Chapter 9: Regression Diagnostics

Regression diagnostics are used after fitting to check if a fitted mean function and assumptions are consistent with observed data. The basic statistics here are the residuals or possibly rescaled residuals.

A related issue is the importance of each case on estimation and other aspects of the analysis. In some datasets, the observed statistics may change in important ways if one case is deleted from the data. Such a case is called influential, and we shall learn to detect such cases.

1 The Residuals

The basic multiple linear regression model is given by

$$\boldsymbol{Y} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \quad \text{Var}(\boldsymbol{e}) = \sigma^2 I,$$

where $\boldsymbol{X}$ is a known matrix with n rows and p' columns, including a columns of 1s for the intercept if the intercept is included in the mean function. We will further assume that $\boldsymbol{X}$ has a full column rank, meaning that $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ exists.

Defining the hat matrix

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}',$$

The vector of residuals $\widehat{\boldsymbol{e}}$ is given by

$$\begin{aligned}\widehat{\boldsymbol{e}} &= \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} \\ &= \boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y} = (I - \boldsymbol{H})\boldsymbol{Y}\,.\end{aligned}$$

1.1 Difference between $\widehat{e}$ and e

The errors $\boldsymbol{e}$ are unobservable random variables, assumed to have zero mean and uncorrelated elements, each with common variance σ^2.

For the residuals $\widehat{\boldsymbol{e}}$, we have

$$E(\widehat{\boldsymbol{e}}) = \mathbf{0}, \quad \text{Var}(\widehat{\boldsymbol{e}}) = \sigma^2(I - \boldsymbol{H}).$$

Although the residuals have mean zero, but each residual may have a different variance, and the residuals are correlated.

If the intercept is included in the mean function, then $\sum \widehat{e}_i = 0$. In scalar form, the variance of the ith residual is

$$\text{Var}(\widehat{e}_i) = \widehat{\sigma}^2(1 - h_{ii}), \tag{1}$$

where h_{ii} is the ith diagonal element of $\boldsymbol{H}$.

1.2 The Hat Matrix

The hat matrix has the following properties:

(1) $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$ or equivalently $(I - \boldsymbol{H})\boldsymbol{X} = \mathbf{0}$.

(2) $\boldsymbol{h}^2 = \boldsymbol{H}$ or equivalently $\boldsymbol{H}(I - \boldsymbol{H}) = \mathbf{0}$.

The second property implies that

$$\begin{aligned} \text{Cov}(\widehat{\boldsymbol{e}}, \widehat{\boldsymbol{Y}}) &= \text{Cov}((I - \boldsymbol{H})\boldsymbol{Y}, \boldsymbol{H}\boldsymbol{Y}) \\ &= \sigma^2 (I - \boldsymbol{H})\boldsymbol{H} = \mathbf{0}. \end{aligned}$$

Another name for $\boldsymbol{H}$ is the orthogonal projection on the column space of $\boldsymbol{X}$. The elements of $\boldsymbol{H}$ are given by

$$h_{ij} = \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_j = h_{ji}.$$

Many helpful relationships can be found between the h_{ij}. For example,

$$\sum_{i=1}^{n} h_{ii} = p'$$

and, if the mean function includes an intercept,

$$\sum_{i=1}^{n} h_{ij} = \sum_{j=1}^{n} h_{ij} = 1.$$

As can be seen from (1), cases with large values of h_{ii} will have small values for Var$(\widehat{e}_i)$; as h_{ii} gets closer to 1, this variance will approach to 0. For such a case, no matter what value if y_i is observed for the ith case, we are nearly certain to get a residual near 0. In addition,

$$\widehat{y_i} = \sum_{j=1}^{n} h_{ij} y_j = h_{ii} y_i + \sum_{j \neq i}^{n} h_{ij} y_j .$$

It shows that as h_{ii} approaches 1, $\widehat{y_i}$ gets closer to y_i. For this reason, h_{ii} is called the leverage of the ith case.

1.3 Residuals and the Hat Matrix with Weights

When $\text{Var}(\boldsymbol{e}) = \sigma^2\boldsymbol{W}^{-1}$ with $\boldsymbol{W}$ a known diagonal matrix of positive weights, all results so far in this section require some modification. A useful version of the hat matrix is given by

$$\boldsymbol{H} = \boldsymbol{W}^{1/2}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}^{1/2}$$

and the leverages are the diagonal elements of this matrix. The fitted values are given as usual by $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}}$ is now the WLS estimator.

The definition of the residuals is a little trickier. The obvious choice $y_i - \widehat{\boldsymbol{\beta}}'\boldsymbol{x}_i$ has important deficiencies. First, the sum of squares of these residuals will not equal the residual sum of squares because the weights are ignored. Second, the variance of the ith residual will depend on the weight of case i. Both of these problems can be solved by defining

$$\widehat{e}_i = \sqrt{w_i}(y_i - \widehat{\boldsymbol{\beta}}'\boldsymbol{x}_i).$$

The sum of squares of these residuals is the residual sum of squares, and the variance of the residuals does not depend on the weight.

1.4 The Residuals When the Model is Correct

Suppose that U is equal to one of the terms in the mean function, or some linear combination of the terms. Residuals are generally used in scatterplots of the residuals $\widehat{e}$ against U. The key features of these residual plots when the correct model is fit are as follows.

1. The mean function $E(\widehat{\boldsymbol{e}}|U) = 0$. This means that the scatterplot of residuals on the horizontal axis versus any linear combination of the terms should have a constant mean function equal to 0.

2. Since Var$(\widehat{e}_i|U) = \sigma^2(1 - h_i i)$, the variance function is not quite constant. The variability will be smaller for high-leverage cases with h_{ii} close to 1.

3. The residuals are correlated, but this correlation is generally unimportant and not visible in residual plots.

When the model is correct, residual plots should look like null plots.

1.5 The Residuals When the Model Is Not Correct

If the fitted model is based on incorrect assumptions, there will be a plot of residuals versus some term or combination of terms that is not a null plot. Here are some generic features of the residual plot for a simple linear regression problem.

(1) No problems: Null plots.

(2) Nonconstant variance.

(3) Curvature: an incorrectly specified mean function.

(4) Both curvature and nonconstant variance.

In models with many terms, we cannot necessarily associate shapes in a residual plot with a particular problem with the assumptions. For example, Figure 1 shows a residual plot for the fit of the mean function $E(Y|X = x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ for a set of artificial data. The plot suggests nonconstant variance. But these data were actually generated using a mean function

$$E(Y|X = x) = \frac{|x_1|}{2 + (1.5 + x_2)^2}$$

with constant variance, with scatterplot matrix given in Figure 2. The real problem is that the mean function is wrong, even though from the residual plot, nonconstant variance appear to be the problem. A nonnull residual plot in multiple regression indicates that something is wrong but does not necessarily tell what is wrong.

1.6 Fuel Data

According to theory, if the mean function and other assumptions are correct, then all possible residual plots of residuals versus any function of the terms should resemble a null plot, so many plots of residuals should be examined. Usual choices include plots versus each of the terms and versus

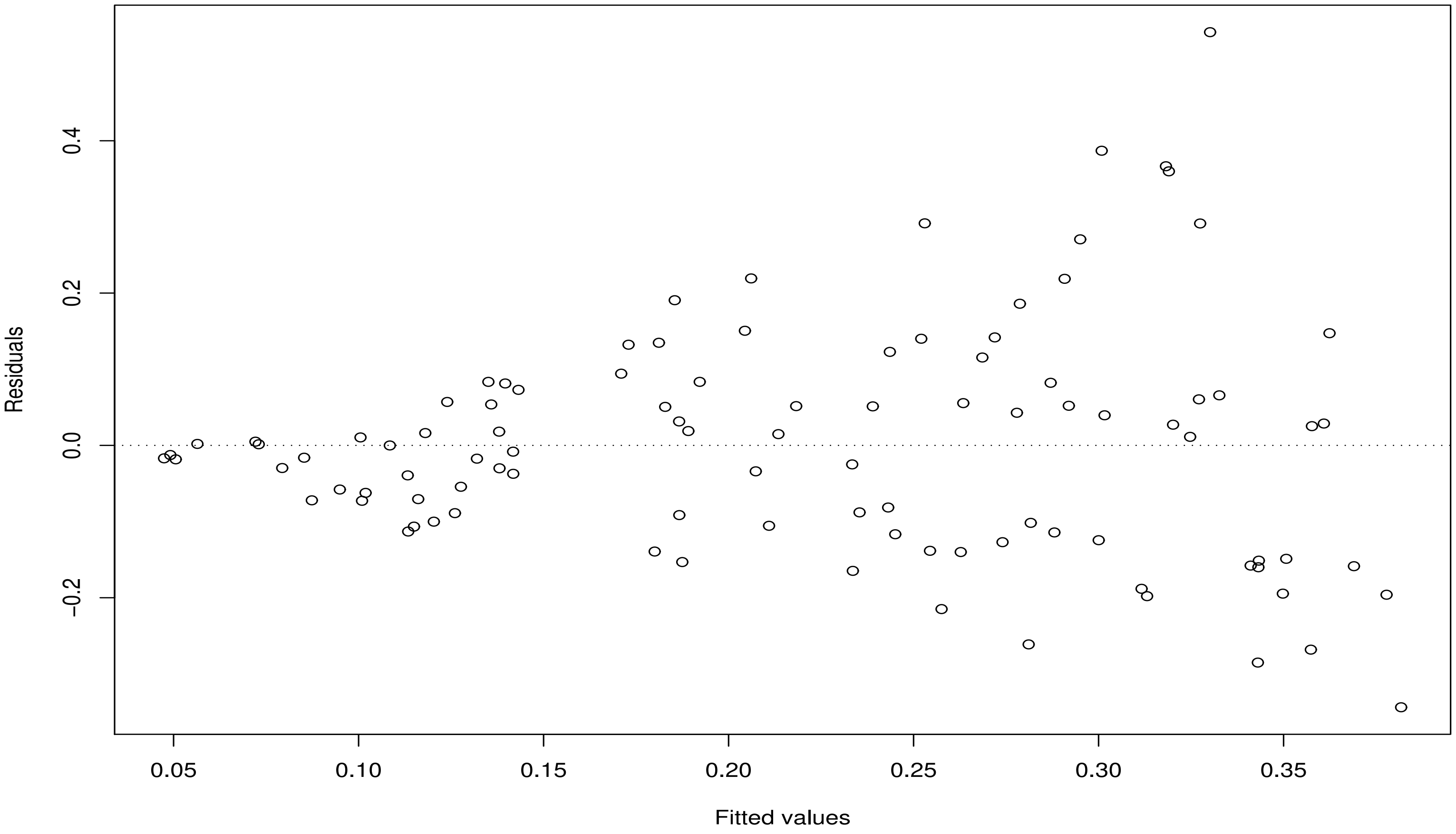


Figure 1: Residual plot for the caution data.

