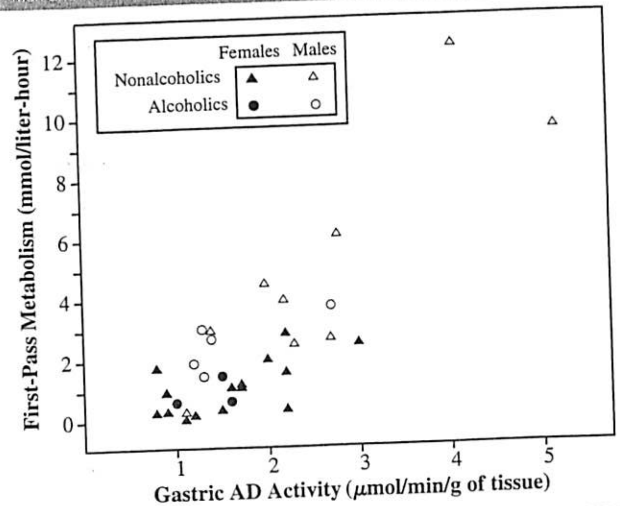
DISPLAY 11.2 First-pass metabolism and gastric alcohol dehydrogenase activity in alcoholic and nonalcoholic men and women



concentration after intravenous administration minus the concentration after oral administration—provides a measure of the "first-pass metabolism" in the stomach. In addition, gastric alcohol dehydrogenase (AD) activity (activity of the key enzyme) was measured in mucus samples taken from the stomach linings. The data are plotted in Display 11.2.

Several questions arise. Do levels of first-pass metabolism differ between men and women? Can the differences be explained by postulating that men have more dehydrogenase activity in their stomachs? Are the answers to these questions complicated by an alcoholism effect?

Statistical Conclusion

The following inferences pertain only to individuals with gastric AD activity levels between 0.8 and 3.0 μmol/min/g. No reliable model could be determined for values greater than 3.0. There was no evidence from these data that alcoholism was related to first-pass metabolism in any way (p-value = 0.93, from an F-test for significance of alcoholism and its interaction with gastric activity and sex.) Convincing evidence exists that first-pass metabolism was larger for males than for females overall (two-sided p-value = 0.0002, from a rank-sum test) and that gastric AD activity was larger for males than for females (two-sided p-value = 0.07 from a rank-sum test). Males had higher first-pass metabolism than females even after accounting for differences in gastric AD activity (two-sided p-value = 0.0003 from a t-test for

equality of male and female slopes when both intercepts are zero). For a given level of gastric dehydrogenase activity, the mean first-pass alcohol metabolism for men is estimated to be 2.20 times as large as the mean first-pass alcohol metabolism for women (approximate 95% confidence interval from 1.37 to 3.04).

Scope of Inference

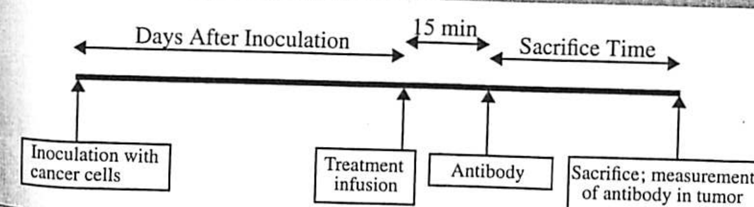
Because the subjects were volunteers, no inference to a larger population is justified. The inference that men and women do have different first-pass metabolism is greatly strengthened, however, by the existence of a physical explanation for the difference. The conclusions about the relationship between first-pass metabolism, gastric AD dehydrogenase activity, and sex are restricted to individuals whose gastric AD activity is less than 3. The sparseness of data for individuals with greater gastric AD activity levels prevents any resolution of the answers in the wider range.

11.1.2 The Blood–Brain Barrier—A Controlled Experiment

The human brain is protected from bacteria and toxins, which course through the bloodstream, by a single layer of cells called the *blood–brain barrier*. This barrier normally allows only a few substances, including some medications, to reach the brain. Because chemicals used to treat brain cancer have such large molecular size, they cannot pass through the barrier to attack tumor cells. At the Oregon Health Sciences University, Dr. E. A. Neuwelt developed a method of disrupting the barrier by infusing a solution of concentrated sugars.

