Block 12

Regression Diagnostics

Regression Diagnostics

Techniques used *before fitting* to check:
(consistent with model assumptions?)
- mean function
- model assumptions

Related issue:
(will develop tools to detect these)
- influential cases
- outliers

The Residuals

The basic multiple linear regression model is given by:

$$E(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta} \quad \text{Var}(\mathbf{Y}|\mathbf{X}) = \sigma^2\mathbf{I}$$

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

Defining the Hat Matrix

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

The vector of residuals $\hat{\mathbf{e}}$ is given by:

$$\hat{\mathbf{e}} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{Y} - \mathbf{H}\mathbf{Y} = (\mathbf{I} - \mathbf{H})\mathbf{Y}.$$

Difference between $\hat{\mathbf{e}}$ and $\mathbf{e}$

The errors $\mathbf{e}$ are unobservable random variables, assumed to have zero mean and uncorrelated elements, each with common variance σ^2.
For the residuals $\hat{\mathbf{e}}$, we have:

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

$E(\hat{\mathbf{e}}) \equiv E(\hat{\mathbf{e}} \mid \mathbf{X}), \quad \text{Var}(\hat{\mathbf{e}}) \equiv \text{Var}(\hat{\mathbf{e}} \mid \mathbf{X})$
Although the residuals have mean zero, each residual may have a different variance, and the residuals are correlated.
If the intercept is included in the mean function, then $\sum \hat{e}_i = 0$. In scalar form, the variance of the ith residual is:

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

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

The Hat Matrix

The hat matrix has the following properties:
(1)$\mathbf{HX} = \mathbf{X}$ or equivalently $(\mathbf{I} - \mathbf{H})\mathbf{X} = 0$.
(2)$\mathbf{H}^2 = \mathbf{H}$ or equivalently $\mathbf{H}(\mathbf{I} - \mathbf{H}) = 0$.
The second property implies that:

$$\text{Cov}(\hat{\mathbf{e}}, \hat{\mathbf{Y}}) = \text{Cov}((\mathbf{I} - \mathbf{H})\mathbf{Y}, \mathbf{HY}) = \sigma^2(\mathbf{I} - \mathbf{H})\mathbf{H} = 0.$$

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

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

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,

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

From:

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

$$\implies \text{cases with large values of } h_{ii} \text{ will have small } \text{Var}(\hat{e}_i)$$

$$\implies \text{As } h_{ii} \to 1, \text{Var}(\hat{e}_i) \to 0$$

$$\implies \text{will get a residual near 0.}$$

Also,

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

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

Example: UN data

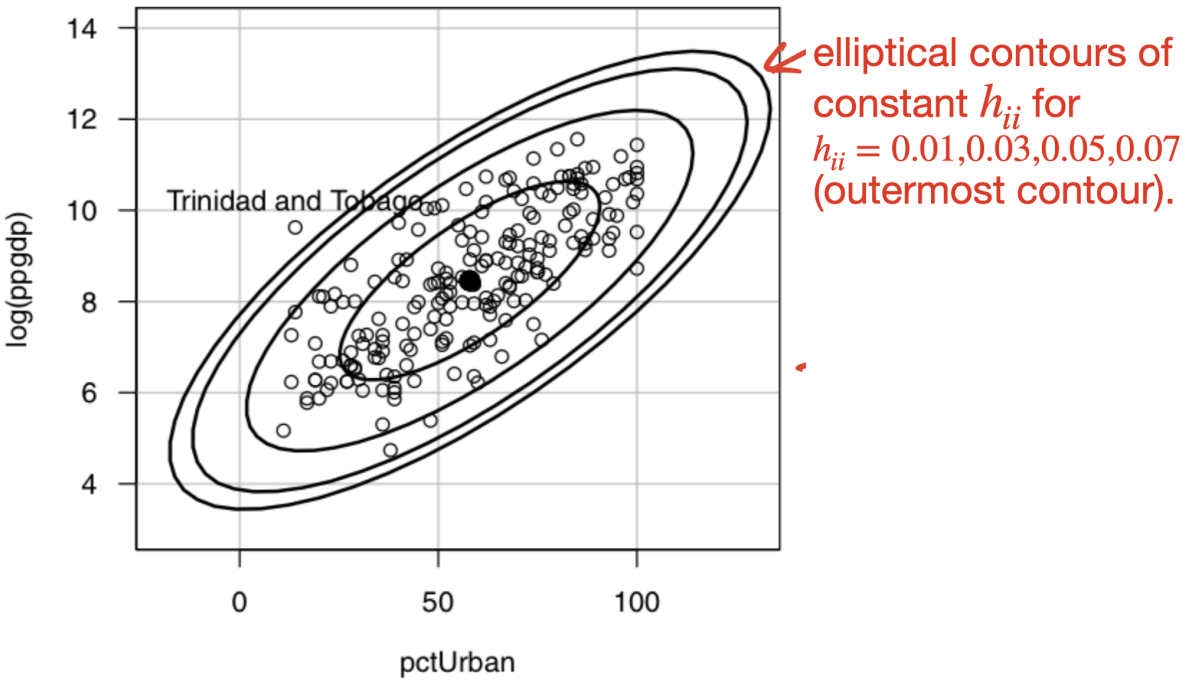


Figure 9.1 Contours of constant leverage in two dimensions.

Some takeaways:

1. Localities with highest urbanization (e.g., 100%) are between $h_{ii} = 0.02$ and 0.04 (not large).

2. Trinidad and Tobago, which has high income for relatively low urbanization, has the highest h_{ii} value.

Residuals and the Hat Matrix with Weights

When $\text{Var}(\mathbf{e}) = \sigma^2\mathbf{W}^{-1}$ with $\mathbf{W}$ 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:

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

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

The definition of the residuals is a little trickier. The obvious choice $y_i - \hat{\boldsymbol{\beta}}'\mathbf{x}_i$ 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:

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

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.

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 $\hat{\mathbf{e}}$ against U. The key features of these residual plots when the correct model is fit are as follows:

1. The mean function $E(\hat{e}|U) = 0$. 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 $\mathrm{Var}(\hat{e}_i|U) = \sigma^2(1 - h_{ii})$, 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.

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.

