Chapter 4: Interpretation of Main Effects

The computations that are done in multiple linear regression, including drawing graphs, creation of terms, fitting models, and performing tests, will be similar in most problems. Interpreting the results, however, may differ by problems, even if the outline of the analysis is the same. Many issues play into drawing conclusions, and some of them are discussed in this chapter.

1 Understanding parameter estimates

Parameters in mean functions have units attached to them. For example, the fitted mean function for the fuel consumption data is

$$R(Fuel|X) = 154.19 - 4.23tax + 0.47Dlic - 6.14Income + 18.54\log(Miles).$$

1.1 Rate of change

The usual interpretation of an estimated coefficient is as a rate of change: increasing Tax rate by one cent should decrease consumption, all other factors being held constant, by about 4.23 gallons per person. This assumes that a predictor can in fact be changed without affecting the other terms in the mean function and that the available data will apply when the predictor is so changed. Other coef-

ficients can be interpreted similarly.

1.2 Signs of estimates

The sign of a parameter estimate indicates the direction of the relationship between the term and the response. In multiple regression, if the terms are correlated, the sign of a coefficient may change depending on the other terms in the model.

1.3 Interpretation depends on other terms in the mean function

The value of a parameter estimate not only depends on the other terms in a mean function but it can also change if the other terms are replaced by linear combinations of the terms.

Berkeley Guidance Study Data from the Berkeley Guidance Study on the growth of boys and girls are given in Problem 3.1. Here we will view Soma as the re-

sponse, and WT2, WT9 and WT18 as predictors. Figure 1 shows the scatterplot matrix of the data. Since each of the two-dimensional plots appear to be well summarized by a straight-line mean function, the regression of the response on the original predictors without transformation is likely to be appropriate. In addition, we consider the following linear combinations of the predictors:

- WT2=Weight at age 2
- DW9=WT9-WT2=Weight gain from age 2 to 9
- DW18=WT18-WT9=Weight gain from age 9 to 18

Table 1 shows the resulting coefficient estimators from two models.

Model 1 leads to the unexpected conclusion that heavier girls at age two may tend to be thinner, have lower expected somatotype, at age 18. The