As a test of the disruption mechanism, researchers conducted a study on rats, which possess a similar barrier. (Data from P. Barnett et al., "Differential Permeability and Quantitative MR Imaging of a Human Lung Carcinoma Brain Xenograft in the Nude Rat," *American Journal of Pathology* 146(2) (1995): 436–49.) The rats were inoculated with human lung cancer cells to induce brain tumors. After 9 to 11 days they were infused with either the barrier disruption (BD) solution or, as a control, a normal saline (NS) solution. Fifteen minutes later, the rats received a standard dose of the therapeutic antibody L6-F(ab′)$_2$. After a set time they were sacrificed, and the amounts of antibody in the brain tumor and in normal tissue were measured. The time line for the experiment is shown in Display 11.3. Measurements for the 34 rats are listed in Display 11.4.

DISPLAY 11.3 Time line for blood–brain barrier disruption experiment



DISPLAY 11.4 Response variable, design variables (explanatory variables associated with the assignment of experimental units to groups), and several covariates (explanatory variables not associated with the assignment) for 34 rats in the blood–brain barrier disruption experiment

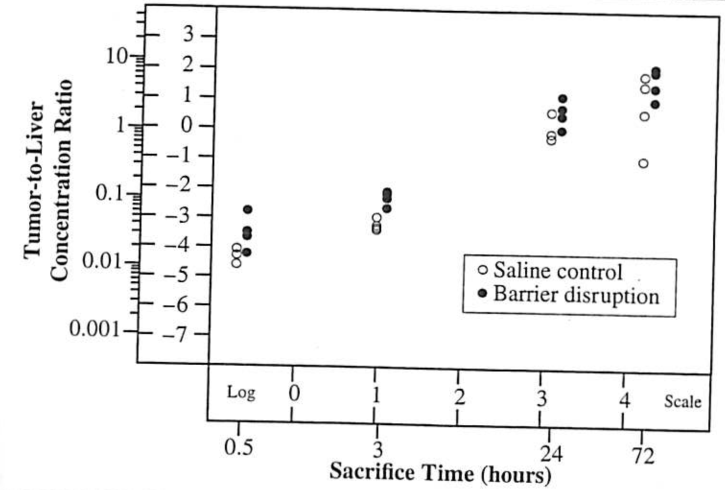
	Response variable	Design variables		Covariates				
Case	Brain tumor count (per gm) / Liver count (per gm)	Sacrifice time (hours)	Treatment	Days post inoculation	Sex	Initial weight (grams)	Weight loss (grams)	Tumor weight (10^{-4} grams)
1	41081/ 1456164	0.5	BD	10	F	239	5.9	221
2	44286/ 1602171	0.5	BD	10	F	225	4.0	246
3	102926/ 1601936	0.5	BD	10	F	224	−4.9	61
4	25927/ 1776411	0.5	BD	10	F	184	9.8	168
5	42643/ 1351184	0.5	BD	10	F	250	6.0	164
6	31342/ 1790863	0.5	NS	10	F	196	7.7	260
7	22815/ 1633386	0.5	NS	10	F	200	0.5	27
8	16629/ 1618757	0.5	NS	10	F	273	4.0	308
9	22315/ 1567602	0.5	NS	10	F	216	2.8	93
10	77961/ 1060057	3	BD	10	F	267	2.6	73
11	73178/ 715581	3	BD	10	F	263	1.1	25
12	76167/ 620145	3	BD	10	F	228	0.0	133
13	123730/ 1068423	3	BD	9	F	261	3.4	203
14	25569/ 721436	3	NS	9	F	253	5.9	159
15	33803/ 1019352	3	NS	10	F	234	0.1	264
16	24512/ 667785	3	NS	10	F	238	0.8	34
17	50545/ 961097	3	NS	9	F	230	7.0	146
18	50690/ 1220677	3	NS	10	F	207	1.5	212
19	84616/ 48815	24	BD	10	F	254	3.9	155
20	55153/ 16885	24	BD	10	M	256	−4.7	190
21	48829/ 22395	24	BD	10	M	247	−2.8	101
22	89454/ 83504	24	BD	11	F	198	4.2	214
23	37928/ 20323	24	NS	10	F	237	2.5	224
24	12816/ 15985	24	NS	10	M	293	3.1	151
25	23734/ 25895	24	NS	10	M	288	9.7	285
26	31097/ 33224	24	NS	11	F	236	5.9	380
27	35395/ 4142	72	BD	11	F	251	4.1	39
28	18270/ 2364	72	BD	10	F	223	4.0	153
29	5625/ 1979	72	BD	10	M	298	12.8	164
30	7497/ 1659	72	BD	10	M	260	7.3	364
31	6250/ 928	72	NS	10	M	272	11.0	484
32	11519/ 2423	72	NS	11	F	226	2.2	168
33	3184/ 1608	72	NS	10	M	249	−4.4	191
34	1334/ 3242	72	NS	10	F	240	6.7	159