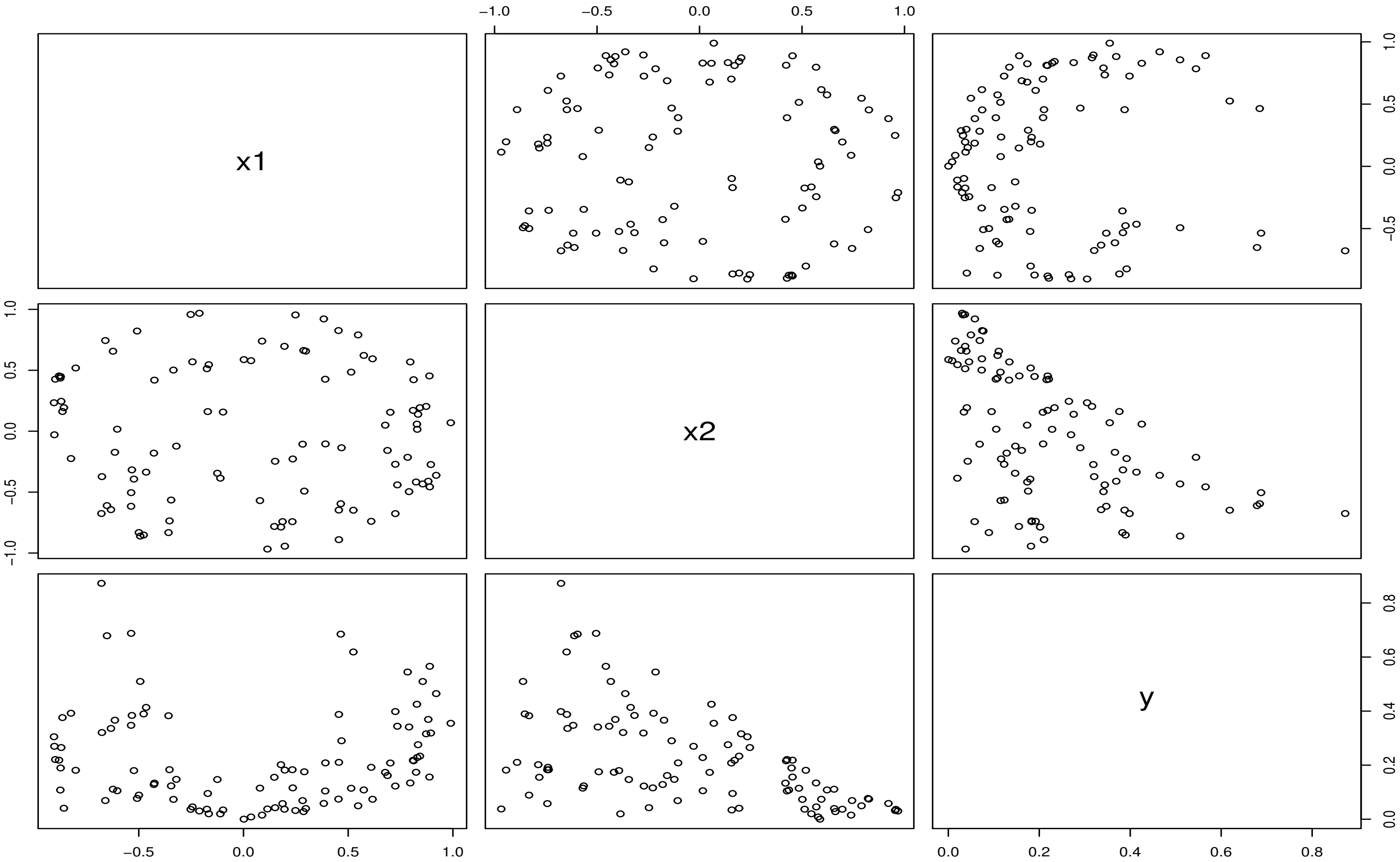


Figure 2: Scatterplot matrix for the caution data.

fitted values, as shown in Figure 8.5 for the fuel data. None of the plots versus individual terms in Figure 3(a)-(d) suggest any particular problems, apart from the relatively large positive residual for Wyoming and large negative residual for Alaska.

Figure 3(e) is a plot of residuals versus fitted values, which are just a linear combinations of the terms. There is a hint of curvature in this plot, possibly suggesting that the mean function is not adequate for the data.

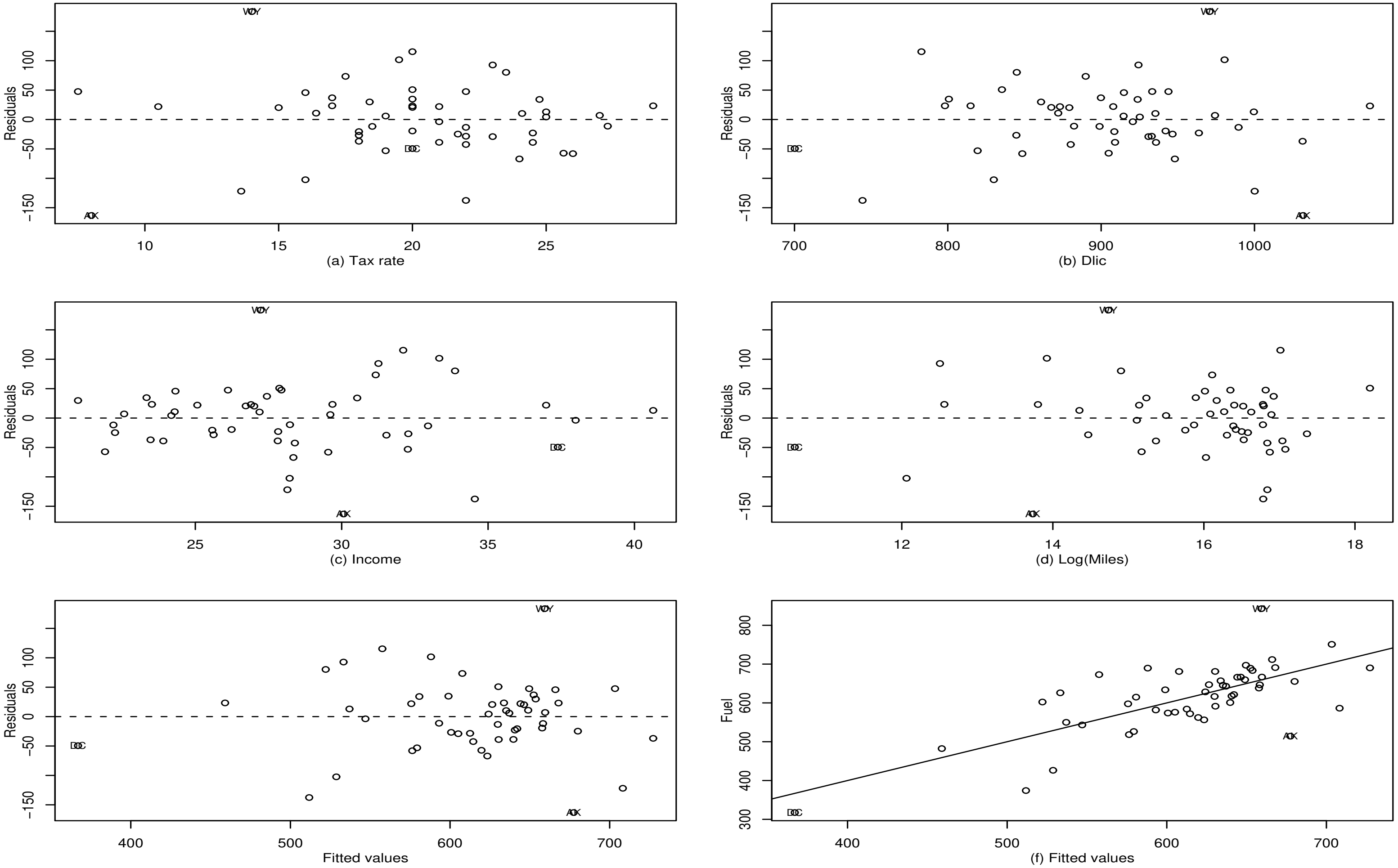


Figure 3: Residual plots for the fuel data.

2 Testing for Curvature

Tests can be computed to help decide if residual plots such as those in Figure 3 are null plots or not. Suppose we have a plot of residuals $\widehat{e}$ versus a quantity U on the horizontal axis, where U could be a term in the mean function or a combination of terms. A simple test for curvature is to refit the original mean function with an additional term for U^2 added. The test for curvature is then based on the t-statistic for testing the coefficient for U^2 to be 0. If U does not depend on estimated coefficients, then a usual t-test of this hypothesis can be used. If U is the fitted mean, Turkey's test for nonadditivity can be used.

Table 1 gives the lack-of-fit tests for the residual plots in Figure 3. None of the tests have small significance levels, providing no evidence against the mean function.

term	Test Stat.	Pr($>|t|$)
tax	-1.08	0.29
Dlic	-1.92	0.06
Income	-0.09	0.93
log(Miles)	-1.35	0.18
Fitted values	-1.45	0.15

Table 1: Significance levels for the lack-of-fit tests for the residual plot in Figure 3.

