BLOOD-BRAIN BARRIER EXPERIMENT

9. Diagnostics

In Section 8, we applied a linear regression model to our data by treating sacrifice time as a factor with four levels and treatment as a factor with two levels. We incorporated the two factors in the model by using three dummy variables to represent four levels of sacrifice time factor, and one dummy variable to represent two levels of the treatment factor. The multiple linear regression model had the form

$$LNRATIO = \beta_0 + \beta_1 * D3 + \beta_2 * D24 + \beta_3 * D72 + \beta_4 * TREAT + ERROR,$$

where *D3, D24, D72,* and *TREAT* are the dummy variables used to represent four levels of sacrifice time and two levels of treatment. We have found before that the covariates are not significant when the design variables, treatment and sacrifice time, are also included in the model.

The random variable *ERROR* is assumed to follow a normal distribution with the mean of zero and an unknown standard deviation σ. The standard deviation is constant at all levels of the response variable *LNRATIO.*

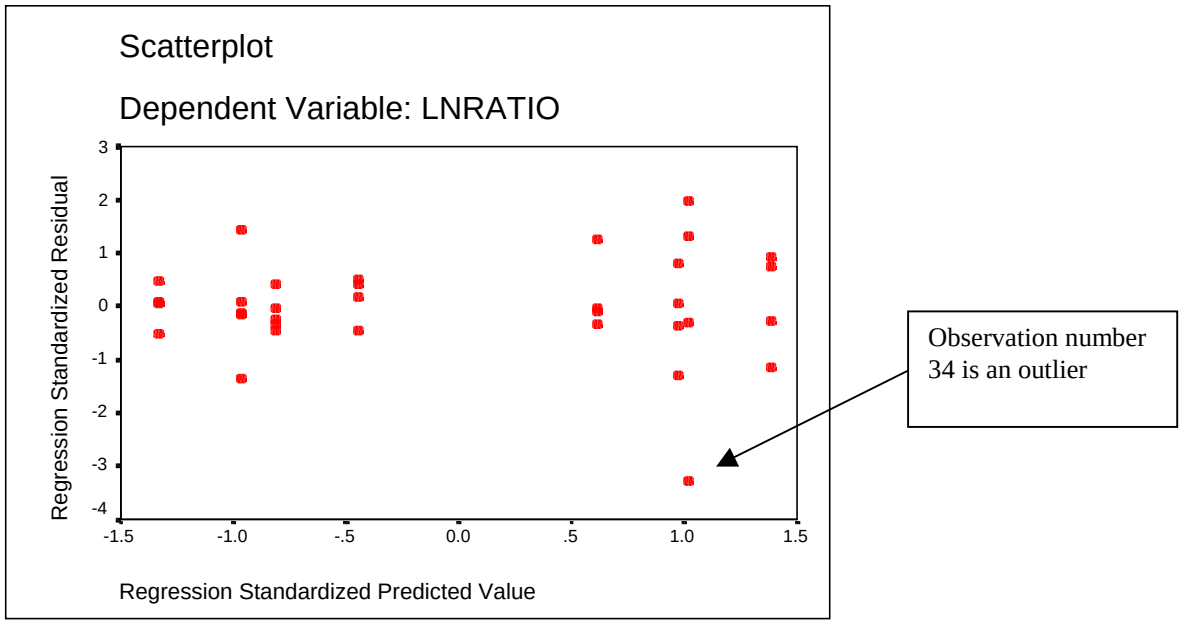
In this section various diagnostic tools will be used to evaluate the adequacy of the regression model.

In order to see whether the assumption of constant variance is not violated, we plot residuals (standardized residuals) against the fitted and also against each predictor variable. If the assumptions of linearity and constant variance appear to be met, then these residual plots should exhibit a random scatter of points with similar spread across all levels of fitted and independent variable values.

The plot displayed below shows the scatterplot of standardized residuals against the corresponding fitted values. No obvious difficulties are revealed in this display. With the exception of the smallest fitted values, the variability appears to be quite similar across all levels of fitted values. A random pattern is apparent in the plot, the linearity assumption is not violated.

The plot displayed below shows the scatterplot of standardized residuals against the corresponding fitted values. There are some difficulties revealed in this display. The variability of the residuals appears to be uneven across all levels of fitted values. More precisely, the spread of residuals appears to increase when the magnitude of the fitted values increases.

