all the counts are small.
$\log(Y)$	Used if Var$(Y\|X) \propto [E(Y\|X)]^2$. In this case, the errors behave like a percentage of the response, $\pm 10\%$, rather than an absolute deviation, ± 10 units.
$1/Y$	The inverse transformation stabilizes variance when Var$(Y\|X) \propto [E(Y\|X)]^4$. It can be appropriate when responses are mostly close to 0, but occasional large values occur.
$\sin^{-1}(\sqrt{Y})$	The arcsine square-root transformation is used if Y is a proportion between 0 and 1, but if can be used more generally if y has a limited range by first transforming Y to the range (0,1), and then applying the transformation.

The Bootstrap

Suppose we have a sample $y_1, \ldots, y_n$ from a particular distribution G, for example a standard normal distribution. What is a confidence interval for the population median?

We can obtain an approximate answer to this question by computer simulation, set up as follows.

1. Obtain a simulated random sample $y_1^*, \ldots, y_n^*$ from the known distribution G.

2. Compute and save the median of the sample in step 1.

3. Repeat steps 1 and 2 a large number of times, say B times. The larger the value of B, the more precise the ultimate answer.

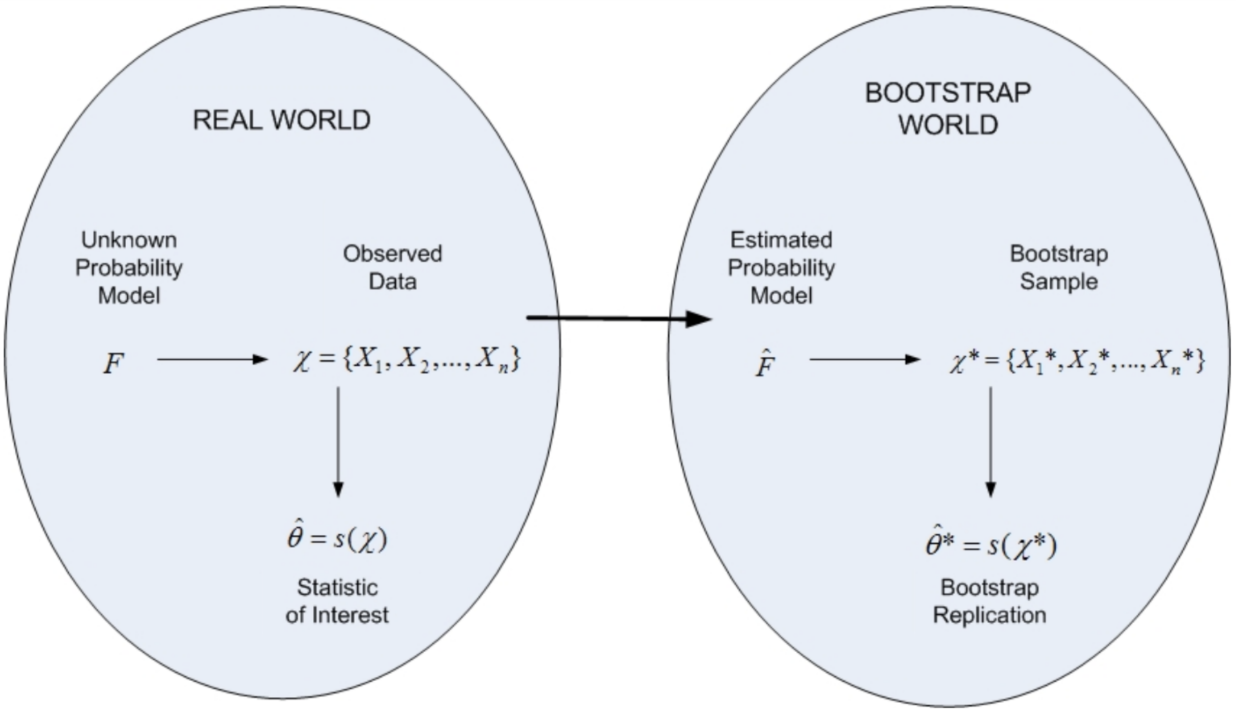
4. If we take $B = 999$, a simple percentile-based 95% confidence interval for the median is the interval between the sample 2.5 and 97.5 percentiles, respectively.