A second example, consider again the United Nations data from Section 3.1 with response log(Fertility) and two predictors log(PPgdp) and Purban. The scatterplot matrix shown in Figure 4 suggests that the mean function

$$E(\log(Fertility) | \log(PPgdp), Purban)$$
$$= \beta_0 + \beta_1 \log(PPgdp) + \beta_2 Purban,$$

should be appropriate for these data Plots of residuals are shown in Figure 5. The visual appearance

term	Test Stat.	Pr($> \|t\|$)
logPPgdp	3.22	0.0015
Purban	3.37	0.0009
Tukey test	3.65	0.0003

Table 2: Significance levels for the lack-of-fit tests for the residual plot in Figure 5.

of these plots is satisfactory. However, the curvature tests given in Table 2 tell a different story.

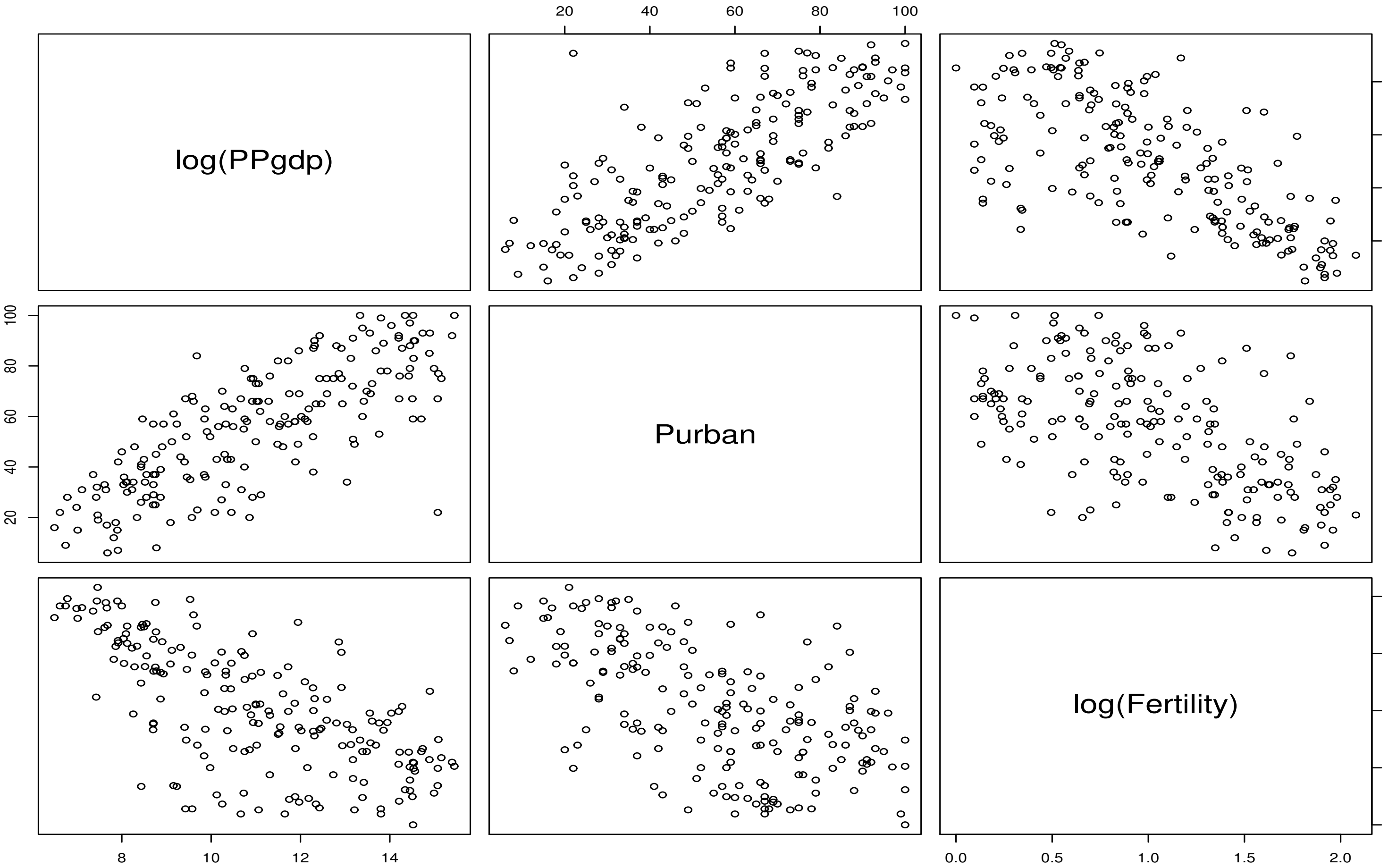


Figure 4: Scatterplot matrix for three variables in the UN data.

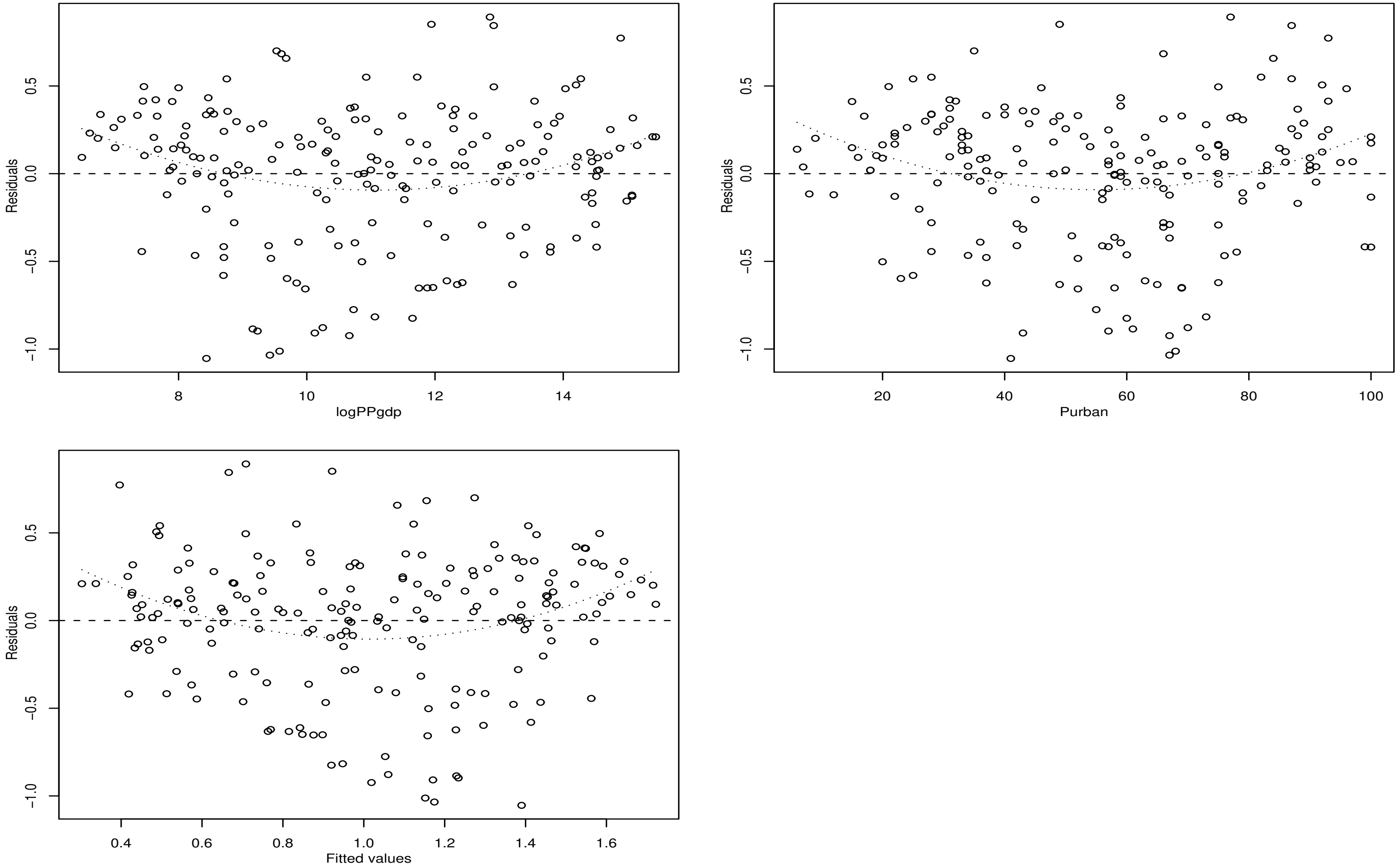


Figure 5: Residual plots for the UN data.

3 Nonconstant Variance

A nonconstant variance function in a residual plot may indicate that a constant variance assumption is false. There are at least four basic remedies for nonconstant variance.

(1) Using a variance stabilizing transformation.

(2) Finding empirical weights that could be used in weighted least squares. If replication is avail-

able, then within group variances may be used to provide approximate weights.

(3) Using the bootstrap method to get more accurate results. Estimates of parameters, given a misspecified variance function, remain unbiased, if somewhat inefficient. Tests and confidence intervals computed with the wrong variance function will be inaccurate, but the bootstrap is helpful in this scenario.

(4) Using generalized linear models, which can account for the nonconstant variance that is a function of the mean.

In this section, we consider primarily the first two options.

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 3 lists the common variance stabilizing transformations.

3.2 A Diagnostic for Nonconstant Variance

Suppose that Var$(Y|X)$ depends on an unknown vector parameter $\boldsymbol{\lambda}$ and a known set of terms Z with observed values for the ith case $\boldsymbol{z}_i$. For ex-

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.

Table 3: Common variance stabilizing transformations.

ample, if $Z = Y$, then variance depends on the response. Similarly, Z may be the same as X or a subset of X. We assume that

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

It says that variance depends on $\boldsymbol{z}$ and $\boldsymbol{\lambda}$ but only through the linear combination $\boldsymbol{\lambda}'\boldsymbol{z}$; and if $\boldsymbol{\lambda} = 0$, then $\text{Var}(Y|X, Z = \boldsymbol{z}) = \sigma^2$. The results of Chen (1983) suggest that the tests described here are not very sensitive to the exact functional form used above, and any form that depends on the linear combination $\boldsymbol{\lambda}'\boldsymbol{z}$ would lead to very similar inference.