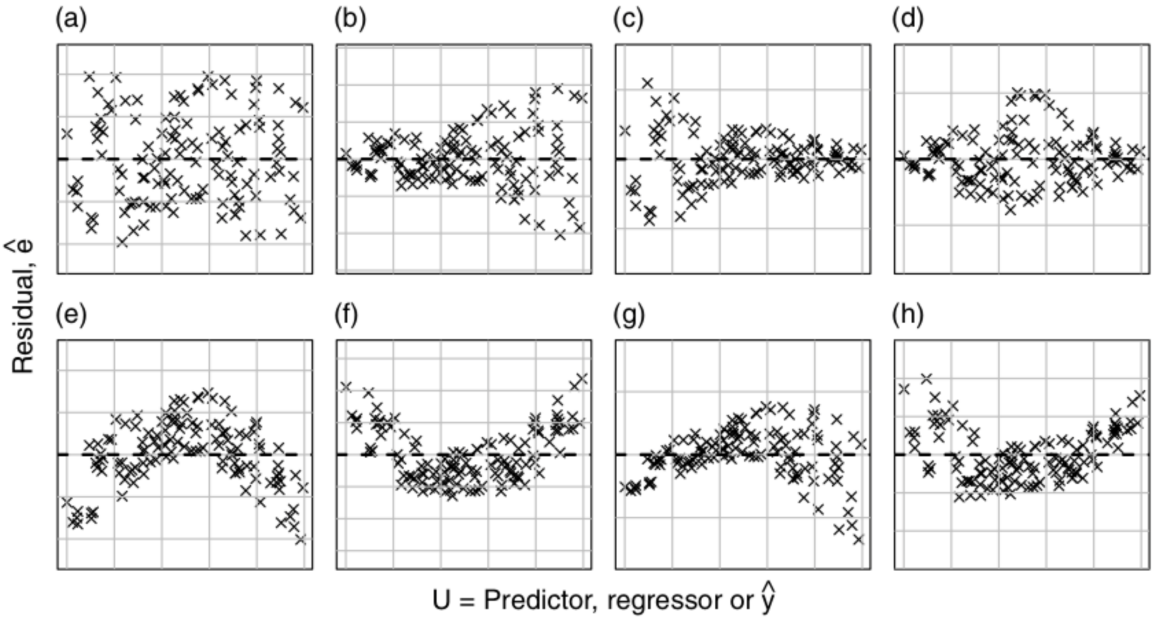
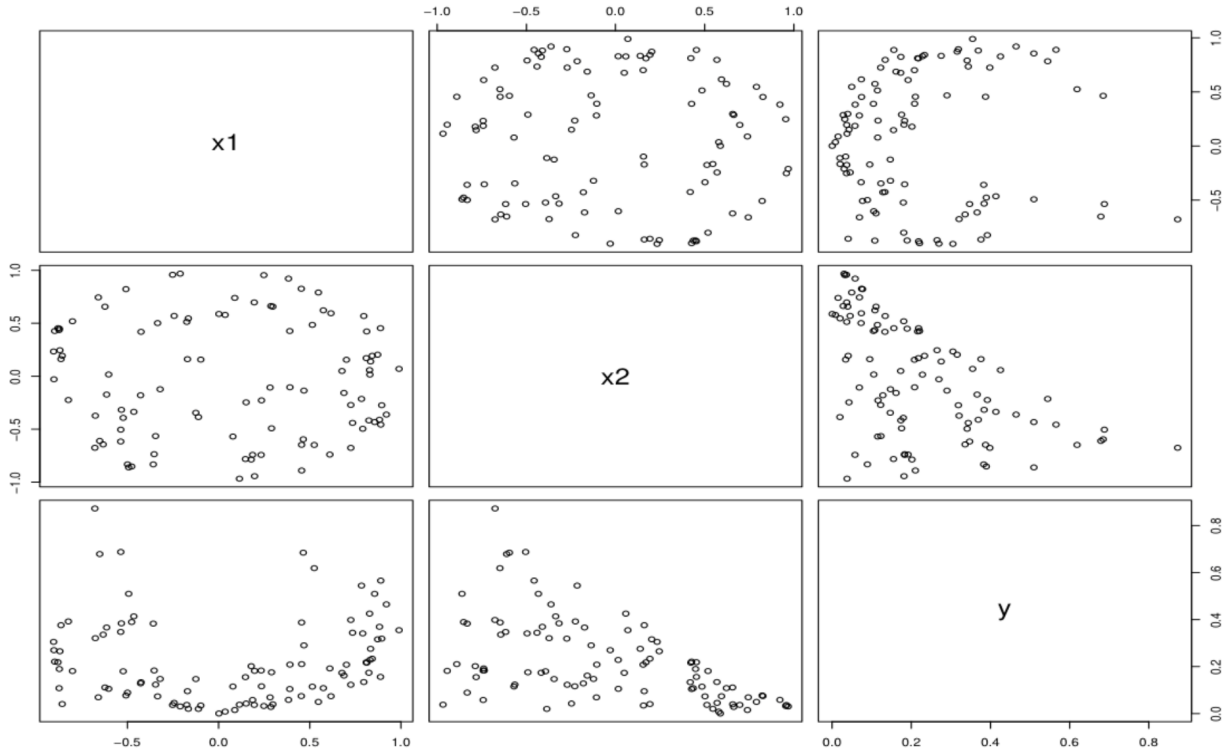


Figure 9.2 Residual plots: (a) null plot; (b) right-opening megaphone; (c) left-opening megaphone; (d) double outward bow; (e) and (f) nonlinearity; (g) and (h) combinations of nonlinearity and nonconstant variance function.

Example: *Caution Data* (ALR Package)



Scatterplot matrix of raw data.

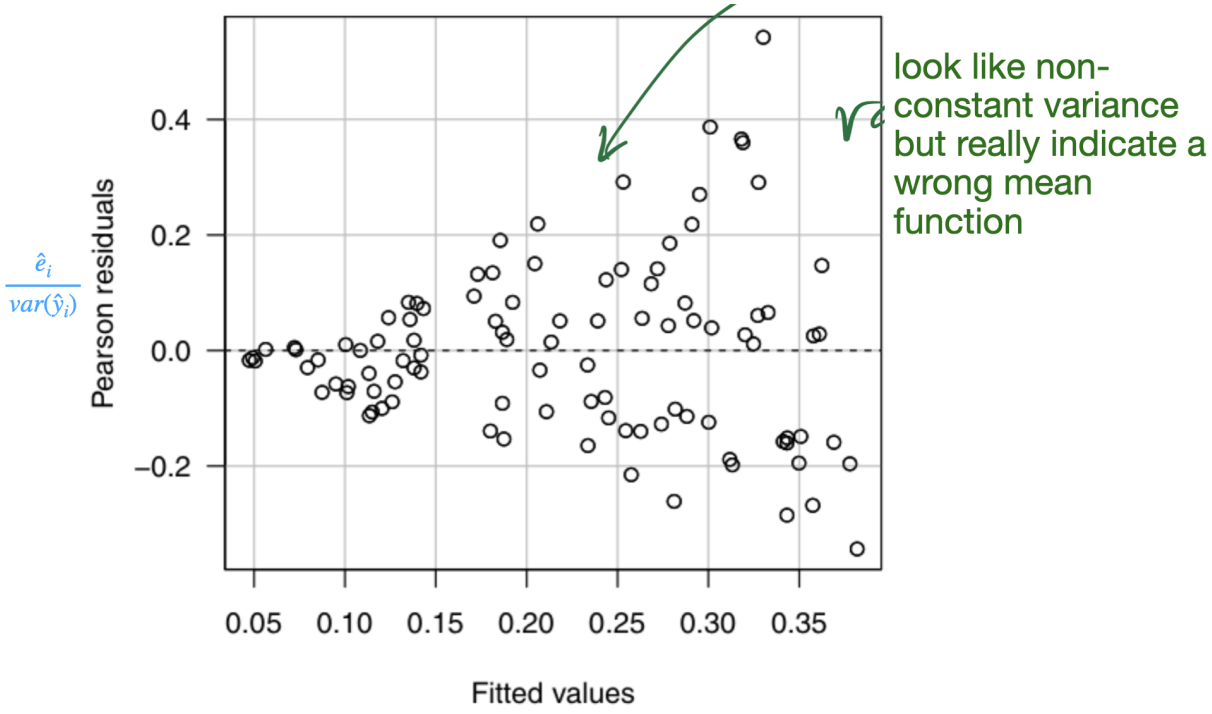


Figure 9.3 Residual plot for the `caution` data.