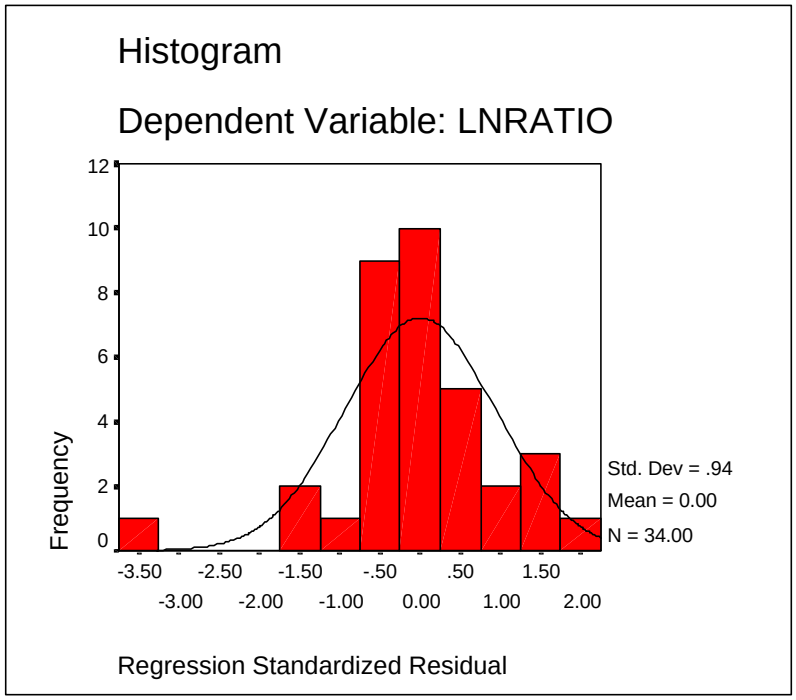
The effect is made even stronger by the presence of the observation 34 that lies far below the main body of the data. As the change in spread refers to a relatively small fraction of observations, it is difficult to make strong claims with respect to non-constant variability.



The plot shows residuals falling randomly, with no strong tendency to be either greater or less than zero.

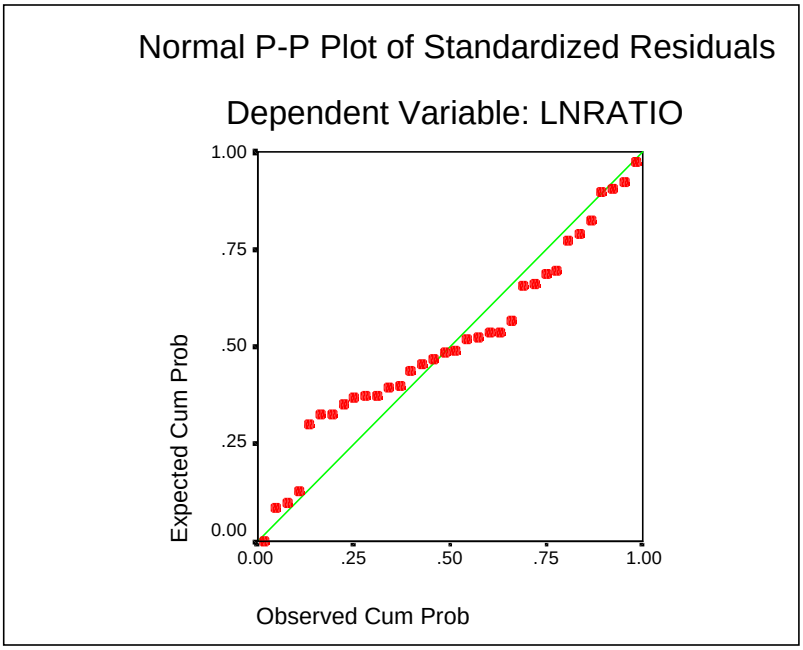
The case 34 is an outlier, and its potential to influence the position of the regression line will be discussed later.

The next two plots, a histogram and a normal p-p plot of the standardized residuals, are used to check whether the assumption of normality of residuals is plausible for the data.