Since the amount of the antibody in normal tissue indicates how much of it the rat actually received, a key measure of the effectiveness of transmission across the blood–brain barrier is the ratio of the antibody concentration in the brain tumor

to the antibody concentration in normal tissue outside of the brain. The brain tumor concentration divided by the liver concentration is a measure of the amount of the antibody that reached the brain relative to the amount of it that reached other parts of the body. This is the response variable: both the numerator and denominator of this ratio are listed in Display 11.4. The explanatory variables in the table comprise two categories: *design variables* are those that describe manipulation by the researcher; *covariates* are those measuring characteristics of the subjects that were not controllable by the researcher.

Was the antibody concentration in the tumor increased by the use of the blood–brain barrier disruption infusion? If so, by how much? Do the answers to these two questions depend on the length of time after the infusion (from 1/2 to 72 hours)? What is the effect of treatment on antibody concentration after weight loss, total tumor weight, and other covariates are accounted for? A coded scatterplot relating to the major questions is shown in Display 11.5.

DISPLAY 11.5 Log-log scatterplot of the ratio of antibody concentration in brain tumor to antibody concentration in liver versus sacrifice time, for 17 rats given the barrier disruption infusion and for 17 rats given a saline (control) infusion



Statistical Conclusion

The median antibody concentration in the tumor (relative to that in the liver) was estimated to be 2.22 times as much for rats receiving the barrier disruption infusion than for those receiving the control infusion (95% confidence interval, from 1.56 to

3.15 times as much). This multiplicative effect appears to be constant between 1/2 and 72 hours after the infusion (the p-value for a test of interaction between treatment and sacrifice time is 0.92, from an F-test on 3 and 26 degrees of freedom).

Scope of Inference

One hitch in this study is that randomization was not used to assign rats to treatment groups. This oversight raises the possibility that the estimated relationships might be related to confounding variables over which the experimenter exercised no control. Including the measured covariates in the model helps alleviate some potential con-concern, and the results appear not to have been affected by these founding variables. Nevertheless, causal implications can only be justified on the tenuous assumption that the assignment method used was as effect-neutral as a random assignment would have been.

11.2 RESIDUAL PLOTS

Faced with analyzing data sets like those involved in the blood–brain barrier and alcohol metabolism studies, a researcher must seek good-fitting models for answering the questions of interest, bearing in mind the model assumptions required for least squares tools, the robustness of the tools against violations of the assumptions, and the sensitivity of these tools to outliers. Since model-building efforts are wasted if the analyst fails to detect problems with nonconstant variance and outliers early on, it is wise to postpone detailed model fitting until after outliers and transformation have been thoroughly considered.

Much can be resolved from initial scatterplots and inspection of the data, but it is almost always worthwhile to obtain the finer picture provided by a residual plot. Creating this plot involves fitting some model in order to get residuals. On the basis of the scatterplots, the analyst can choose some tentative model or models and conduct residual analysis on these, recognizing that further modeling will follow.

Selecting a Tentative Model

A tentative model is selected with three general objectives in mind: The model should contain parameters whose values answer the questions of interest in a straightforward manner; it should include potentially confounding variables; and it should include features that capture important relationships found in the initial graphical analysis.