From ALR package

We fit: $E(Y|X = x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

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

Example: Fuel Consumption 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. Therefore, many plots of residuals should be examined.

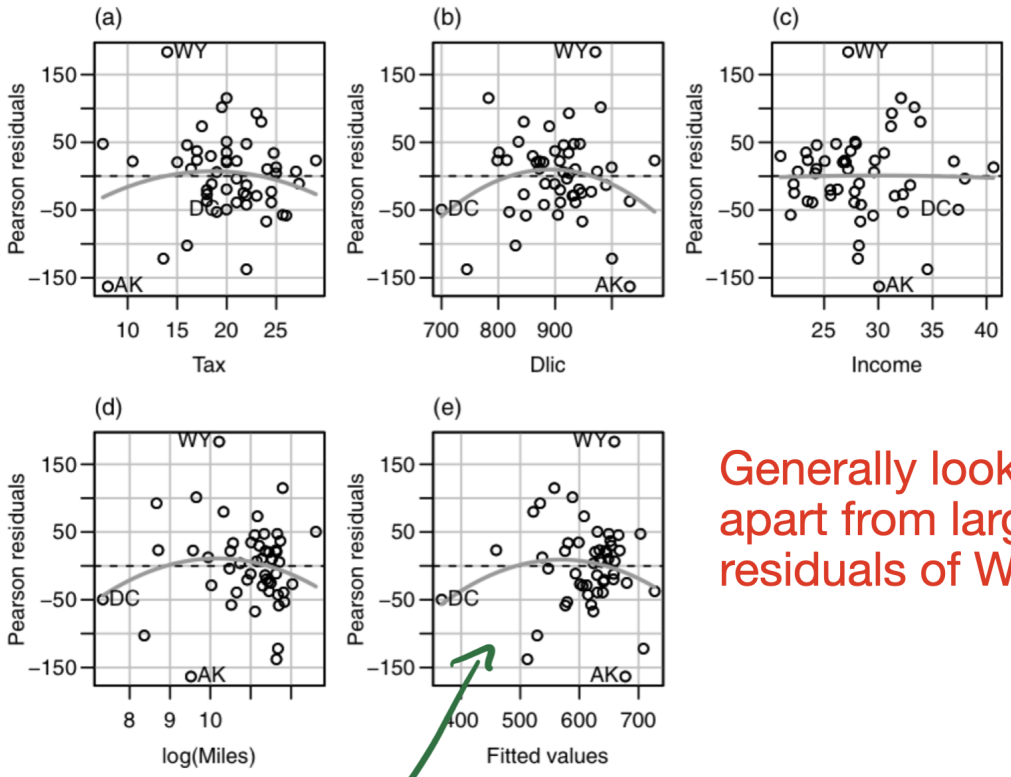


Figure 9.4 Residual plots for the fuel consumption data. The curves are quadratic fits used in lack-of-fit testing.

Testing for Curvature

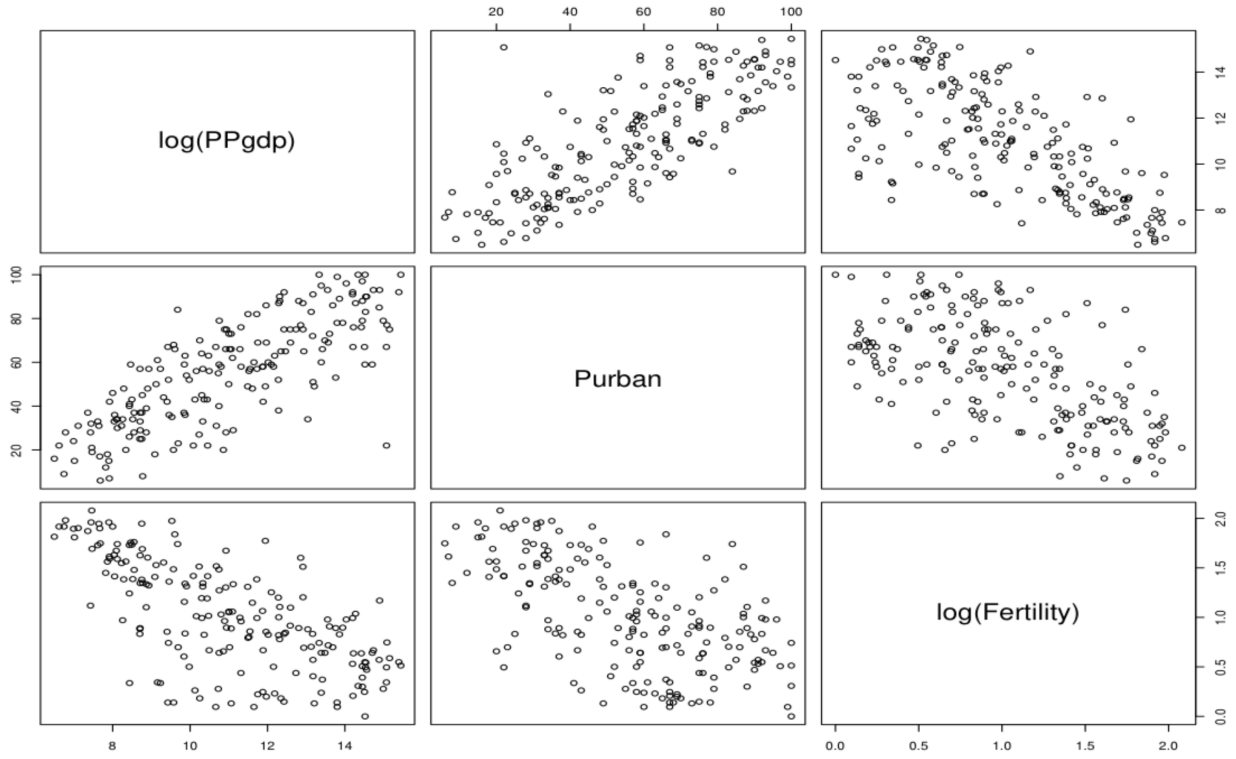
Suppose we have a plot of residuals $\hat{\mathbf{e}}$ versus a quantity U on the horizontal axis, where U could be a term in the mean function or a combination of terms. A simple test for curvature is to refit the original mean function with an additional term for U^2 added. The test for curvature is 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, Tukey's test for nonadditivity can be used.

	Test Stat	p-Value
Tax	-1.08	0.29
Dlic	-1.92	0.06
Income	-0.08	0.93
log(Miles)	-1.35	0.18
Tukeytest	-1.45	0.15

Example: UN Data

Potential model:

$$E(\text{fertility} \mid \log(\text{ppgdp}), \text{pcturban}) = \beta_0 + \beta_1 \log(\text{ppgdp}) + \beta_2 \text{pcturban}.$$



raw scatterplot matrix.