As you can see, the standardized residuals follow approximately a normal distribution, although there is some slight skewness in the data.

SPSS also provides a normal plot of standardized residuals to verify the assumption of normality.



The plot displays some skewness in the data. However, most points in the plot are lying close to a straight line. Taking into account a relatively small number of observations, there is no sufficient evidence that the normality assumption is seriously violated.

Now we discuss the collinearity and multicollinearity.

It is well known that collinearity and multicollinearity can have harmful effects on multiple regression, both in the interpretation of the results and in how they are obtained. In particular, collinearity affects parameter estimates and their standard errors, and consequently t ratios. Inflated standard errors mean wider confidence intervals for the regression coefficients and a diminished ability of tests to find significant results.

Two measures for assessing both pairwise and multiple variable collinearity available in SPSS are the tolerance and the variance inflation factor (VIF). **Tolerance** is the amount of variability of the selected independent variable not explained by the other independent variables. It is obtained by making each independent variable a dependent variable and regressing it against the remaining independent variables. Tolerance values approaching zero indicate that the variable is highly collinear with the other predictor variables. The **variance inflation factor** (VIF) is inversely related to the tolerance value: $VIF = 1/TOLERANCE$. Large VIF values (a usual threshold is 10.0, which corresponds to a tolerance of .10) indicate a high degree of collinearity or multicollinearity among the independent variables.

SPSS provides the collinearity statistics for each variable when requested. The statistics for our data is displayed below:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	-4.302	.205		-21.010	.000	-4.720	-3.883		
	D3	1.134	.252	.226	4.501	.000	.619	1.650	.676	1.480
	D24	4.257	.259	.815	16.431	.000	3.727	4.787	.691	1.447
	D72	5.154	.259	.987	19.892	.000	4.624	5.684	.691	1.447
	TREAT	.797	.183	.180	4.346	.000	.422	1.172	.993	1.007

a. Dependent Variable: LNRATIO

No VIF value exceeds 10.0, and the tolerance values show that collinearity does not explain more than 10 percent of any independent variable's variance. There is no evidence of a significant collinearity in the problem.

SPSS regression collinearity diagnostics includes also the condition indices and the regression coefficient variance-decomposition matrix. A large condition index (over 30) indicates a high degree of collinearity. The regression coefficient variance-decomposition matrix shows the proportion of variance for each regression coefficient (and its associated variable) attributable to each condition index.

In order to examine collinearity, we first identify all condition indices above the threshold value of 30. Then for all condition indices exceeding the threshold, we identify variables with variance proportions above 0.90. A collinearity problem is indicated when a condition index identified as above the threshold value accounts for a substantial proportion of variance (.90 or above) for *two or more coefficients*. Thus each row in the matrix with the proportions exceeding 0.90 for at least two coefficients indicates significant correlations among the corresponding variables.

The collinearity diagnostics table for the brain-barrier data is displayed below:

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions				
				(Constant)	D3	D24	D72	TREAT
1	1	2.437	1.000	.03	.02	.02	.02	.06
	2	1.001	1.560	.00	.33	.09	.09	.00
	3	1.000	1.561	.00	.00	.26	.26	.00
	4	.439	2.357	.02	.09	.13	.13	.75
	5	.123	4.455	.95	.55	.49	.49	.19

a. Dependent Variable: LNRATIO

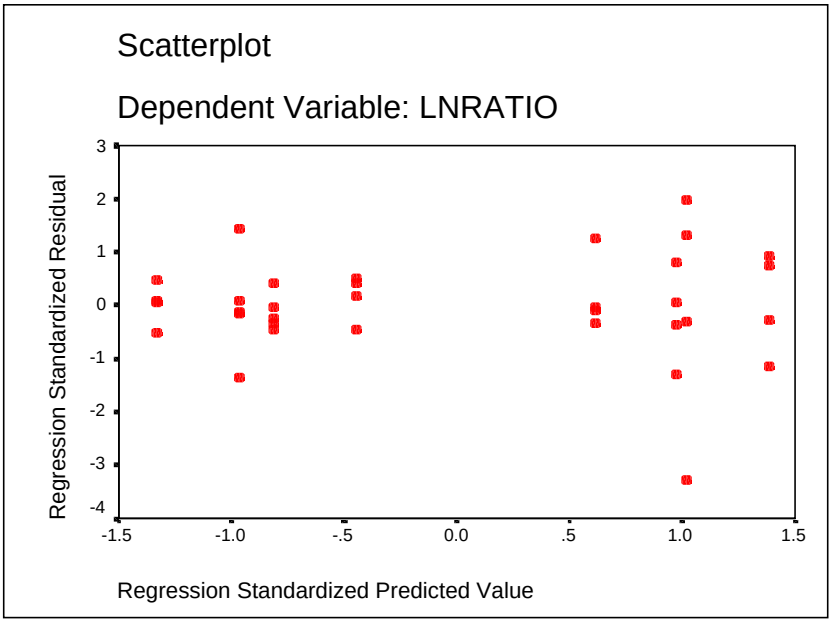
As you can see, none of the condition indices exceeds the threshold value of 30. Thus, we can find no support for the existence of multicollinearity.