It is disadvantageous to start with either too many or too few explanatory variables in the tentative model. With too few, outliers may appear simply because of omitted relationships. With too many (lots of interactions and quadratic terms, for example), the analyst risks overfitting the data—causing real outliers to be explained away by complex, but meaningless, structural relationships. Overfitting becomes less of a problem when the sample sizes are substantially larger than the number of model parameters.

For large sample sizes, therefore, the initial tentative model for residual analysis can err on the side of being rich, including potential model terms that may not be retained in the end. For small sample sizes, several tentative models may be needed for residual analysis; and the data analyst must guard against including terms whose significance hinges on one or two observations. As evident in the strategy for data analysis laid out in Display 9.9, the process of trying a model and plotting residuals is often repeated until a suitable inferential model is determined.

Example—Preliminary Steps in the Analysis of the Blood–Brain Barrier Data

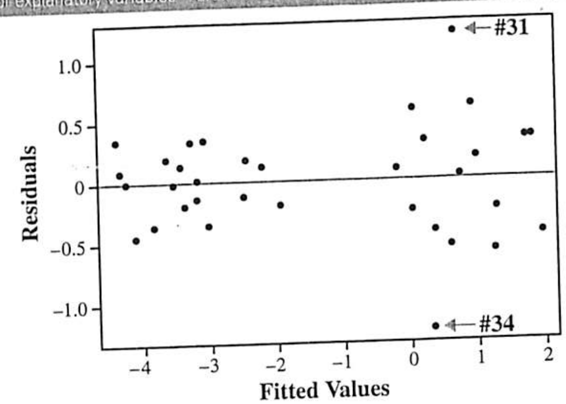
The coded scatterplot in Display 11.5 is a good starting point for the analysis. Apparently, the disruption solution does allow more antibody to reach the brain than the control solution does; this effect is about the same for all sacrifice times (time between antibody treatment and sacrifice); an increasing proportion of antibody reaches the brain with increasing time after infusion; and this increasing relationship appears to be slightly nonlinear. A matrix of scatterplots and a correlation matrix (an array showing the sample correlation coefficients for all possible pairs of variables), which are not shown here, indicate further that the covariates—days after inoculation, initial weight, and sex of the rat—are associated with the response. These covariates are also related to the treatment given. (Recall that randomization was not used.) In particular, rats treated at longer days after inoculation were also assigned to the longer sacrifice times. Furthermore, all male rats were assigned to the longer sacrifice times.

This initial investigation suggests the following tentative regression model (using the shorthand model specification of Section 9.3.5):

$$\mu\{antibody \mid SAC, TREAT, DAYS, FEM, weight, loss, tumor\}$$
$$= SAC + TREAT + (SAC \times TREAT) + DAYS + FEM + weight + loss + tumor,$$

where *antibody* is the logarithm of the ratio of antibody in the brain tumor to that in the liver. *SAC* is the sacrifice time factor with four levels; *TREAT* is treatment, with two levels; *DAYS* is days after inoculation, with three levels; and *FEM* is sex, with two levels. *Weight*, *loss*, and *tumor* are the initial weight, weight loss, and tumor weight variables. Display 11.5 shows a strong linear effect of log sacrifice time on the response, but some additional curvature may be present as well. To avoid mismodeling the effect of sacrifice time at the start, it is treated as a factor with four levels. Similarly, the coded scatterplot suggests that the difference between the two treatments may be greater for the shorter sacrifice times than for the longer ones. Consequently, the sacrifice time by treatment interaction terms are included in the tentative model. Although more terms may be added to this model later, it captures the most prominent features of the scatterplot. Display 11.6 shows the plot of residuals versus the fitted values from the regression model. (*Note:* Even if prior experience or initial inspection had not led the researchers to consider the logarithms of the response, the coded scatterplot and residual plot would have revealed that the variability increases with increasing response, leading them to the same consideration.)