Assuming that errors are normally distributed, a score test of $\boldsymbol{\lambda} = 0$ can be carried out using the following steps:

(1) Compute the OLS fit with the mean function

$$E(Y|X = x) = \boldsymbol{\beta}'\boldsymbol{x},$$

as if $\boldsymbol{\lambda} = 0$, i.e., constant variances. Save the residuals $\widehat{e}_i$.

(2) Compute scaled squared residuals $u_i = n\widehat{e}_i^2 / \sum_{i=1}^n \widehat{e}_j^2$. Combine the u_i into a variable U.

(3) Compute the regression with the mean function $E(U|Z = \boldsymbol{z}) = \lambda_0 + \boldsymbol{\lambda}'\boldsymbol{z}$. Obtain SS_{reg} for this regression with $df = q$, the number of components in Z. If the variance is thought to be a function of the responses, then SS_{reg}

will have 1 df.

(4) Compute the score test, $S = SS_{reg}/2$. The significance level for the test can be obtained by comparing S with its asymptotic distribution, which, under the hypothesis $\boldsymbol{\lambda} = \mathbf{0}$, is $\chi^2(q)$. If $\boldsymbol{\lambda} \neq 0$, then S will be too large, so large values of S provide evidence against the hypothesis of constant variance.

Snow Geese The relationship between $photo$=photo count, $obs1$=count by observer 1, and $obs2$=count by observer 2 of flocks of snow geese in the Hudson bay area of Canada is discussed in chapter 5. The data are displayed in Figure 6. We see in the graph that (1) there is substantial disagreement between the observers; (2) the observers cannot predict the photo count very well, and (3) the variability appears to be large for larger flocks.

Use the first observer only, we illustrate computation of the score test for constant variance. The first step is to fit the OLS regression of photo on obs1. The fitted mean function is $\widehat{E}(photo|obs1) = 26.55 + 0.88obs1$. From this, we can compute the residuals and u_is, and regress U on obs1. The score test for nonconstant variance is $S = SS_{reg}/2 = 81.41$, which, when compared with the chi-squared distribution with one df, gives an extremely small p-value. The nonconstant variance evident is almost certain in the data.

4 Graphs for Model Assessment

Residual plots are used to examine regression models to see if they fail to match observed data. If systematic failures are found, then models may need to reformulated to find a better fitting model. A closely related problem is assessing how well a

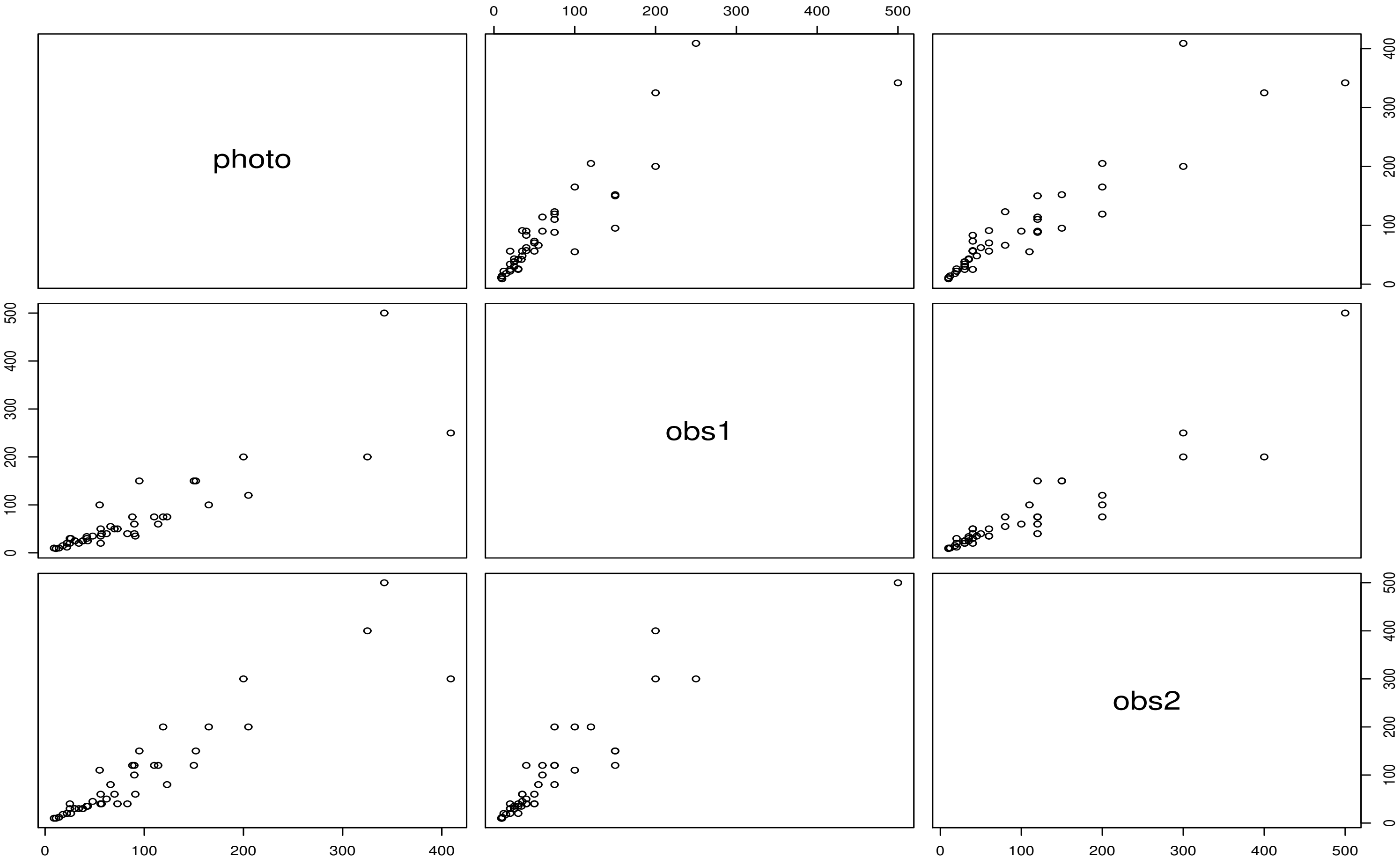


Figure 6: The snow geese data. The line on each plot is a loess smooth with smoothing parameter 2/3.

model matches the data. We now look at this issue from a graphical point of view using marginal model plots.

We illustrate the idea first with a problem with just one predictor. In Section 7.1, we discussed the regression of Height on Dbh for a sample of western red cedar trees. The mean function

$$E(Height|Dbh) = \beta_0 + \beta_1 Dgh$$

was shown to be a poor summary of these data, as can be seen in Figure 7. If we judge these two fits, the OLS fit and the loess fit, to be different, then we have visual evidence against the simple linear regression mean function. The loess fit is clearly curved, so the linear mean function is not a very good summary of this regression problem.

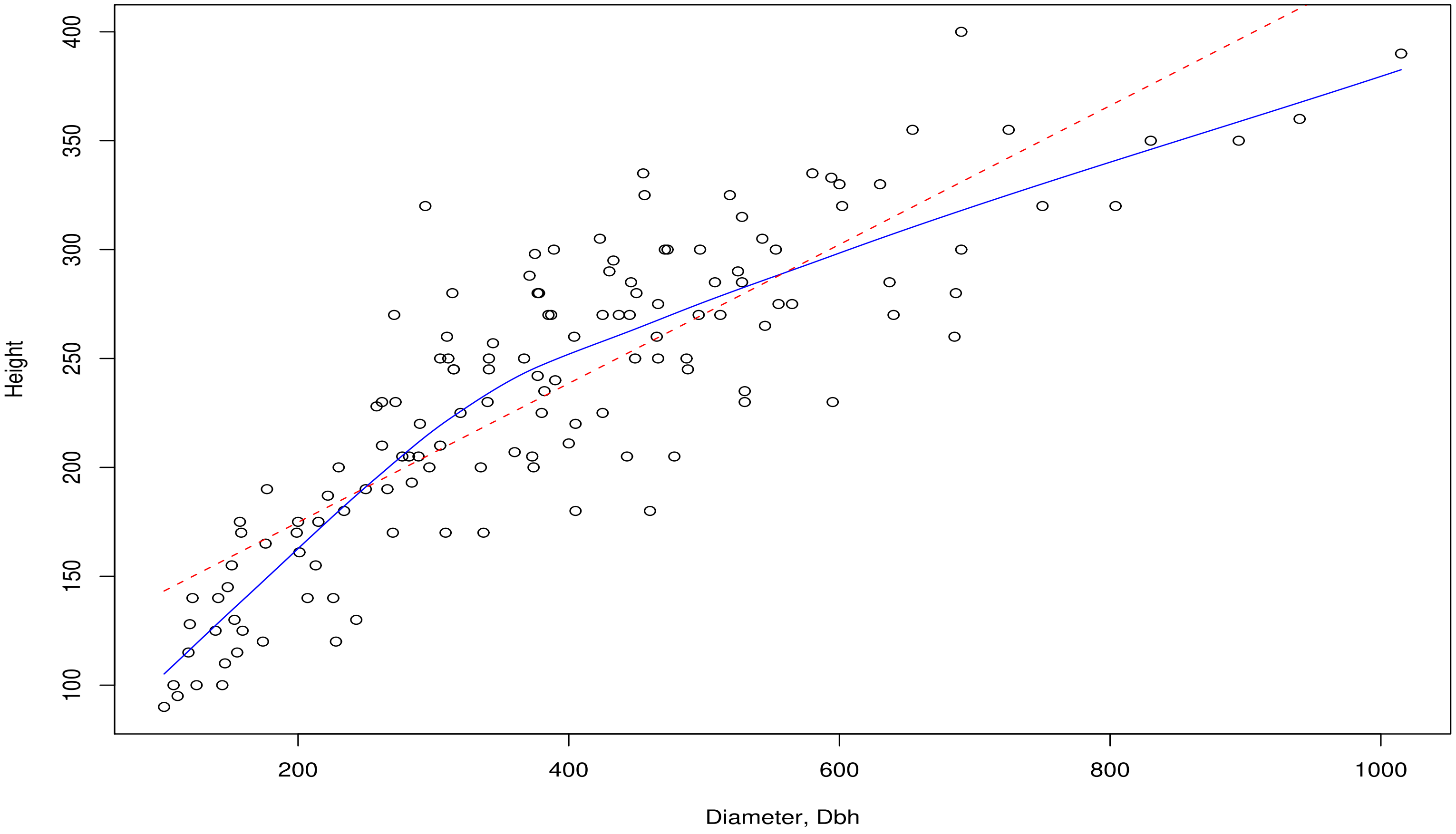


Figure 7: Model checking plot for the simple linear regression for western red cedar trees at upper flat creek.