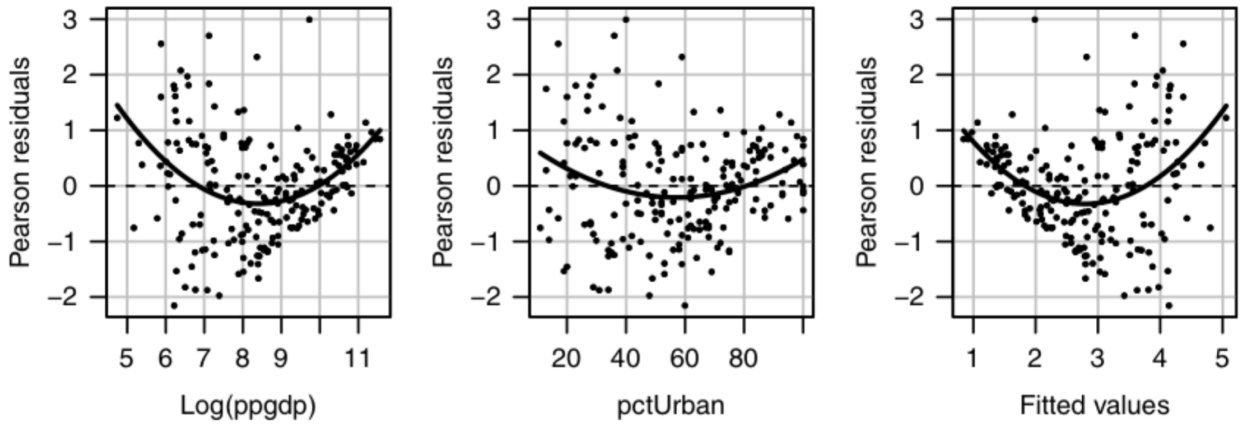


Figure 9.5 Residual plots for the UN data. The curved lines are quadratic polynomials fit to the residual plot and do not correspond exactly to the lack-of-fit tests that add a quadratic regressor to the original mean function.

Curvature suggested.

Lack-of-Fit Tests (Curvature)

	Test Stat	p-Value
log(ppgdp)	5.41	0.000
pctUrban	3.29	0.001
Tukey test	5.42	0.000

All p-values ≈ 0:

$\Longrightarrow$ Assumed mean function is not adequate for these data.

Model that Worked

$$E(\text{fertility} \mid \log(\text{ppgdp}), \text{pcturban})$$

$$= \beta_0 + \beta_1 \log(\text{ppgdp}) + \beta_2 \text{pcturban} + \beta_3 \log(\text{ppgdp}) \cdot \text{pcturban}$$

$\log(\text{ppgdp}) \cdot \text{pcturban}$: 2-way interaction

- Adding quadratic terms resulted only in minor improvements.
- But made interpretation of predictors challenging.

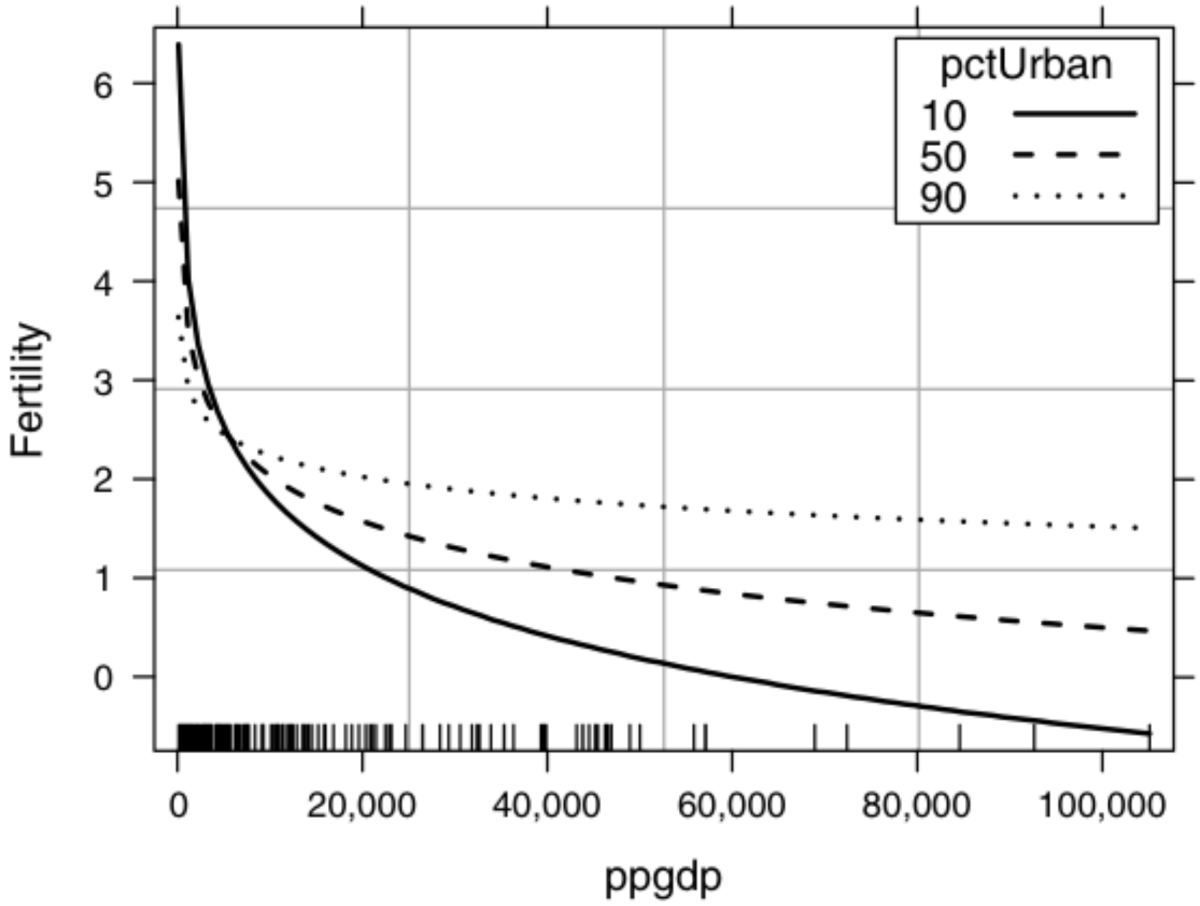


Figure 9.6 Effects plot for the UN data with an interaction.

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 available, 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.

Variance Stabilizing Transformations

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

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

where $g(E(Y \mid X = x))$ is a function that is increasing with the value of its argument. For example, if the distribution of $Y \mid X$ has a Poisson distribution, then $g(E(Y \mid X = x)) = E(Y \mid X = x)$.

For distributions in which the mean and variance are functionally related, Scheffé (1959) provides a general theory for determining transformations that can stabilize variance.

Y_T	**Comments**
$\sqrt{Y}$	Used when $\text{Var}(Y \mid X) \propto \mathbb{E}(Y \mid X)$, 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 $\text{Var}(Y \mid X) \propto [\mathbb{E}(Y \mid X)]^2$. 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 $\text{Var}(Y \mid X) \propto [\mathbb{E}(Y \mid X)]^4$. 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 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 1: Common variance stabilizing transformations.

A Diagnostic for Nonconstant Variance

Suppose that $\text{Var}(Y \mid X)$ depends on an unknown vector parameter λ and a known set of terms Z with observed values for the i-th case z_i. For example, 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

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

This implies that the variance depends on z and λ only through the linear combination $\lambda' z$; and if $\lambda = 0$, then $\text{Var}(Y \mid X, Z = z) = \sigma^2$. 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 $\lambda' z$ would lead to very similar inference.
Assuming that errors are normally distributed, a score test of $\lambda = 0$ can be carried out using the following steps:

1. Compute the OLS fit with the mean function

$$\mathbb{E}(Y \mid X = x) = \beta' x,$$

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

2. Compute scaled squared residuals

$$u_i = n\hat{e}_i^2 / \sum_{i=1}^{n} \hat{e}_j^2.$$

   Combine the u_i into a variable U.