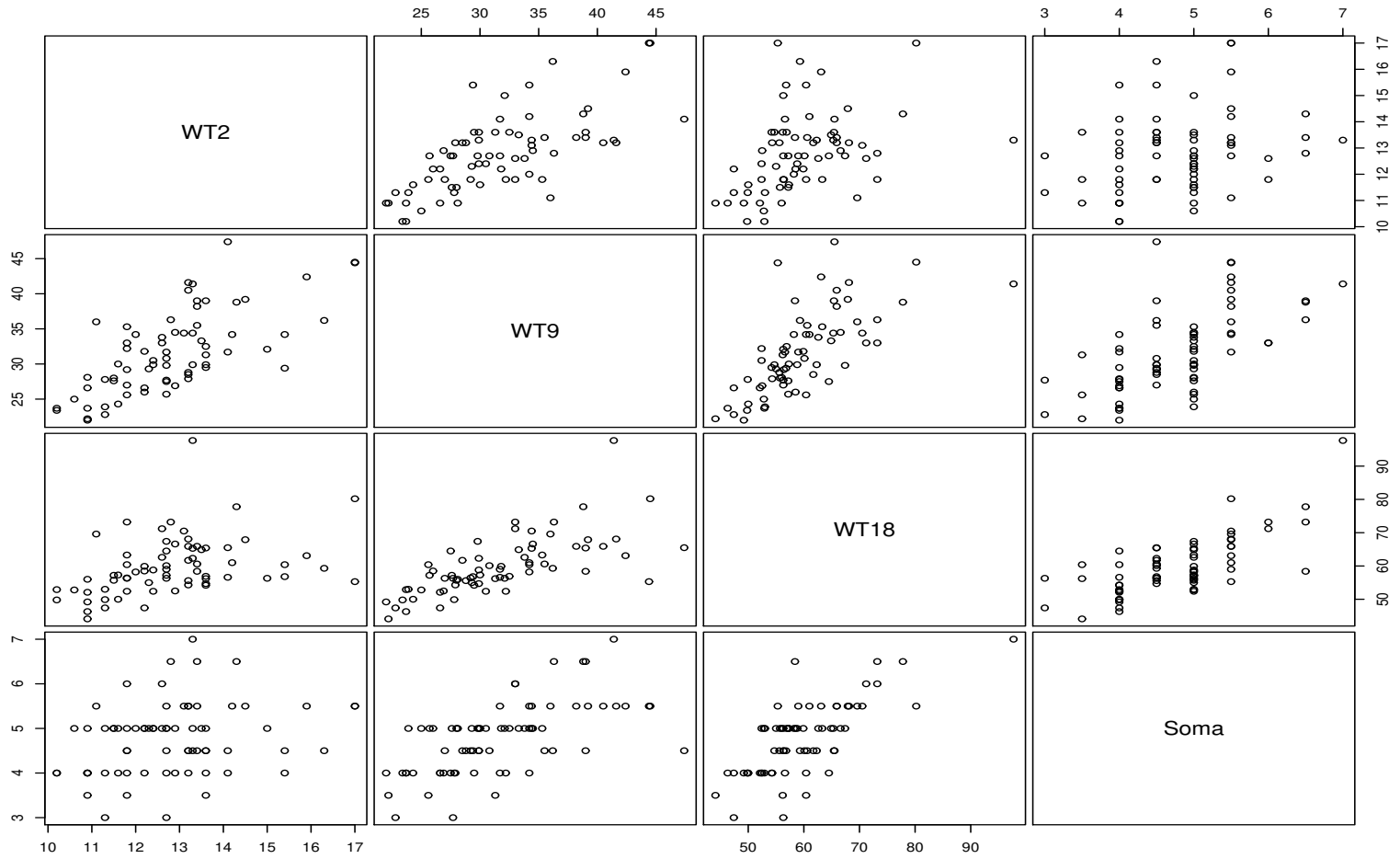

Figure 1: Scatterplot matrix for the girls in the Berkeley Guidance Study.

Table 1: Regression of Soma on different combinations of three weight variables for the $n = 70$ girls in the Berkeley Guidance Study.

Term	Model 1	Model 2	Model 3
(intercept)	1.5921	1.5921	1.5921
WT2	-0.1156	-0.0111	-0.1156
WT9	0.0562		0.0562
WT18	0.0483		0.0483
DW9		0.1046	NA
DW18		0.0483	NA

t-statistic for testing this coefficient equal to zero has a significance level of about 0.06. The sign, and the weak significance may be due to the correlations between the terms.

Model 2 is much better in this respect. The estimate is close to 0, and the corresponding t-statistics is -0.21. Thus, we can conclude that the effect of WT2 is negligible.

1.4 Rank deficient and over-parameterized mean functions

Model 3 in Table 1 gives the estimates produced when we tried to fit using an intercept and the five terms WT2, WT9, WT18, DW9, and DW18. The program set some coefficients to “NA”, a code for a missing value, as the predictors are linearly dependent. The maximum number of linearly independent terms that could be included in a mean function is called the rank of the data matrix.

Mean functions that are over-parameterized occur most often in designed experiments. Suppose that a unit is assigned to one of three treatment groups, and let X_1, X_2 and X_3 be the indicator variable of the three groups, respectively. We therefore cannot fit the model

$$E(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

because $X_1 + X_2 + X_3 = 1$. To fit a model, we must do something else. The options are

- Place a constraint like $\beta_1 + \beta_2 + \beta_3 = 0$ on the parameters.
- Exclude one of the X_j from the model.
- Leave out an explicit intercept.

All of these options will in some sense be equivalent, since the same R^2, σ^2 and overall F-test and predictions will result.

1.5 Tests

Even if the fitted model were correct and errors were normally distributed, tests and confidence statements for parameters are difficult to interpret because correlations among the terms lead to a multiplicity of possible tests. Sometimes, tests of effects adjusted for other variables are clearly desirable, such as in assessing a treatment effect after adjusting for other variables to reduce variability.

Suppose that the true mean function is

$$E(Y|X_1 = x_1, X_2 = x_2) = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1'\boldsymbol{x}_1 + \boldsymbol{\beta}_2'\boldsymbol{x}_2$$

but we want to fit a mean function with X_1 only. The mean function for $Y|X_1$ is obtained by averaging over X_2,

$$\begin{aligned} E(Y|X_1 = \boldsymbol{x}_1) &= E(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1'X_1 + \boldsymbol{\beta}_2'X_2 + \boldsymbol{e}|X_1 = \boldsymbol{x}_1 \\ &= \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1'\boldsymbol{x}_1 + \boldsymbol{\beta}_2'E(X_2|X_1 = \boldsymbol{x}_1). \end{aligned}$$

We cannot, in general, simply drop a set of terms from a correct mean function, but we need to substitute the conditional expectation of the terms dropped given the terms that remain in the mean function.