4.1 Checking mean functions

With more than one predictor, we will look at marginal models to get a sequence of two-dimensional plots to examine. We will draw a plot with the response Y on the vertical axis. On the horizontal axis, we will plot a quantity U that will consist of any function of X we think is relevant, such as fitted values, any of the individual terms in X, or even transformations of them.

Under the model, we have

$$E(Y|U = \boldsymbol{u}) = E[E(Y|X = x)|U = \boldsymbol{u}].$$

This implies

$$E(Y|U = \boldsymbol{u}) \approx E[\widehat{Y}|U = \boldsymbol{u}].$$

Hence, we can estimate $E(Y|U = \boldsymbol{u})$ by smoothing the scatterplot with U on the horizontal axis, and the fitted values $\widehat{Y}$ on the vertical axis. If the model is correct, then the smooth of Y versus U and the smooth of $\widehat{Y}$ versus U should agree; if the model is not correct, these smooths may not agree.

As an example, we return to the United Nations data, starting with the mean function given below,

$$E(\log(Fertility)|\log(PPgdp), Purban)$$
$$= \beta_0 + \beta_1 \log(PPgdp) + \beta_2 Purban,$$

and suppose that $U = \log(PPgdp)$. Figure 8 suggests that the above mean function is not adequate.

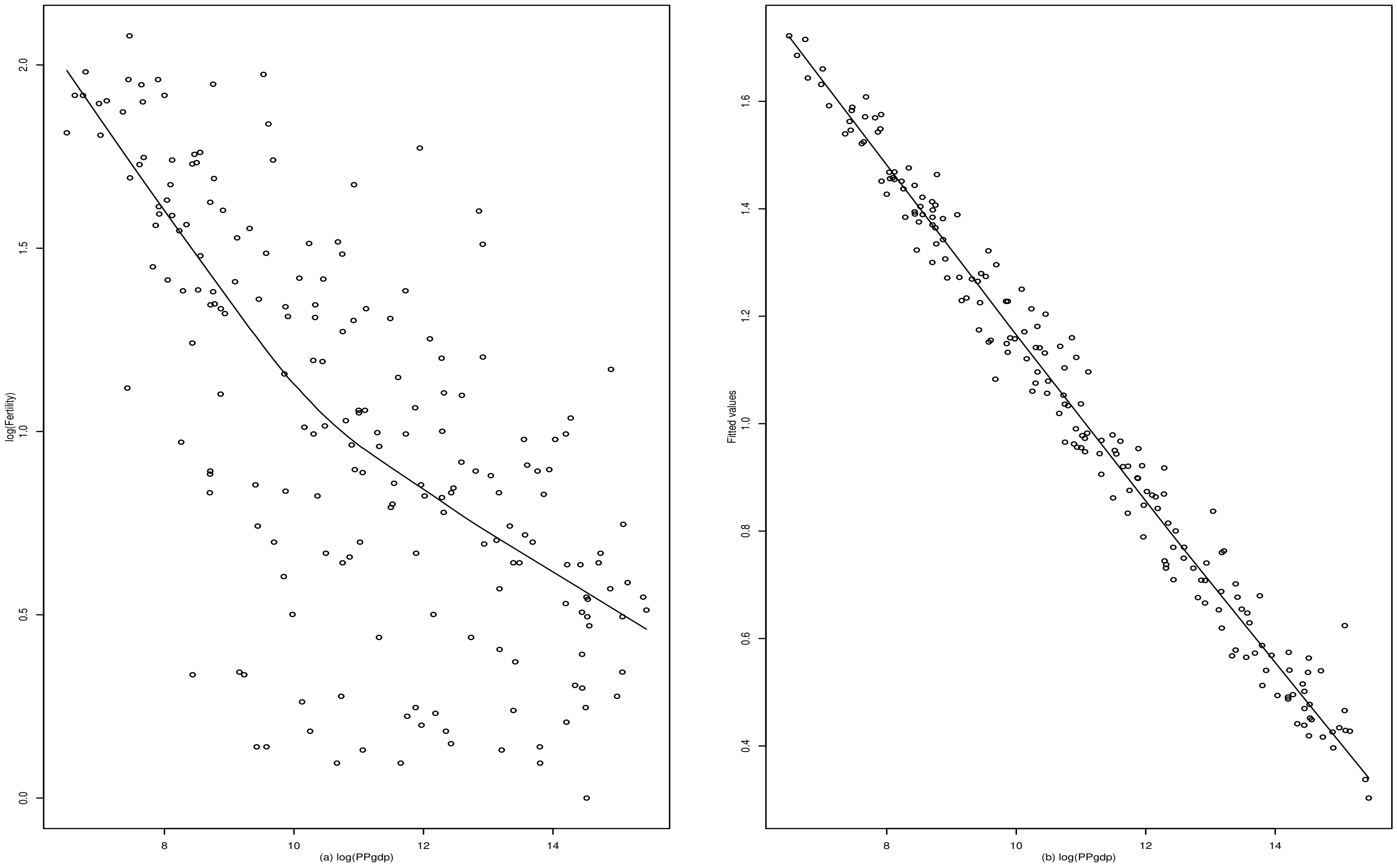


Figure 8: Plots for $\log(Fertility)$ versus $\log(PPgdp)$ and $\widehat{y}$ versus $\log(PPgdp)$. In both plots, the curves are loess smooths. If the model has the correct mean function, then these two smooths estimate the same quantity.

4.2 Checking variance functions

Model checking plots can also be used to check for model inadequacy in the variance function, which for the multiple linear regression problem means checking the constant variance assumption. The checking is based on the following calculation:

$$\begin{aligned}\text{Var}(Y|U) &= E[\text{Var}(Y|X)|U] + \text{Var}[E(Y|X)|U] \\ &\approx E(\sigma^2|U) + \text{Var}(\widehat{Y}|U) \\ &= \sigma^2 + \text{Var}(\widehat{Y}|U)\end{aligned} \qquad (2)$$

Equation (2) holds for the linear regression model in which the variance function $\text{Var}(Y|X) = \sigma^2$ is constant. According to this result, we can estimate $\text{Var}(Y|U)$ under the model by getting a variance smooth of $\widehat{Y}$ versus U, and then adding to this an estimate of $\widehat{\sigma}^2$ from the OLS fit of the model. We will call the square root of this estimated variance function $SD_{model}(Y|U)$. If the model is appropriate for the data, then apart from sampling error, $SD_{data}(Y|U) = SD_{model}(Y|U)$, but if the model is wrong, these two functions need not be equal.

For visual display, we show the mean function estimated from the plot $\pm SD_{data}(Y|U)$ using solid lines and the mean function estimated from the model $\pm SD_{model}(Y|U)$ using dashed lines.