3. Compute the regression with the mean function

$$\mathbb{E}(U \mid Z = z) = \lambda_0 + \lambda' z.$$

   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.

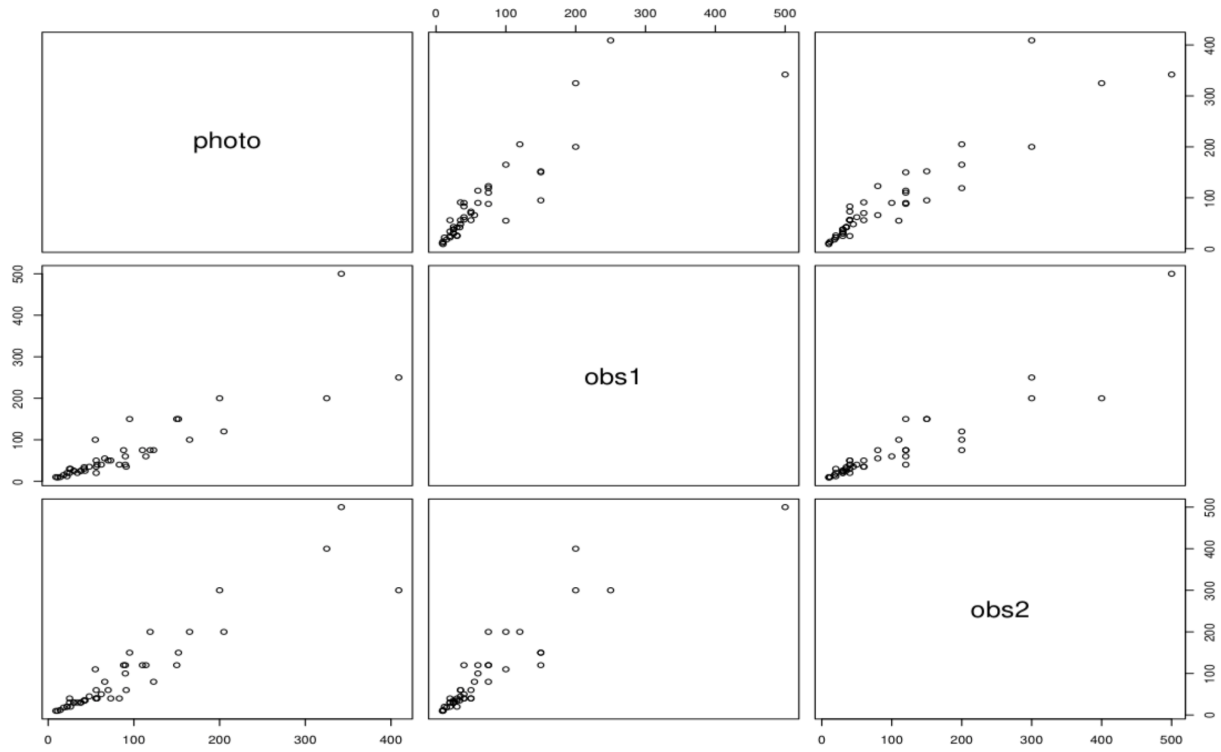
4. Compute the score test,

$$S = SS_{\text{reg}}/2.$$

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

Example: Snow Geese

The relationship between *photo* = photo count, *obs*1 = count by observer 1, *obs*2 = count by observer 2 of flocks of snow geese in the Hudson Bay area of Canada is explored.



- Substantial disagreement between observers.
- Observers cannot predict count well.
- Variance is larger for larger flocks.

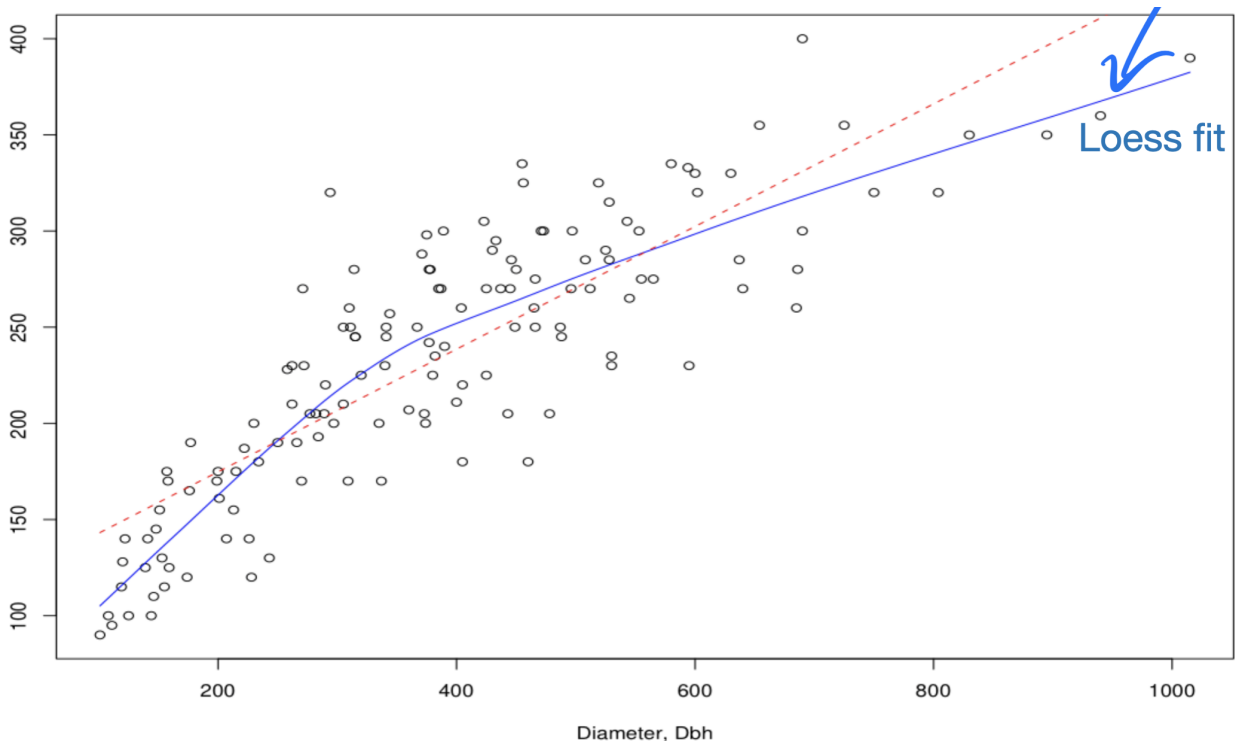
Use the first observer only, we illustrate computation of the score test for constant variance. The first step is to fit the OLS regression of photo on obs1. The fitted mean function is $\hat{E}(\text{photo} \mid \text{obs1}) = 26.55 + 0.88\text{obs1}$. From this, we can compute the residuals and u_is, and regress U on obs1. The score test for nonconstant variance is $S = \frac{SS_{\text{reg}}}{2} = 81.41$, 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.

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 be reformulated to find a better fitting model. A closely related problem is assessing how well a model matches the data. We now look at this issue from a graphical point of view using marginal model plots.