Now we consider diagnostics for outliers and influential cases. An outlier is not necessarily an influential point, nor do all influential points have to be outliers. Thus, different statistical tools are used to identify outliers and influential observations. Studentized residuals are used for flagging outliers, and leverages and Cook's distances for flagging influential cases.

A studentized residual is a residual divided by its estimated standard deviation. The standardization makes the residuals directly comparable (larger predicted values have larger residuals). The studentized residual is the primary indicator of an observation that is an outlier on the dependent variable. With a fairly large sample size (50 or above), we may use a rule of thumb that studentized residuals smaller than -2 or larger than 2 are substantial. Observations falling outside the range can be considered *potential* outliers.

Instead of using cutoffs based on distributional assumptions, many researchers plot the standardized residuals, looking for points that stand apart from the others.

The following plot is a plot of standardized residuals versus case number for the brain-barrier data.



Only one observation, the case 34 stands apart from the others. SPSS provides the statistics for the case.

Casewise Diagnostics[a]

Case Number	Std. Residual	LNRATIO	Predicted Value	Residual
34	-3.266	-.89	.8522	-1.7402

a. Dependent Variable: LNRATIO

The standardized residual for the case is -3.266, and thus the observation can be treated as an outlier. Its potential to influence the position of the regression line will be discussed below.

The leverage of a case is a measure of the distance between its explanatory variable values and the average of the explanatory variable values in the entire data set. This observation has substantial impact on the regression results due to its differences from other observations. Leverages are greater than 1/n and less than 1, and the average of all leverages in a data set is always p/n, where p is the number of regression variables. While a large leverage does not necessarily indicate that the case is influential, it does imply that the case has a high potential for influence. Statisticians use (2*p)/n as a lower cutoff point for flagging potential influential cases (if p>10 and n>50), (3*p)/n otherwise. Instead of using cutoffs, many researchers are looking for points that stand apart from the others.

In the blood-brain barrier problem, the threshold leverage value is (3*4)/34=0.353. All leverage values lie below the threshold value.

In order to get some overall assessment of influence and see whether the case 34 is indeed influential, we will look at some other case influence statistics. One of these statistics is the Cook's distance.

Cook's Distance measures overall influence of a single case on the estimated regression coefficients when the case is deleted from the estimation process. Large values (usually greater than 1) indicate substantial influence by the case in affecting the estimated regression coefficients. However, even if no observations exceed this threshold, additional attention is dictated if a small set of observations has substantially higher values than the rest.

The values of Cook's distances are provided by an SPSS output when requested. Although the value of Cook's Distance for case 34 is equal to 0.46152, which is smaller than 1, it is obviously substantially larger than the rest. Therefore, it is worthy to rerun the regression without the case to see its influence on the regression results.

Running multiple regression without the case changes the coefficient of determination (it is equal to .972 without the case and .951 for all observations), the estimate of standard error (0.5328 without the case, and 0.4074 for all observations), and the coefficients of the regression equation. The regression equation without the case is

$$\mu\{LNRATIO\} = 1.121 * D3 + 4.250 * D24 + 5.404 * D72 + .675 * TREAT - 4.234.$$

Notice that the estimate of the multiplicative effect of the diffusion treatment has changed from exp(0.797) = 2.22 to exp(0.675) = 1.964. We can treat the case 34 as an influential observation.