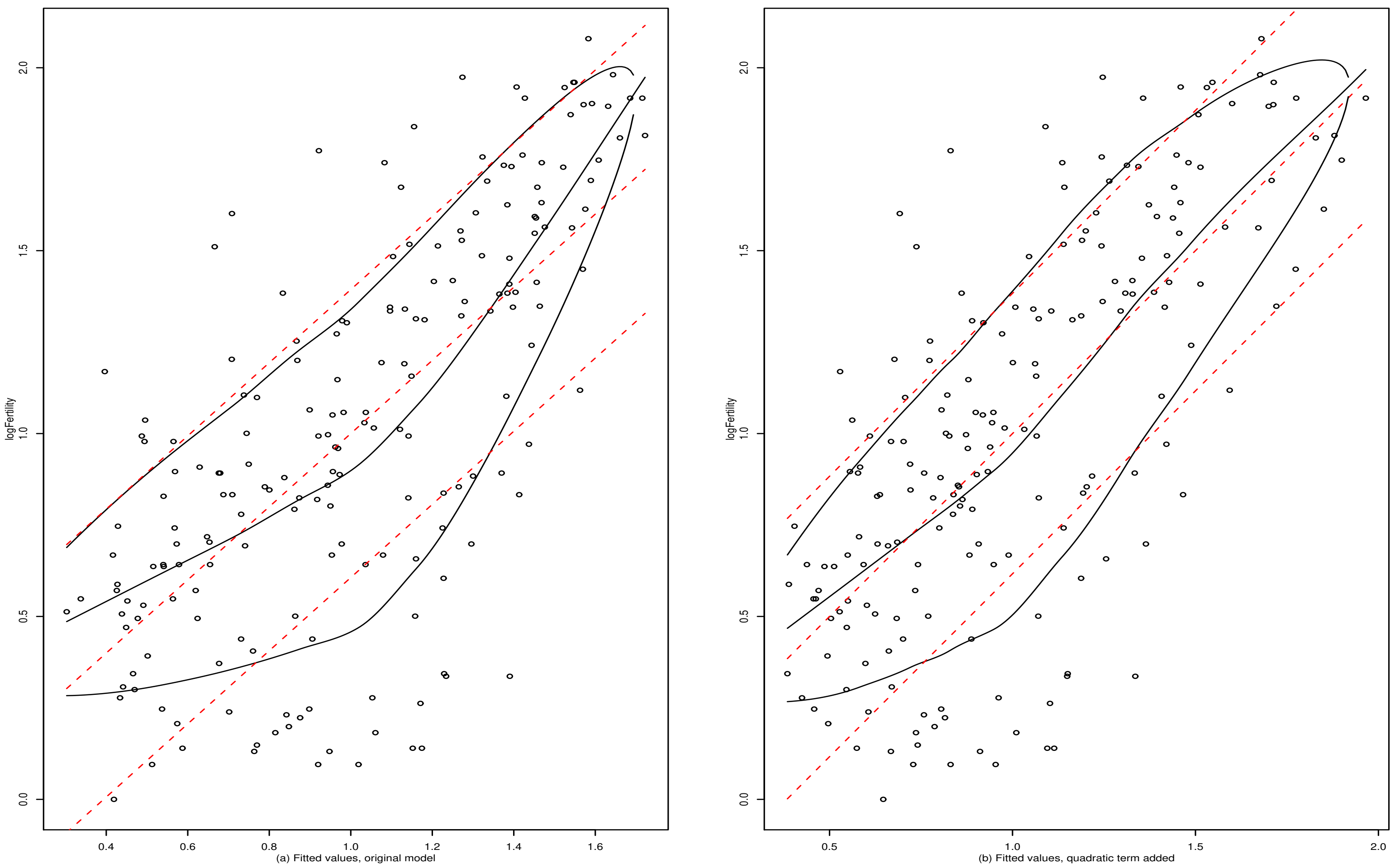


Figure 9: Marginal model plots with standard deviation smooths added.

5 Outliers

In some problems, the observed response for a few of the cases may not seem to correspond to the model fitted to the bulk of the data. Cases that do not follow the same model as the rest of the data are called outliers, and identifying these cases can be useful.

We use the mean shift outlier model to define outliers. Suppose that the ith case is a candidate for an outlier. We assume that the mean function for all other cases is $E(Y|X = \boldsymbol{x}_j) = \boldsymbol{x}_j'\boldsymbol{\beta}$, but for case i the mean function is $E(Y|X = \boldsymbol{x}_i) = \boldsymbol{x}_i'\boldsymbol{\beta} + \delta$. The expected response for the ith case is shifted by an amount δ, and a test of $\delta = 0$ is a test for a single outlier in the ith case. In this development, we assume Var$(Y|X) = \sigma^2$.

1. Cases with large residuals are candidates for outliers. Whatever testing procedure we de-

velop must offer protection against declaring too many cases to be outliers.

2. Outlier identification is done relative to a specified model. If the form of the model is modified, the status of individual cases as outliers may change.

3. Some outliers will have greater effect on the regression estimates than will others, a point that is pursued shortly.

5.1 An outlier test

Suppose that the ith case is suspected to be an outlier. Define a dummy variable

$$u_j = \begin{cases} 0 & j \neq i, \\ 1 & j = i. \end{cases}$$

Then, simply compute the regression of the response on both the terms in X and U. The estimated coefficient for U is the estimate of the mean shift δ. The t-statistic can then be used to test if

$\delta = 0$.

We will now consider an alternative approach that will lead to the same test, but from a different point of view. Again suppose that the ith case is suspected to be an outlier. We can proceed as follows:

1. Delete the ith case from the data, so $n - 1$ cases remain in the reduced data set.

2. Using the reduced data set, estimate $\boldsymbol{\beta}$ and

σ^2. Call these estimates $\widehat{\boldsymbol{\beta}}_{(i)}$ and $\widehat{\sigma}^2_{(i)}$ to remind us that case i was not used in estimation. The estimator $\widehat{\sigma}^2_{(i)}$ has $n - p' - 1$ df.

3. For the deleted case, compute the fitted value $\widehat{y}_{i(i)} = \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(i)}$. Since the ith case was not used in estimation, y_i and $\widehat{y}_{i(i)}$ are independent. Then

$$\text{Var}(y_i - \widehat{y}_{i(i)}) = \sigma^2 + \sigma^2 \boldsymbol{x}_i'(\boldsymbol{X}'_{(i)}\boldsymbol{X}_{(i)})^{-1}\boldsymbol{x}_i,$$

where $\boldsymbol{X}_{(i)}$ is the matrix $\boldsymbol{X}$ with the ith row deleted.

4. Assuming normal errors, a Student's t-test for the hypothesis $\delta = 0$ is given by
$$t_i = \frac{y_i - \widehat{y}_{i(i)}}{\widehat{\sigma}_{(i)}\sqrt{1 + \boldsymbol{x}_i'(\boldsymbol{X}_{(i)}'\boldsymbol{X}_{(i)})^{-1}\boldsymbol{x}_i}}. \qquad (3)$$
This test has $n - p' - 1$ df.

There is a simple computational formula for t_i in (3). We first define an intermediate quantity often called a standardized residual, by

$$r_i = \frac{\widehat{e}_i}{\widehat{\sigma}\sqrt{1-h_{ii}}},$$

where the h_{ii} is the leverage for the ith case. It is easy to verify that r_i has mean 0 and variance 1. With the aid of Appendix A.12, one can show that t_i can be computed as

$$t_i = r_i\left(\frac{n-p'-1}{n-p'-r_i^2}\right)^{1/2} = \frac{\widehat{e}_i}{\widehat{\sigma}_{(i)}\sqrt{1-h_{ii}}}.$$

The residual t_i is called a studentized residual.

This result shows that t_i can be computed from the residuals, the leverages and $\widehat{\sigma}^2$, so we don't need to delete the ith case, or to add a variable U, to get the outlier test.

5.2 Significance levels for the outlier test

Testing the case with the largest value of $|t_i|$ to be an outlier is like performing n significance tests, one for each of n cases. If, for example, $n = 65$, $p' = 4$, the probability that a t statistic with df 60 exceeds 2.0 in absolute is 0.05; however, the probability that the largest of 65 independent t-tests exceeds 2.0 is 0.964, suggesting quite clearly the need for a different critical value for a test based on the maximum of many tests. Since tests based on the t_i are correlated, this computation is only a guide.

The technique we use to find critical values is based on the Bonferroni inequality, which states that for n tests each of size a, the probability of falsely labeling at least one case as an outlier is no greater than na. Hence, choosing the critical value to be the $(\alpha/n) \times 100\%$ point of t will give a significance level of no more than $n(\alpha/n) = \alpha$. We would choose a level of $0.05/65 = 0.00077$ for each test to give an overall level of no more than $65(.00077) = 0.05$.

In Forbes' data, case 12 was suspected to be an outlier because of its large residual. The outlier test is $t_{12} = 12.4$. The nominal two-sided p-value corresponding to this test statistic when compared with the $t(14)$ distributed is 6.13×10^{-9}. The

Bonferroni-adjusted p-value is $17 \times 6.13 \times 10^{-9} = 1.04 \times 10^{-7}$. This very small value supports case 12 as an outlier.

The test locates an outlier, but it does not tell us what to do about it.

- If we believe the case is an outlier because of a blunder, e.g., an unusually large measurement error, or a recording error, then we might delete the outlier and reanalyze the data with-out the outlier.

- Try to figure out why a particular case is outlying. This may be the most important part of the analysis.

6 Influence of Cases

Single cases or small groups of cases cab strongly influence the fit of a regression model. Recall Anscombe's example, the fitted model depends entirely on the one point with $x = 19$.

The general idea of influence analysis is to study changes in a specific part of the analysis when the data are slightly perturbed. Whereas statistics such as residuals are used to find problems with a model, influence analysis is done as if the model were correct. The most useful and important method of perturbing the data is deleting the cases from the data one at a time. Cases whose removal causes major changes in the analysis are called influential.

Figure 10 is a scatterplot matrix of coefficient estimates for the three parameters in the UN data obtained by deleting cases one at a time. Every time a case is deleted, different coefficient estimates may be obtained.

6.1 Cook's Distance

The influence on the estimates of $\boldsymbol{\beta}$ (the information shown in Figure 10) can be summarized by comparing $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}_{(i)}$. This can be done using