Example



$$E(\text{Height} \mid \text{Dbh}) = \beta_0 + \beta_1 \text{Dbh}$$

Looks like a poor fit (red dashed line).

Checking Mean Functions

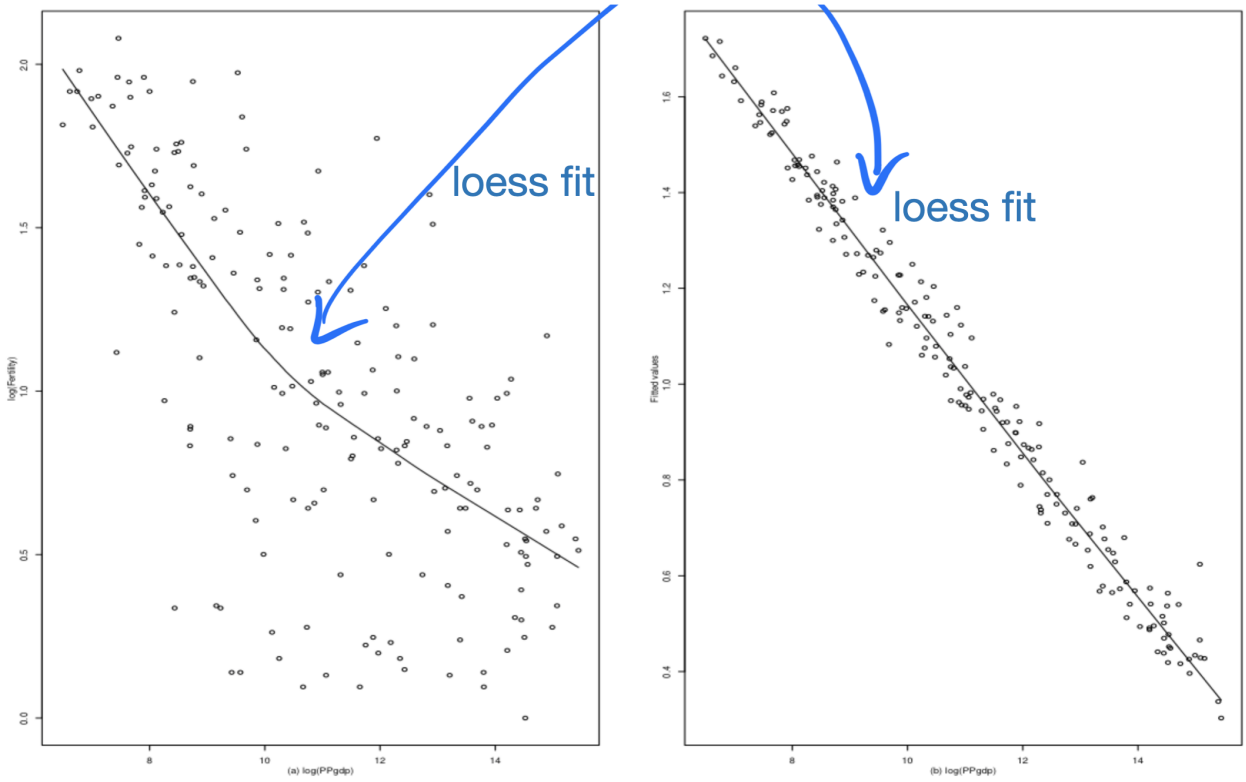
With more than one predictor, we will look at marginal models to get a sequence of two-dimensional plots to examine. We will draw a plot with the response Y on the vertical axis. On the horizontal axis, we will plot a quantity U that will consist of any function of X we think is relevant, such as fitted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

Under the model, we have $E(Y \mid U = u) = E\left[E(Y \mid X = x) \mid U = u\right]$.

This implies $E(Y \mid U = u) \approx E[\hat{Y} \mid U = u]$.

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

Example: UN Data



If the correct mean function is specified, these two smooths estimate the same quantity.

Checking Variance Functions

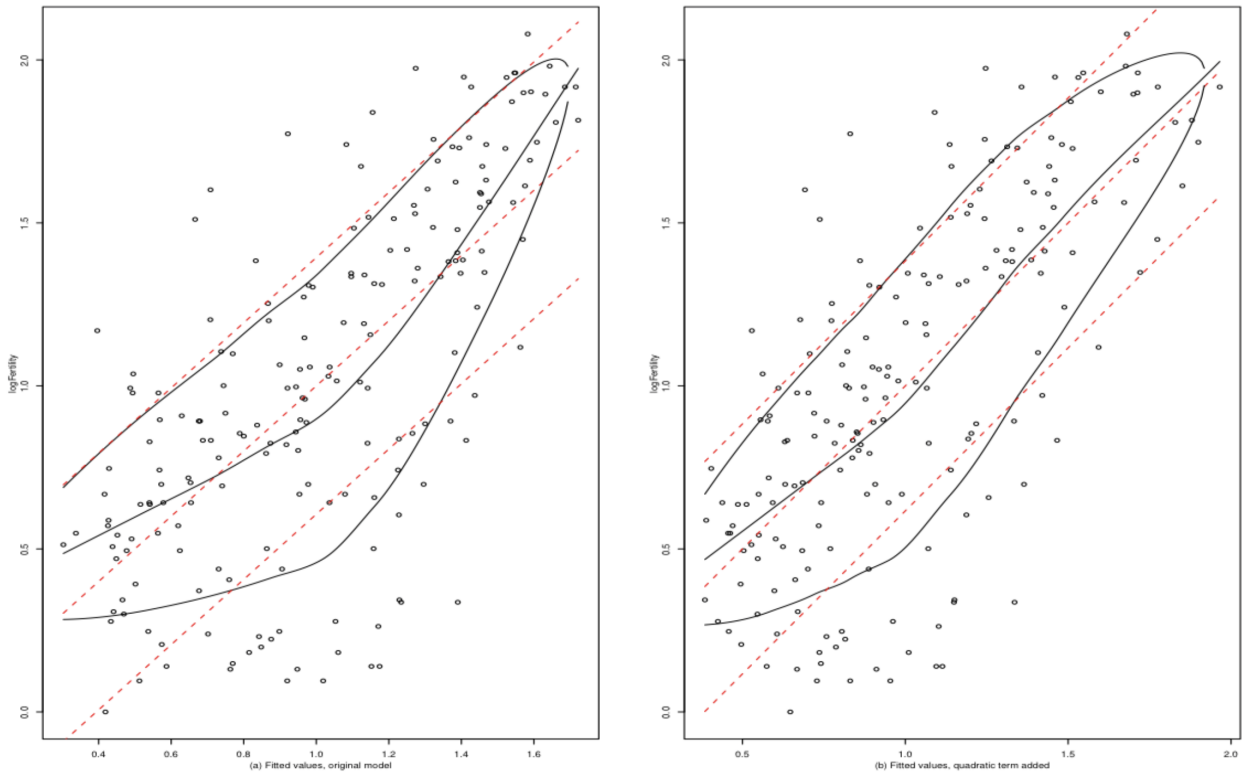
Model checking plots can also be used to check for model inadequacy in the variance function, which for the multiple linear regression problem means checking the constant variance assumption. The checking is based on the following calculation:

$$\begin{aligned}\mathrm{Var}(Y \mid U) &= \mathbb{E}[\mathrm{Var}(Y \mid X) \mid U] + \mathrm{Var}[\mathbb{E}(Y \mid X) \mid U] \\ &\approx \mathbb{E}(\sigma^2 \mid U) + \mathrm{Var}(\hat{Y} \mid U) \\ &= \sigma^2 + \mathrm{Var}(\hat{Y} \mid U)\end{aligned}$$

holds for linear regression model where $\text{var}(Y \mid X) = \sigma^2$ (constant).