DISPLAY 11.6 Scatterplot of residuals versus fitted values from the fit of the logged response on a rich model for explanatory variables—brain-barrier data



The residual plot in Display 11.6 exemplifies the ambiguity that can arise with small data sets. Is there a funnel-shaped pattern, or is the apparent funnel only due to a few outliers? The usual course of action consists of three steps:

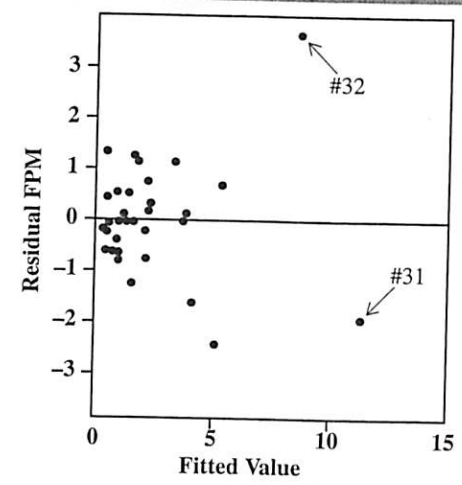
1. Examine the outliers for recording error or contamination.
2. Check whether a standard transformation resolves the problem.
3. If neither of these steps works, examine the outliers more carefully to see whether they influence the conclusions (following the strategy suggested in Section 11.3).

The residual plot is based on a response that has already been transformed into its logarithm. A reciprocal transformation corrects more pronounced funnel-shaped patterns than does the log. Here, however, it does not help, and there is no suggestion of a recording error. Consequently, the analyst must proceed with further model fitting, paying careful attention to the roles of observations 31 and 34.

Example—Preliminary Steps in the Analysis of the Alcohol Metabolism Data

Refer to the coded scatterplot in Display 11.2. The next step is to examine a residual plot for outliers and to assess the need for transformation. The plot in Display 11.7 is a residual plot from the regression of first-pass metabolism on gastric AD activity (*gast*), an indicator variable for females (*fem*), an indicator for alcoholics (*alco*), and the interaction terms *gast* × *fem*, *fem* × *alco*, *gast* × *alco*, and *gast* × *fem* × *alco*. (The last term is a *three-factor interaction* term, formed as the product of three explanatory variables.)

DISPLAY 11.7 Residual plot from the regression of first-pass metabolism on gastric activity, sex indicator, alcoholism indicator, and all two- and three-factor interactions



The plot draws attention to two observations: one that has a considerably larger residual than the rest and one that has a fitted value quite a bit larger than the rest. These are cases 31 and 32, and they appear in the coded scatterplot of Display 11.2 in the upper right-hand corner, separated from the rest of the points. There appears to be a downward trend in the residual plot, excluding cases 31 and 32. This could reflect a model that is heavily influenced by one or two observations and consequently does not fit the bulk of the observations well.

11.3 A STRATEGY FOR DEALING WITH INFLUENTIAL OBSERVATIONS

Least squares regression analysis is not resistant to outliers. One or two observations can strongly influence the analysis, to the point where the answers to the questions of interest change when these isolated cases are excluded. Although any influential observation that comes from a population other than the one under investigation should be removed, removing an observation simply because it is influential is not justified. In any circumstance, it is unwise to state conclusions that hinge on one or two data points. Such a statistical study should be considered extremely fragile.

There are two approaches for dealing with excessively influential observations in regression analysis. One is to use a robust and resistant regression procedure. The other is to use least squares but to examine outliers and influence closely to

The meaning of the partial residuals is clearer in the first version, but the calculation is often more straightforward with the second. Because of this calculating formula, partial residual plots are sometimes referred to as *component plus residuals plots*.

Notes About Partial Residuals

When Should Partial Residual Plots Be Used? Partial residuals are primarily useful when analytical interest centers on one explanatory variable whose effect is expected to be small relative to the effects of others. They are also useful when uncertainty exists about a particular explanatory variable that needs to be modeled carefully or when the underlying explanation for why an observation is influential on the estimate of a single coefficient needs to be understood.

Augmented Partial Residuals. Rather than using $\beta_2 lgest$ as an approximation to $f(lgest)$, some statisticians prefer to use $\beta_2 lgest + \beta_3 lgest^2$. Here, partial residuals are obtained just as in the preceding algorithms, except that $lgest^2$ is also included as an explanatory variable in step 1. (In step 2 of the component-plus-residual version, $pres = res + \hat{\beta}_2 lgest + \hat{\beta}_3 lgest^2$.) If they are equally convenient to use, the augmented partial residuals are preferred. In many cases, however, the difference between the partial residual and the augmented partial residual is slight.