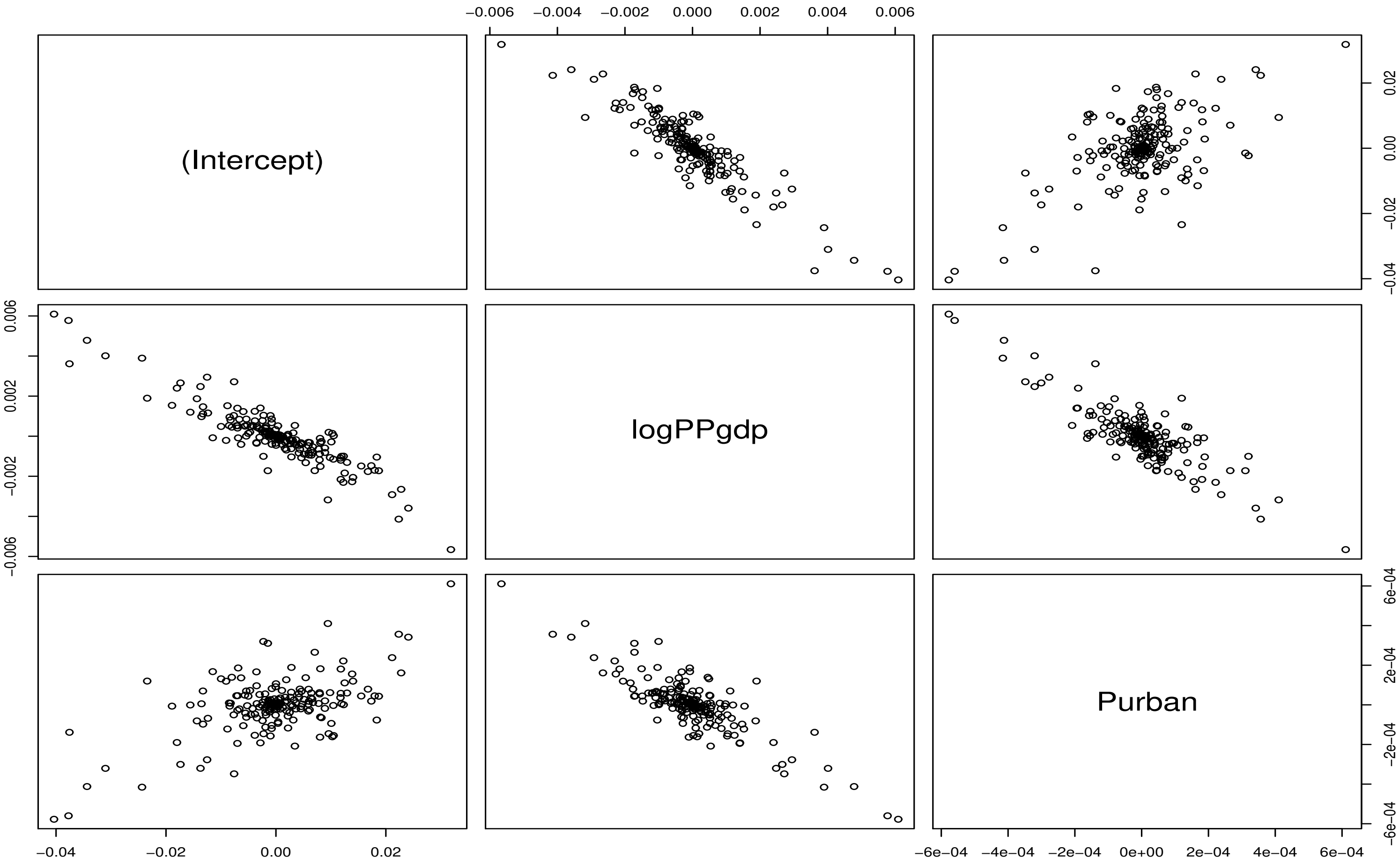


Figure 10: Estimates of parameters in the UN data obtained by deleting one case at a time.

Cook's distance (Cook, 1977) given by

$$D_i = \frac{(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})'(\boldsymbol{X}'\boldsymbol{X})(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})}{p'\widehat{\sigma}^2}$$

$$= \frac{(\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})'(\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})}{p'\widehat{\sigma}^2}.$$

Case for which D_i is large have substantial influence on both the estimate of $\boldsymbol{\beta}$ and on fitted values, and deletion of them may result in important changes in conclusions.

Typically, the case with the largest D_i, or in large data sets the cases with the largest few D_i, will be of interest. To investigate the influence of a case more closely, the analyst should delete the largest D_i case and recompute the analysis to see exactly what aspects of it have changed.

D_i can be computed in a simple form,

$$D_i = \frac{1}{p'} r_i^2 \frac{h_{ii}}{1 - h_{ii}}.$$

D_i is a product of the square of the ith standardized residual r_i and a monotonic function of h_{ii}. A large value of D_i may be due to large r_i, large h_{ii}, or both.

Rat Data An experiment was conducted to investigate the amount of a particular drug present in the liver of a rat. Nineteen rats were randomly selected, weighted, and given an oral dose of the drug. Because large livers would absorb more of a given dose than smaller livers, the actual dose an animal received was approximately determined as 40 mg of the drug per kilogram of body weight. Liver weight is known to be strongly related to body weight. After a fixed length of time, each rat was sacrificed, the liver weighted, and the percent of the dose in the liver determined. The experimental hypothesis was that, for the method of determining the dose, there is no relationship between the percentage of the dose in the live (Y) and the body weight, liver weight and relative dose.

The fitted regression summaries for the regression of Y on the three predictors are shown below.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.265922   0.194585   1.367   0.1919
BodyWt      -0.021246   0.007974  -2.664   0.0177 *
LiverWt      0.014298   0.017217   0.830   0.4193
Dose         4.178111   1.522625   2.744   0.0151 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.07729 on 15 degrees of freedom
Multiple R-Squared: 0.3639,     Adjusted R-squared: 0.2367
F-statistic:  2.86 on 3 and 15 DF,  p-value: 0.07197
```

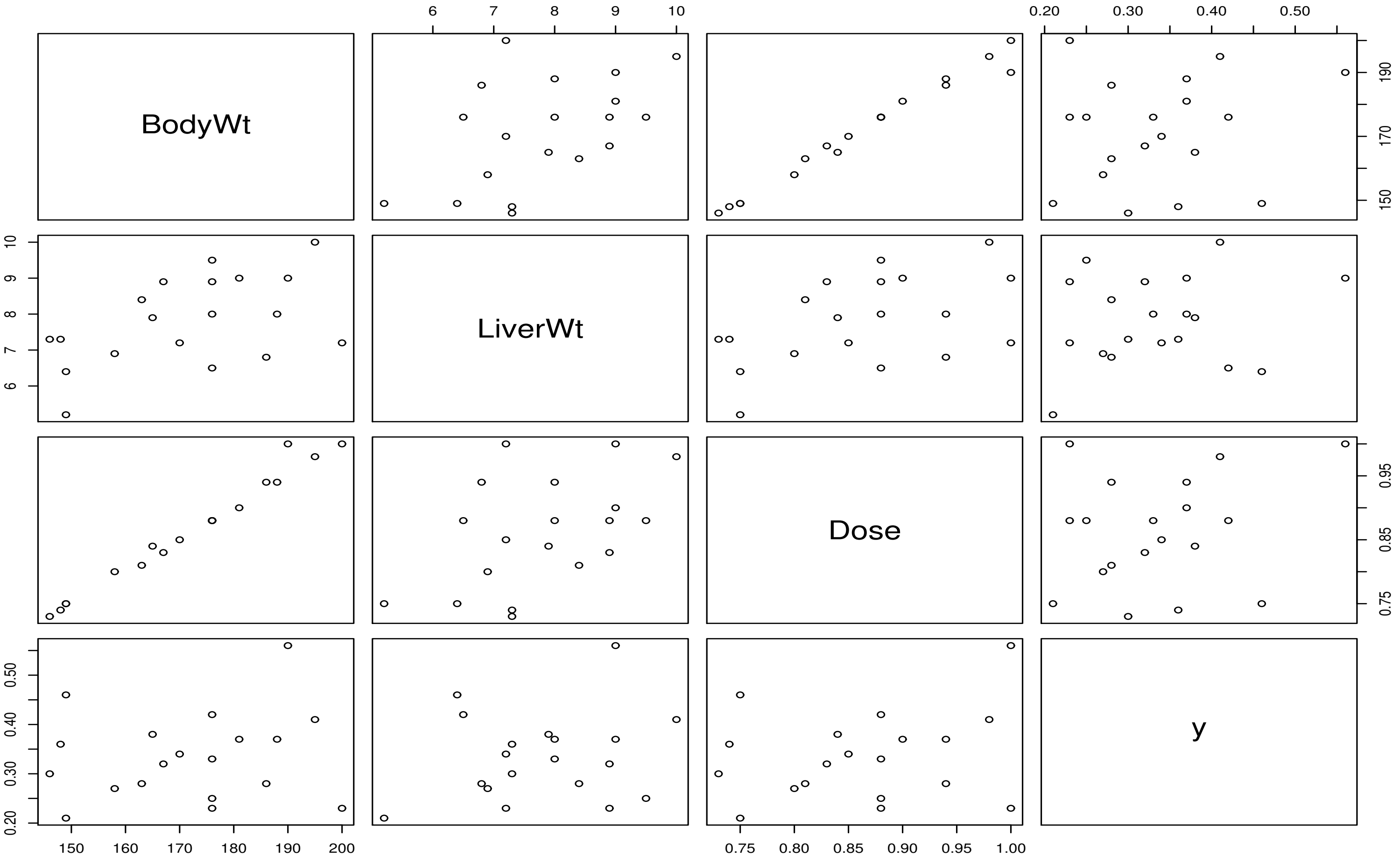


Figure 11: Scatterplot matrix for the rat data.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.1783569  0.2277755   0.783    0.445
BodyWt      0.0003535  0.0015138   0.234    0.818
LiverWt     0.0123260  0.0204127   0.604    0.554

Residual standard error: 0.09172 on 16 degrees of freedom
Multiple R-Squared: 0.0446,     Adjusted R-squared: -0.07483
F-statistic: 0.3735 on 2 and 16 DF,  p-value: 0.6942
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.117366   0.219094   0.536    0.600
Dose        0.173551   0.286261   0.606    0.553
LiverWt     0.008742   0.020085   0.435    0.669

Residual standard error: 0.09084 on 16 degrees of freedom
Multiple R-Squared: 0.06287,    Adjusted R-squared: -0.05427
F-statistic: 0.5367 on 2 and 16 DF,  p-value: 0.5948
```

The analysis might lead to the conclusion that while neither BodyWt nor Dose are associated with the response when the other is ignored, in combination they are associated with the response. But, from Figure 11, Dose and BodyWt are almost perfectly linearly related.