According to this result, we can estimate $\text{Var}(Y \mid U)$ under the model by getting a variance smooth of $\hat{Y}$ versus U, and then adding to this an estimate of $\hat{\sigma}^2$ from the OLS fit of the model. We will call the square root of this estimated variance function $SD_{\text{model}}(Y \mid U)$. If the model is appropriate for the data, then apart from sampling error, $SD_{\text{data}}(Y \mid U) = SD_{\text{model}}(Y \mid U)$, 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 $\pm SD_{\text{data}}(Y \mid U)$ using solid lines and the mean function estimated from the model $\pm SD_{\text{model}}(Y \mid U)$ using dashed lines.



Marginal model plots with SD smooths added.

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 $E(Y \mid X = x_j) = x_j'\beta$, but for case i the mean function is $E(Y \mid X = x_i) = x_i'\beta + \delta$. 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 $\mathrm{Var}(Y \mid X) = \sigma^2$.

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

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

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

$$\mathrm{Var}(y_i - \hat{y}_{i(i)}) = \sigma^2 + \sigma^2 x_i'(X_{(i)}' X_{(i)})^{-1} x_i,$$

4. Assuming normal errors, a Student's t-test for the hypothesis $\delta = 0$ is given by

$$t_i = \frac{y_i - \hat{y}_{i(i)}}{\hat{\sigma}_{(i)} \sqrt{1 + x_i'(X_{(i)}' X_{(i)})^{-1} x_i}}.$$

   This test has $n - p' - 1$ df.

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

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

where 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.13, one can show that

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

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

A.13: Case Deletion in Linear Regression(from ALR text)

$$(X'_{(i)}X_{(i)})^{-1} = (X'X)^{-1} + \frac{(X'X)^{-1}x_i x_i'(X'X)^{-1}}{1 - h_{ii}},$$

Based on the famous Sherman–Morrison–Woodbury formula.

$$h_{ii} = x_i'(X'X)^{-1}x_i$$

$$\Longrightarrow \hat{\beta}_{(i)} = \hat{\beta} - \frac{(X'X)^{-1}x_i\hat{e}_i}{1 - h_{ii}}.$$

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 $65(.00077) = 0.05$.

Example: Forbes Data

Recall: James D. Forbes (1809–1868) did a series of experiments exploring the relationship between atmospheric pressure and boiling point of H_2O.

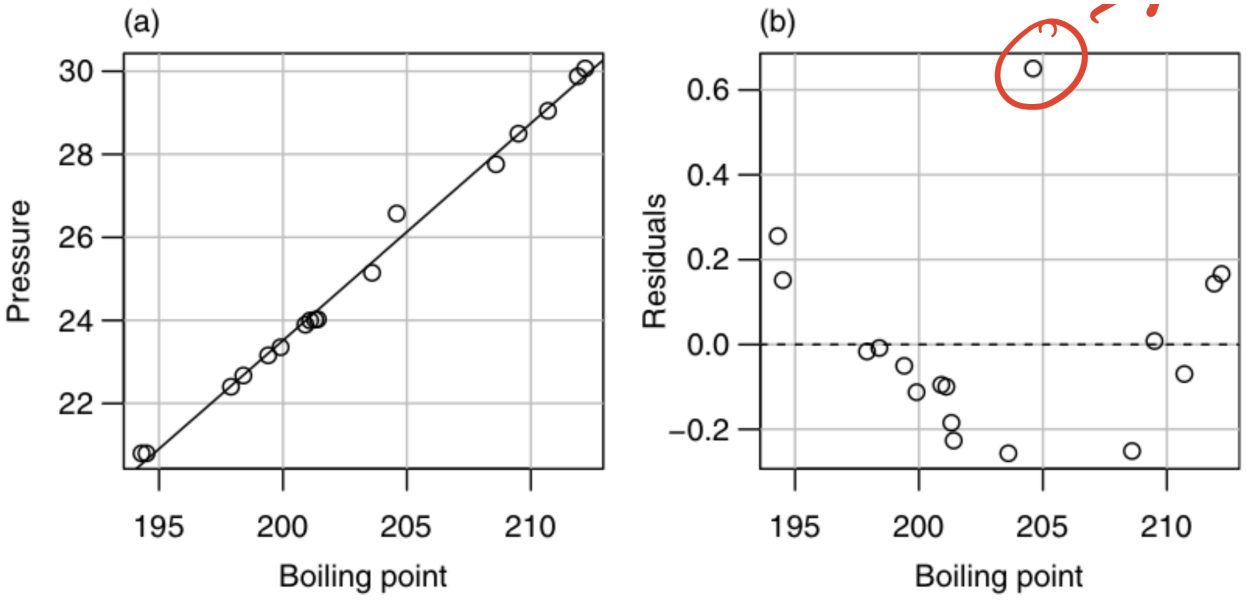


Figure 1.3 Forbes data: (a) `pres` versus `bp`; (b) residuals versus `bp`.

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 without the outlier.

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

Influence of Cases

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.