Variances are also affected when terms are dropped,

$$
\begin{aligned}
&\mathsf{Var}(Y|X_1 = \boldsymbol{x}_1) \\
&= \mathsf{Var}(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1' X_1 + \boldsymbol{\beta}_2' X_2 + \boldsymbol{e}|X_1 = \boldsymbol{x}_1) \\
&= \sigma^2 + \boldsymbol{\beta}_2' \mathsf{Var}(X_2|X_1 = \boldsymbol{x}_1)\boldsymbol{\beta}_2.
\end{aligned}
$$

If we starts with the simple regression mean function,

$$E(Y|X = x) = \beta_0 + \beta_1 x.$$

A useful way to interpret the coefficient β_1 is as the first derivative of the mean function with respect to x,

$$\frac{dE(Y|X = x)}{dx} = \beta_1.$$

When the predictor is replaced by $\log(x)$, the mean function $\beta_0 + \beta_1 \log(x)$ is no longer a straight line, but rather it is a curve. The tangent at the point $x > 0$ is

$$\frac{dE(Y|X=x)}{dx} = \frac{\beta_1}{x}.$$

When the response is in log scale, we have $E(\log(Y)|Y = x) = \beta_0 + \beta_1 x$ and $E(Y|X = x) \approx e^{\beta_0+\beta_1 x}$.

Differentiating it gives

$$\frac{dE(Y|X=x)}{dx} = \beta_1 E(Y|X=x).$$

2 Experimentation versus Observation

There are fundamentally two types of predictors that are used in a regression analysis, *experimental and observational*. Experimental predictors have values that are under the control of the experimenter, while for observational predictors, the values are observed rather than set.

Consider, for example, a hypothetical study of factors determining the yield of a certain crop. Ex-

perimental variables might include the amount and type of fertilizers used, and the space of plants. Observational predictors might include characteristics of the plots in the study, such as soil fertility and weather variables.

The primary difference between experimental and observational predictors is in the inferences we can make. From experimental data, we can often infer causation. If we assign the level of fer-

tilizer to plots, usually on the basis of a randomization scheme, and observe differences due to levels of fertilizer, we can infer that the fertilizer is causing the difference. Observational predictors allow weaker inferences. We might say that weather variables are associated with yield, but the causal link is not available for variables that are not under the experimenter's control. Some experimental designs, including those that use random-

ization, are constructed so that the effects of observational factors can be ignored or used in analysis of covariance.

3 Computationally Intensive Methods

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 $\widehat{G}$. The algorithm becomes:

1. Obtain a random sample $y_1^*, \ldots, y_n^*$ from $\widehat{G}$ 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 popu-

lation 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.

3.1 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 mislead-

ing, 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 transac-

tions. The mean response function is

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

possibly with $\beta_0 = 0$ because zero transactions should imply zero time spent. The data are displayed in Figure 2. 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, variabil-

ity appears to increase from left to right.

The errors in this problem probably have a skewed distribution. Occasional 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