In most interesting problems, G is unknown and so this simulation is not available. Efron (1979) pointed out that the observed data can be used to estimate G, and then we can sample from the estimate $\hat{G}$. The algorithm becomes:

1. Obtain a random sample $y_1^*, \ldots, y_n^*$ from $\hat{G}$ by sampling with replacement from the observed values $y_1, \ldots, y_n$. In particular, the i-th element of the sample y_i^* is equally likely to be any of the original $y_1, \ldots, y_n$. Some of the y_i will appear several times in the random sample, while others will not appear at all.

2. Continue with steps 2-4 of the first algorithm. A test at the 5% level concerning the population median can be rejected if the hypothesized value of the median does not fall in the confidence interval computed in step 4.

Efron called this method the *bootstrap*.



The bootstrap approach (Efron and Tibshirani, 1993).

Regression Inference without Normality

For regression problems, when the sample size is small and the normality assumption does not hold, standard inference methods can be misleading, and in these cases a bootstrap can be used for inference.

Transactions Data

Each branch makes transactions of two types, and for each of the branches we have recorded the number of transactions T_1 and T_2, as well as *Time*, the total number of minutes of labor used by the branch in type 1 and type 2 transactions. The mean response function is

$$E(\text{Time}|T_1, T_2) = \beta_0 + \beta_1 T_1 + \beta_2 T_2$$

possibly with $\beta_0 = 0$ because zero transactions should imply zero time spent. The data are displayed in Figure 1. The marginal response plots in the last row appear to have reasonably linear mean functions; there appear to be a number of branches with no T_1 transactions but many T_2 transactions; and in the plot of *Time* versus T_2, variability appears to increase from left to right.