Example—Blood–Brain Barrier

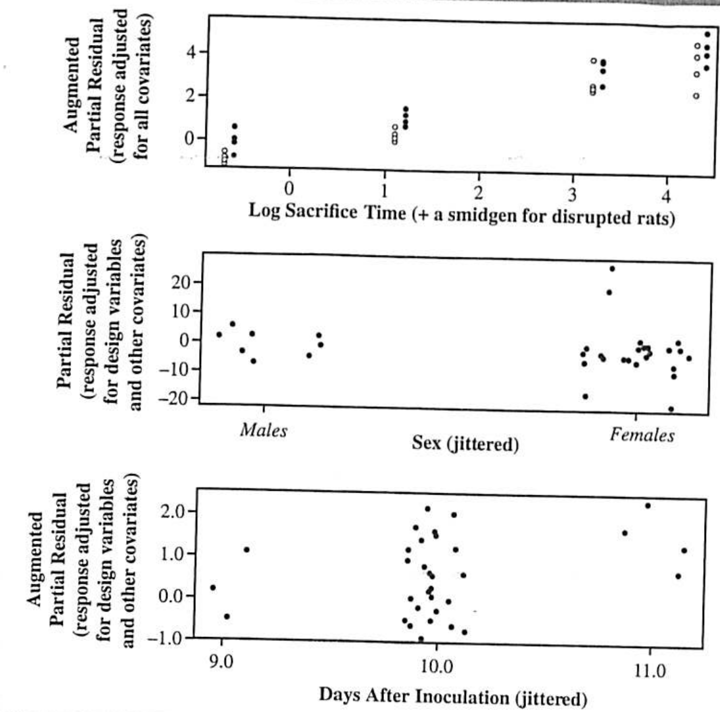
The residual plot in Display 11.6 indicated some potential outliers, but further investigation does not show that these points are influential in determining the structure of the model or in answering the questions of interest (see Exercise 11.18).

The key explanatory variables are the indicator variable for whether the rat received the disruption infusion or the control infusion, the length of time after infusion that the rat was sacrificed, and the interaction of these. The additional covariates should be given a chance to be included in the model, for two reasons. First (and most importantly), since randomization was not used, it behooves the researchers to demonstrate that the differences in treatment effects cannot be explained by differences in the types of rats that received the various treatments. Second, even if randomization had been used, including important covariates can yield higher resolution. If the covariates have some additional association with the response, smaller standard errors and more powerful tests should result from their inclusion.

Among the covariates, sex and days after inoculation are associated with both the response and the design variables. To some extent the effects of these variables are confounded, since their effects on the response cannot be separated. On the other hand, the effects of the design variables can be examined after the covariates are accounted for, and the effects of the covariates can be examined after the design variables are accounted for. This is shown graphically in the partial residual plots of Display 11.16.

The top scatterplot indicates that the relationship between the response and the design variables (sacrifice time and treatment) is much the same when the

DISPLAY 11.16 Some partial residual plots for the blood–brain barrier data, with log of antibody concentration ratio (brain tumor-to-liver) as response



effects of the covariates are included as when they are ignored (Display 11.5). The lower two plots show that, after the effects of the design variables are accounted for, little evidence exists of a sex effect, although slight visual evidence exists of a days-after-inoculation effect.

This conclusion is further investigated through model fitting. A search through possible models that contain covariates shows that sex and days after inoculation (treated as a factor) are the only ones associated with the response. When the design variables are included as well, three conclusions are supported:

1. The covariates are not significant when the design variables are also included in the model.
2. The design variables *are* significant when the covariates are also included in the model.