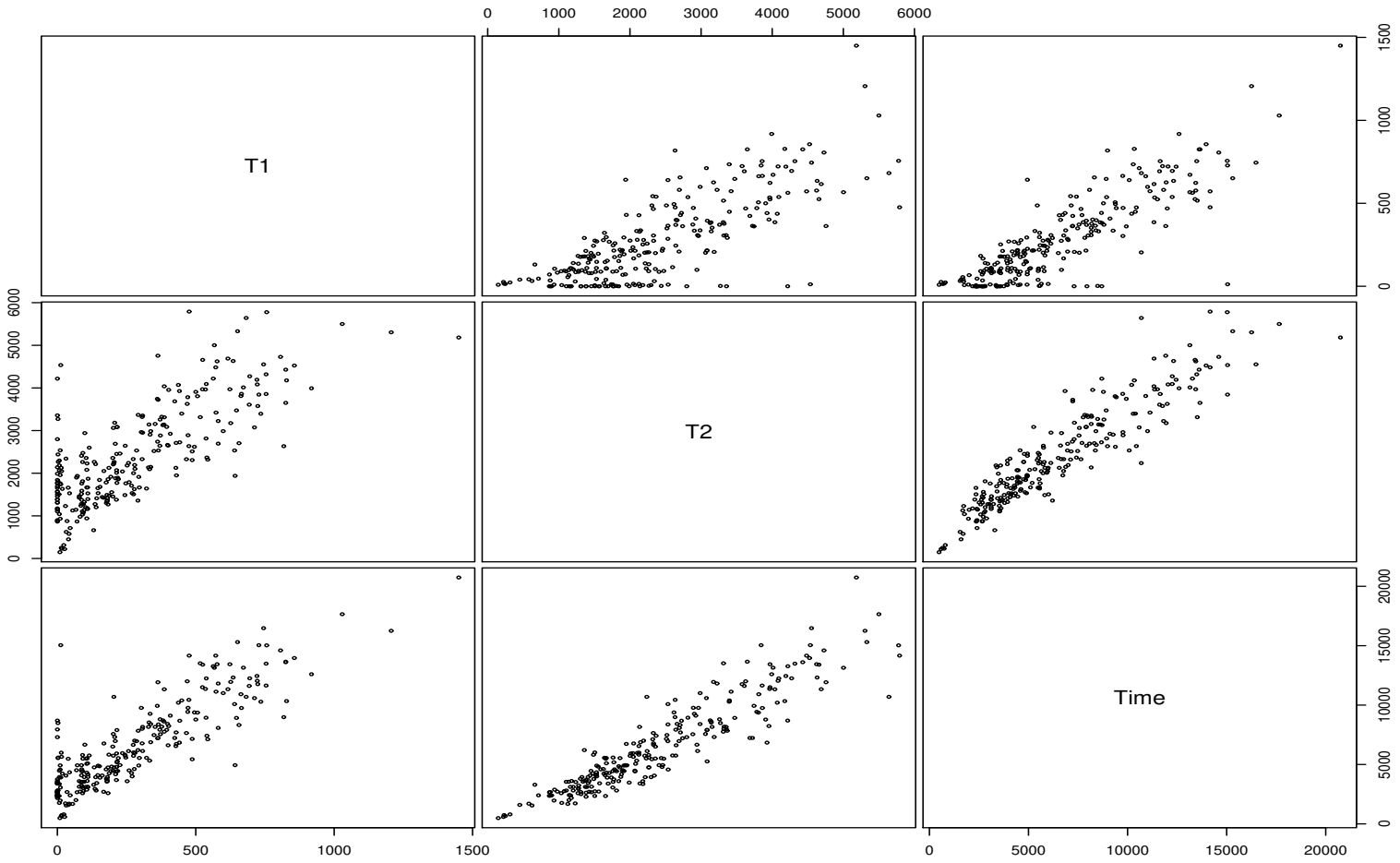

Figure 2: Scatterplot matrix for the transactions data.

n. Take a random sample 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

Table 2: Summary for $B = 999$ case bootstraps for the transactions data.

	Normal Theory			Bootstrap		
	Estimate	Lower	Upper	Estimate	Lower	Upper
Intercept	144.37	-191.47	480.21	136.09	-254.73	523.36
T_1	5.46	4.61	6.32	5.48	4.08	6.77
T_2	2.03	1.85	2.22	2.04	1.74	2.36

the estimates by the 2.5 and 97.5 percentiles of the sample of B bootstrap samples.

Table 2 summarizes the percentile bootstrap for the transaction data.

The 95% bootstrap intervals are consistently wider than the corresponding normal intervals, indicating that the normal-theory confidence intervals are probably overly optimistic.

3.2 Nonlinear functions of parameters

One of the important uses of the bootstrap is to get estimates of error variability in problems where standard theory is either missing, or, equally of-

ten, unknown to analyst. Suppose, for example, we wanted to get a confidence interval for the ratio β_1/β_2 in the transactions data. The point estimate for this ratio is just $\widehat{\beta}_1/\widehat{\beta}_2$, but we will not learn how to get a normal-theory confidence interval for a nonlinear function of parameters like this until Chapter 6.

Using bootstrap, this computation is easy: just compute the ratio in each of the bootstrap samples

and then use the percentiles of the bootstrap distribution to get the confidence interval. For these data, the point estimate is 2.68 with 95% confidence interval from 1.76 to 3.86.