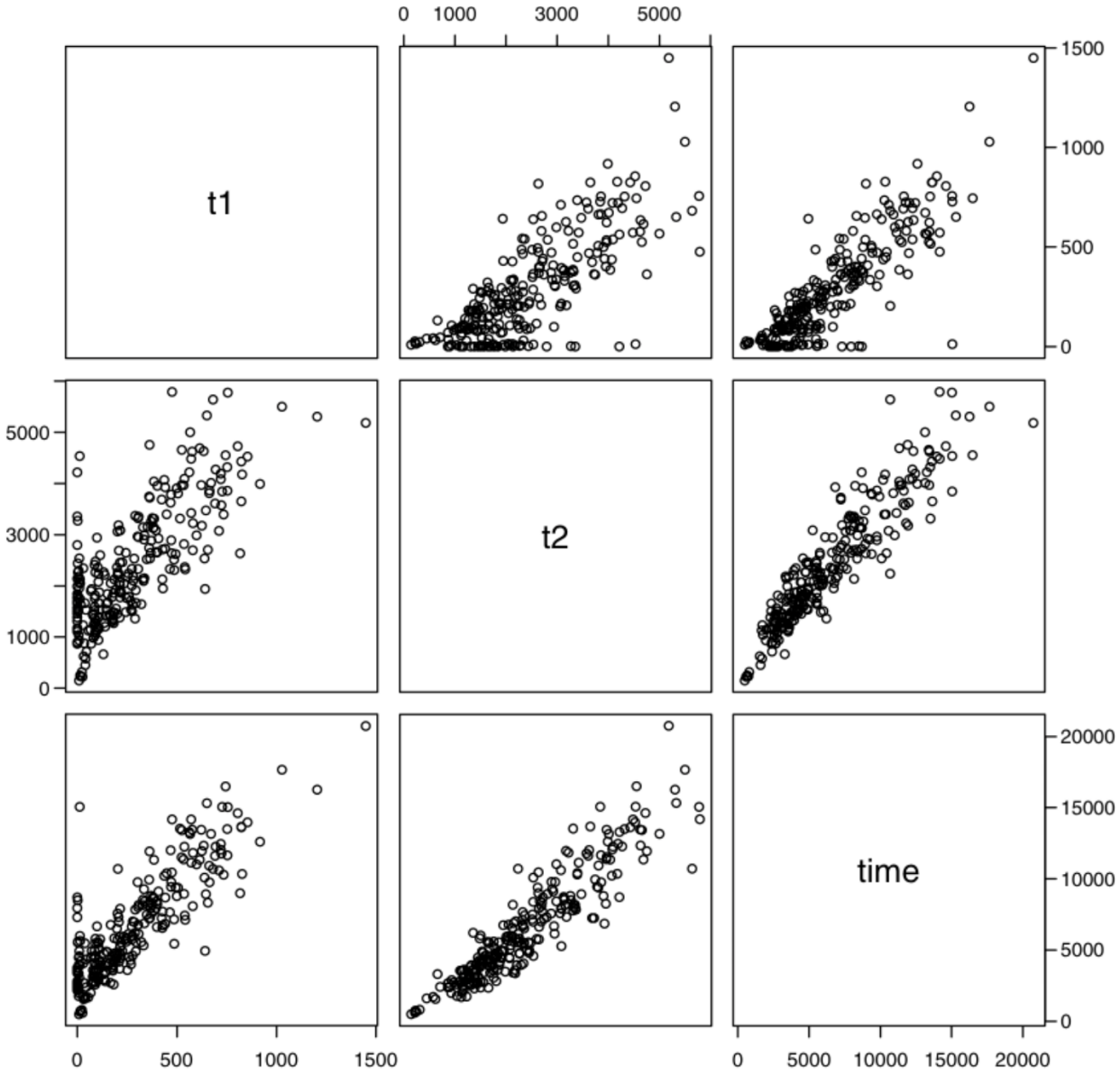


Figure 7.6 Scatterplot matrix for the transactions data.

The errors in this problem probably have a skewed distribution. Oc-

casional transactions take a very long time, but since transaction time is bounded below by zero, there cannot be any really extreme "quick" transactions. Inferences based on normal theory are therefore questionable.

A bootstrap is computed as follows.

1. Number the cases in the dataset from 1 to n. Take a random sample (from $\{(x_i, y_i); i = 1, \ldots, n\}$) with replacement of size n from these case numbers.

2. Create a dataset from the original data, but repeating each row in the dataset the number of times that row was selected in the random sample in step 1.

3. Repeat steps 1 and 2 a large number of times, say, B times.

4. Estimate a 95% confidence interval for each of the estimates by the 2.5 and 97.5 percentiles of the sample of B bootstrap samples.

Table 6 summarizes the percentile bootstrap for the transaction data.

The 95% bootstrap intervals are consistently wider than the corresponding normal intervals, indicating...

Table 7.6 Summary Statistics for Case Bootstrap in the Transactions Data

	OLS	boot	bias	OLS.SE	boot.SE
(Intercept)	144.37	159.20	–14.83	170.54	188.54
t1	5.46	5.51	–0.04	0.43	0.66
t2	2.03	2.02	0.01	0.09	0.15

Table 7.7 95% Confidence Intervals for the Transactions Data

Method	(Intercept)	t1	t2
Normal theory	(–191.47, 480.21)	(4.61, 6.32)	(1.85, 2.22)
Normal with boot SE	(–240.00, 499.08)	(4.12, 6.72)	(1.75, 2.34)
Percentile	(–204.33, 538.92)	(4.20, 6.80)	(1.73, 2.32)
BCa	(–259.44, 487.18)	(3.88, 6.64)	(1.79, 2.38)

Residual Bootstrap

1. Given data (x_i, y_i), $i = 1, \ldots, n$, fit the linear regression model $E(Y|X = x) = \beta' x$ and compute the OLS estimator $\hat{\beta}$, and the residuals, $\hat{e}_i = y_i - \hat{\beta}' x_i$.
2. Randomly sample from the residuals to get a new sample $(e_1^*, \ldots, e_n^*)$, where e_i^* is equally likely to be any of $(\hat{e}_1, \ldots, \hat{e}_n)$. A modified definition of residuals can be used that may slightly improve performance by correcting for unequal variances of the residuals; see Davison and Hinkley (1997, p. 262).
3. Create a bootstrap response with elements $y_i^* = \hat{\beta}' x_i + e_i^*$. Compute the regression of the bootstrap response on X, and get the coefficient estimates or other summary statistic of interest.
4. Repeat steps 2 and 3 B times. Confidence intervals are obtained as with the case bootstrap.