We turn to case analysis to attempt to resolve this paradox. The results are shown in Figure 12. The outlier statistics are not particularly large. However, Cook's distance immediately locates a possible cause: case three has $D_3 = 0.93$; no other case has D_i bigger than 0.27, suggesting that case 3 alone may have large enough influence on the fit to induce the anomaly. Rat number 3, with weight 190 g, was reported to have received a full dose of 1.0, which was a larger dose than it should have received according to the rule for assigning doses; for example, rat 8 with weight of 195g got a lower dose of 0.98. Deleting this case from the study, the following regression results are produced.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.311427   0.205094   1.518    0.151
BodyWt      -0.007783   0.018717  -0.416    0.684
LiverWt      0.008989   0.018659   0.482    0.637
Dose         1.484877   3.713064   0.400    0.695

Residual standard error: 0.07825 on 14 degrees of freedom
Multiple R-Squared: 0.02106,    Adjusted R-squared: -0.1887
F-statistic: 0.1004 on 3 and 14 DF,  p-value: 0.9585
```

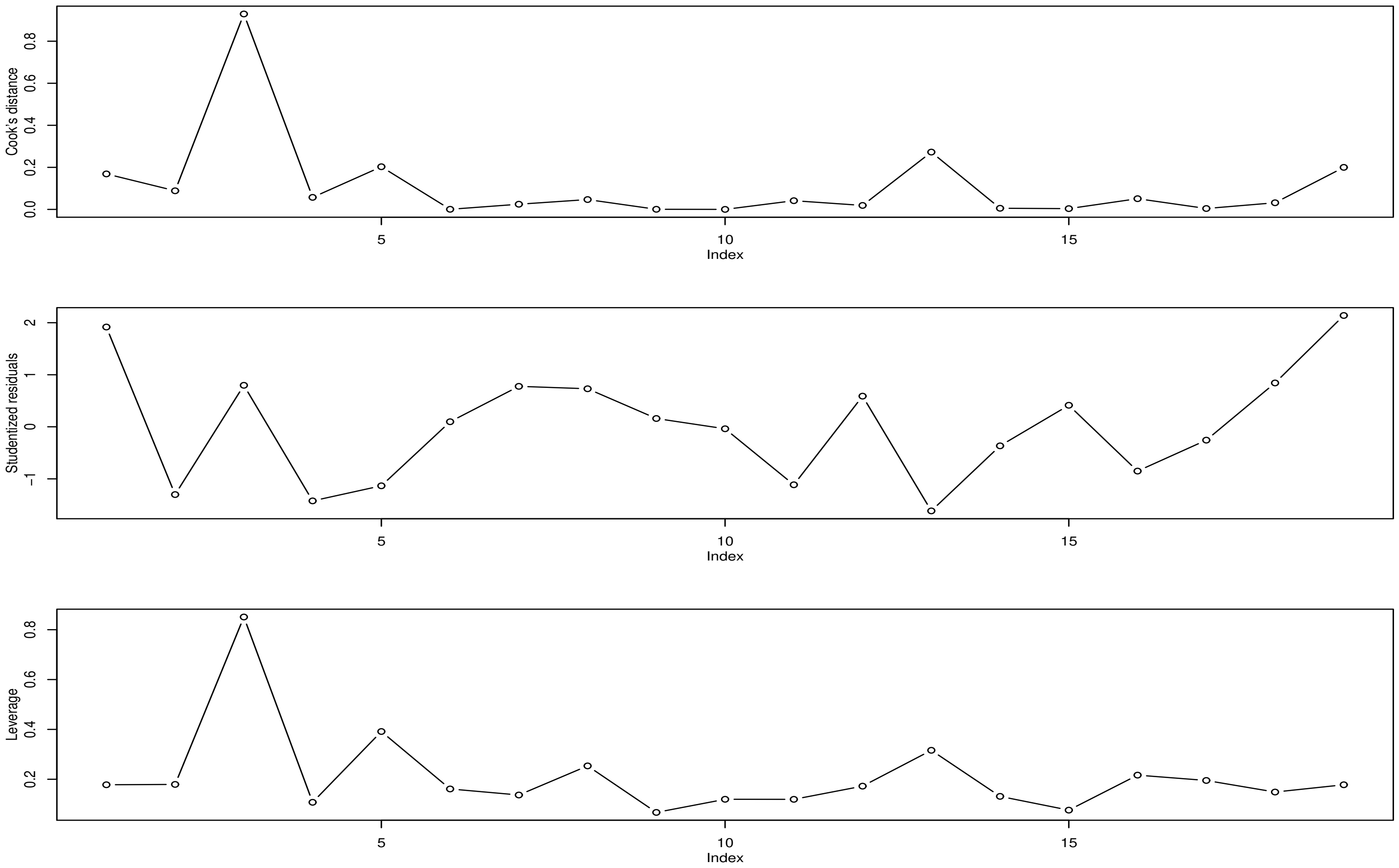


Figure 12: Diagnostic statistics for the rat data.

7 Normality Assumption

The assumption of normal errors plays only a minor role in regression analysis. It is needed primarily for inference with small samples, and even then the bootstrap can be used for inference. Furthermore, nonnormality of the unobservable errors is very difficult to diagnose in small samples by examination of residuals. The relationship between the errors and the residuals is

$$\begin{aligned}\widehat{\boldsymbol{e}} &= (I - H)\boldsymbol{Y} \\ &= (I - H)(\boldsymbol{x\beta} + \boldsymbol{e}) \\ &= (I - H)\boldsymbol{e}.\end{aligned}$$

In scalar form, the ith residual is

$$\widehat{e}_i = e_i - \sum_{j=1}^{n} h_{ij} e_j. \tag{4}$$

By the central limit theorem, the sum in (4) will be nearly normal even if the errors are not normal.

With a small or moderate sample size n, the second term can dominate the first, and the residuals can behave like a normal sample even if the errors are not normal.

As n increases for fixed p', the second term in (4) has small variance compared to the first term, so it becomes less important, and residuals can be used to assess normality. Should a test of normality be desirable, a normal probability plot can be used.

Suppose we have a sample of n numbers z_1, z_2, . . ., z_n, and we wish to examine the hypothesis that the z's are a sample from a normal distribution with known mean μ and variance σ^2. A useful way to proceed is as follows:

1. Order the z's to get $z_{(1)} \leq \cdots \leq z_{(n)}$.

2. Consider a standard normal sample of size n. Let $\mu_{(1)} \leq \cdots \leq \mu_{(n)}$ be the mean values

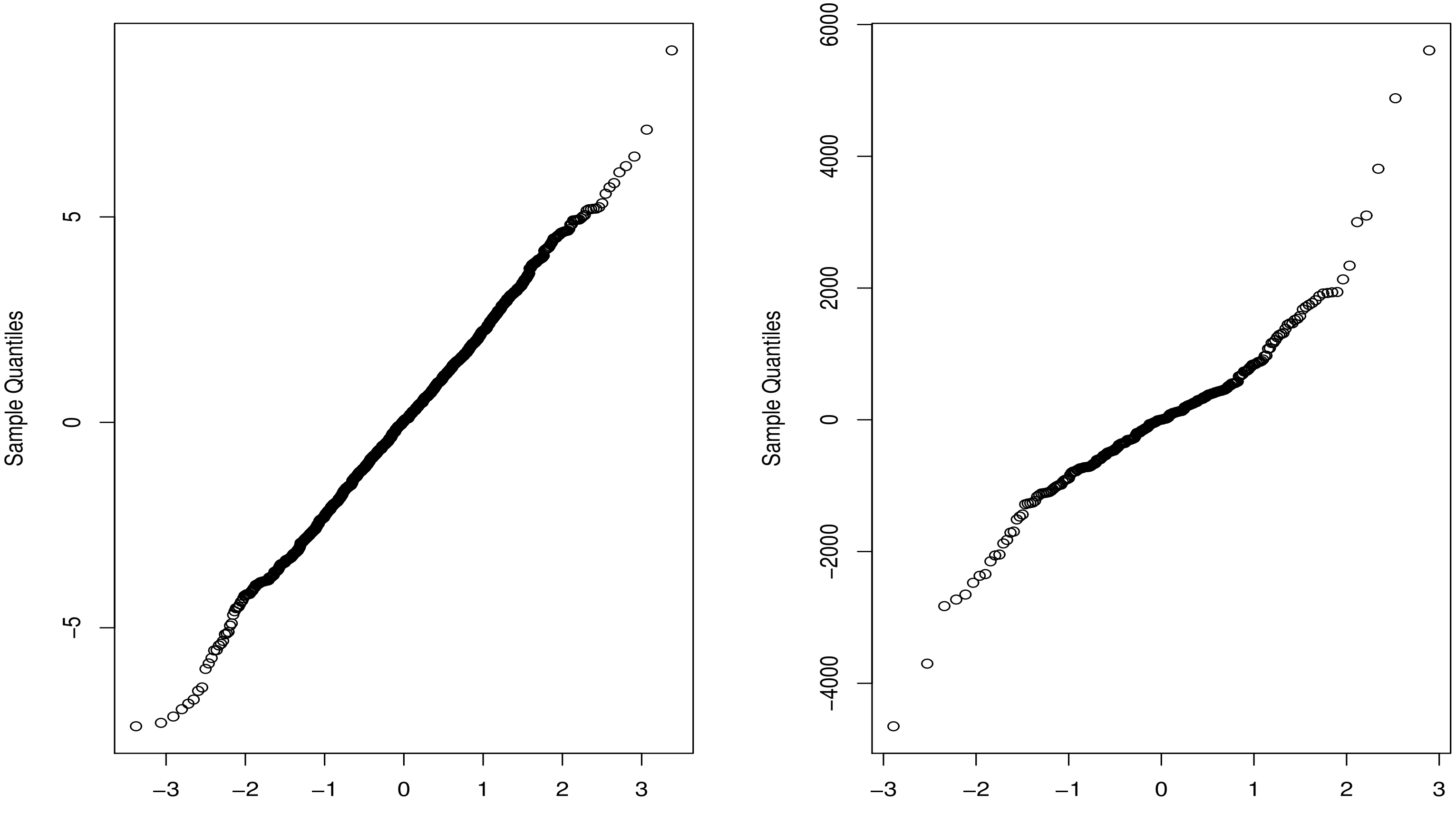
of the order statistics. The $\mu_{(i)}$ are available in printed tables or can be well approximated using a computer program.

3. If zs are normal, then

$$E(z_{(i)}) = \mu + \sigma\mu_{(i)},$$

   so that the regression of $z_{(i)}$ on $\mu_{(i)}$ will be a straight line. If it is straight, we have evidence against normality.

Figure 13 shows normal probability plots of the residuals for the height data and for the transactions data. Both have large enough samples for normal probability plots to be useful. For heights data, the plot indicates no evidence against normality. For the transactions data, normality is doubt because the plot is not straight. This supports the earlier claim that the errors for this problem are likely to be skewed with too many large values.



(a) Heights data

(a) Transaction data

Figure 13: Normal probability plots of residuals for (a) the heights data and (b) the transaction data.