DISPLAY 11.17 Results from the regression of log ratio of antibody concentration (brain tumor-to-liver) on sacrifice time (treated as a factor) and treatment

Variable	Estimate	Standard error	t-statistic	Two-sided p-value
Constant	−4.302	0.205	−21.01	<0.0001
Indicator for time = 3	1.134	0.252	4.50	0.0001
Indicator for time = 24	4.257	0.259	16.43	<0.0001
Indicator for time = 72	5.154	0.259	19.89	<0.0001
Indicator for treatment = BD	0.797	0.183	4.35	0.0002

3. The conclusions regarding the design variables depend very little on whether the covariates are in the model.

These results suggest that the conclusions can be based satisfactorily on the model without the covariates.

Since the effect of log sacrifice time is not linear (and since the addition of a quadratic term does not remedy the lack-of-fit), sacrifice time is treated as a factor with four levels. Therefore, the final model used to estimate the treatment effect has the following terms: *TIME*+*TREAT*. The estimates and standard errors are shown in Display 11.17. The coefficient of the indicator variable for the blood–brain barrier disruption treatment is 0.797. So, expressed in accordance with the interpretation for log-transformed responses, the median ratio of antibody concentration in the brain tumor to antibody concentration in the liver is estimated to be exp(0.797) = 2.22 times greater for the blood–brain barrier diffusion treatment than for the control.

11.6 RELATED ISSUES

11.6.1 Weighted Regression for Certain Types of Nonconstant Variance

Although nonconstant variance can sometimes be corrected by a transformation of the response, in many situations it cannot. If enough information is known about the form of the nonconstant variance, the method of *weighted least squares* may be used.

The *weighted regression* model, written here with two explanatory variables, is

$$\mu\{Y_i \mid X_{1i}, X_{2i}\} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$$

$$\text{Var}\{Y_i \mid X_{1i}, X_{2i}\} = \sigma^2/w_i,$$

where the w_i's are known constants called *weights* (because cases with larger w_i's have smaller variances and should be weighted more in the analysis).

This model arises in at least three practical situations:

1. *Responses are estimates; SEs are available.* Sometimes the response values are measurements whose estimated standard deviations, SE(Y_i), are available. In the preceding model, the w_i's are taken to be $1/[\text{SE}(Y_i)]^2$; that is, the responses with smaller standard errors should receive more weight.
2. *Responses are averages; only the sample sizes are known.* If the responses are averages from samples of different sizes and if the ordinary regression model applies for the individual observations (the ones going into the average), then the weighted regression model applies to the averages, with weights equal to the sample sizes. The averages based on larger samples are given more weight.
3. *Variance is proportional to X.* Sometimes, while the regression of a response on an explanatory variable is a straight line, the variance increases with increases in the explanatory variable. Although a log transformation of the response might correct the nonconstant variance, it would induce a nonlinear relationship. A weighted regression model, with $w_i = 1/X_i$ (or possibly $w_i = 1/X_i^2$) may be preferable.

The weighted regression model can be estimated by *weighted least squares* within the standard regression procedure in most statistical computing programs. The estimated regression coefficients are chosen to minimize the weighted sum of squared residuals (see Exercise 21 for the calculus). It is necessary for the user to specify the response, the explanatory variables, and the weights.

11.6.2 The Delta Method

When, as in the alcohol metabolism study, there is a quantity of interest that is a nonlinear function of model parameters, calculating a standard error for the estimate of the quantity requires advanced methods. One such method—the *delta method*—requires some calculus and is therefore presented as an optional topic.

Taking the alcohol metabolism study's example, suppose interest centers on the parameter $\theta = \beta_1/(\beta_1 + \beta_2)$, where estimates of β_1 and β_2 are available. Substituting the estimates into the equation for θ produces an estimate for θ. Two inputs are required for calculating its standard error: (1) the variance–covariance matrix of the β-estimates, which should be available from the computer, and (2) the partial derivatives of θ with respect to each of the β's. Display 11.18 illustrates how these pieces combine to produce the standard error for this θ.

11.6.3 Measurement Errors in Explanatory Variables

Sometimes a theoretical model specifies that the mean response depends on certain explanatory variables that cannot be measured directly. This is called the *errors-in-variables problem.* If, for example, a study is examining the relationship between blood cholesterol (Y) and the dietary intake of polyunsaturated fat (X), and if the intake of polyunsaturated fat is estimated from a questionnaire individuals supply on what they eat in a typical week, then the questionnaire results will not measure X precisely.