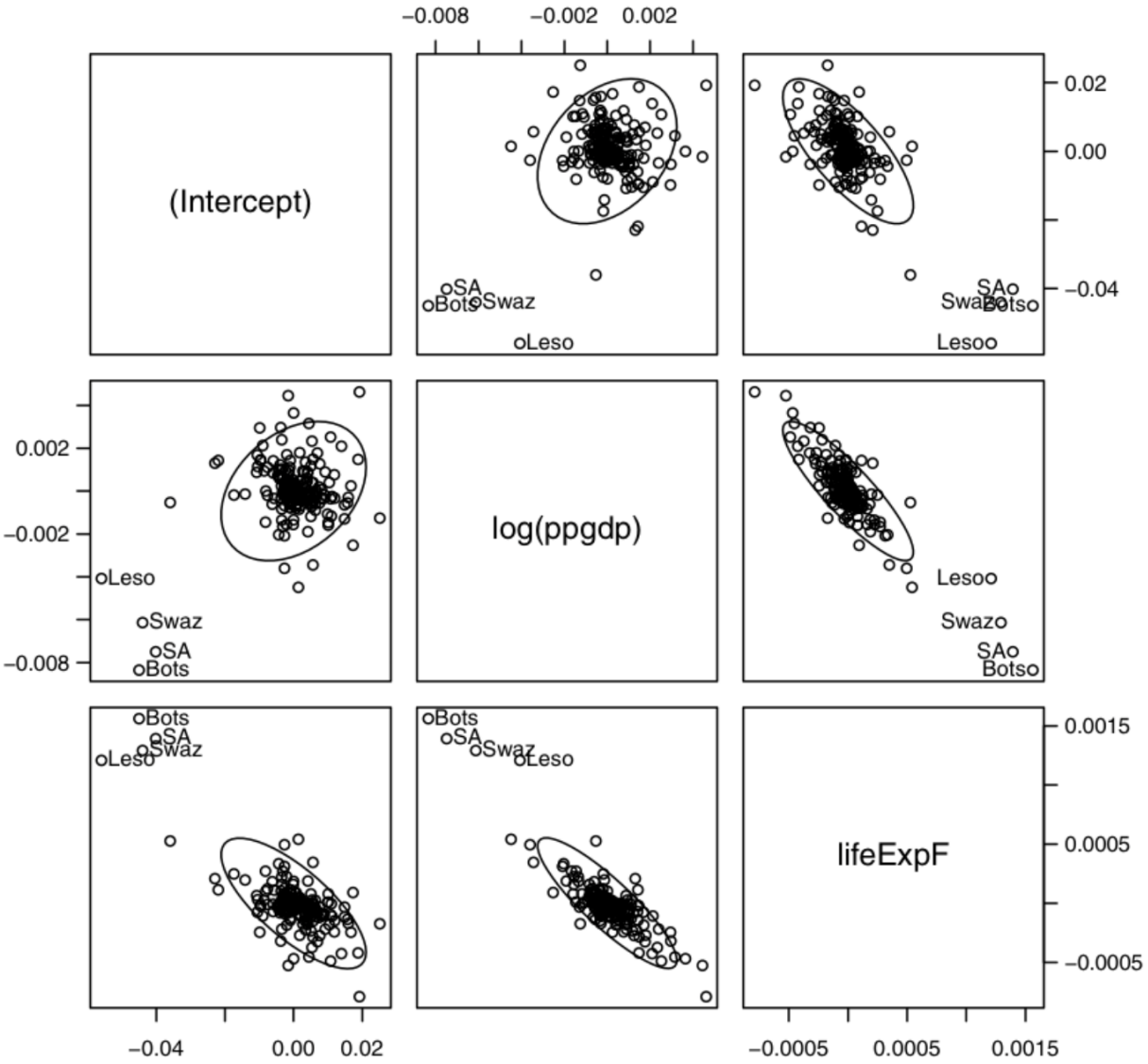


Figure 9.7 Estimates of parameters in the UN data obtained by deleting one case at a time. The ellipses shown on the plots would be 95% confidence regions for the bivariate mean in each plot if the points in the plot were a sample from a bivariate normal distribution.

Cook's Distance

The influence on the estimates of β (info in plot above) can be summarized by comparing $\hat{\beta}$ and $\hat{\beta}_{(i)}$.

Cook's distance (Cook, 1977) is given by

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})'(X'X)(\hat{\beta}_{(i)} - \hat{\beta})}{p'\hat{\sigma}^2} = \frac{(\hat{Y}_{(i)} - \hat{Y})'(\hat{Y}_{(i)} - \hat{Y})}{p'\hat{\sigma}^2}.$$

Cases for which D_i is large have substantial influence on both the estimate of β 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 datasets, 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.

Example: 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 liver (Y) and the body weight, liver weight, and relative dose.

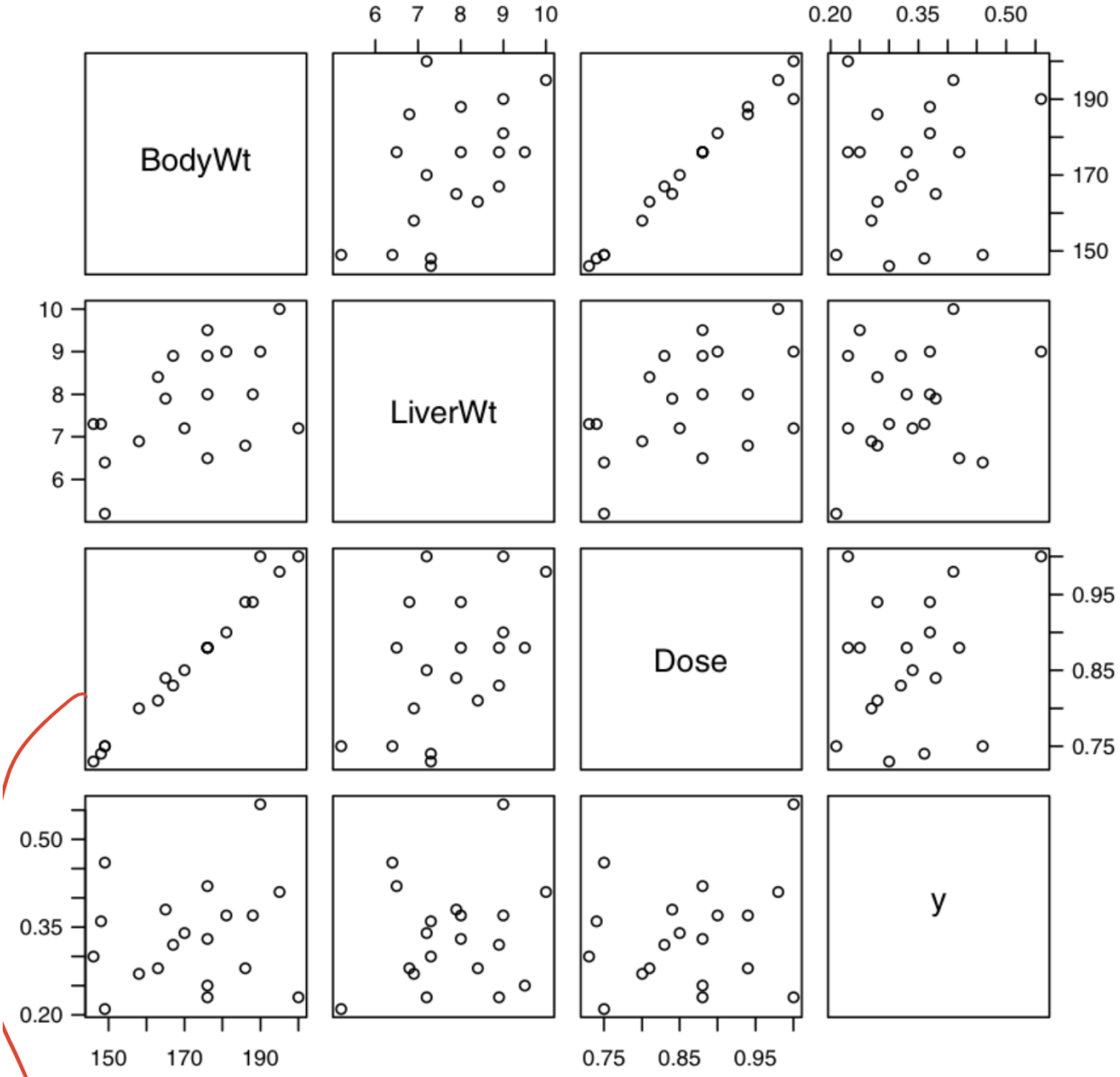


Figure 9.8 Scatterplot matrix for the rat data.

Nearly perfectly linearly related!

Table 9.2 Regression Summary for the Rat Data

	Estimate	Std. Error	t Value	Pr(>|t|)
`(Intercept)`	0.2659	0.1946	1.37	0.1919
`BodyWt`	–0.0212	0.0080	–2.66	0.0177
`LiverWt`	0.0143	0.0172	0.83	0.4193
`Dose`	4.1781	1.5226	2.74	0.0151

$\hat{\sigma} = 0.0773$ with 15 df, $R^2 = 0.3639$.

But SLR models all NS!

Using Diagnostics to Resolve the "Paradox"

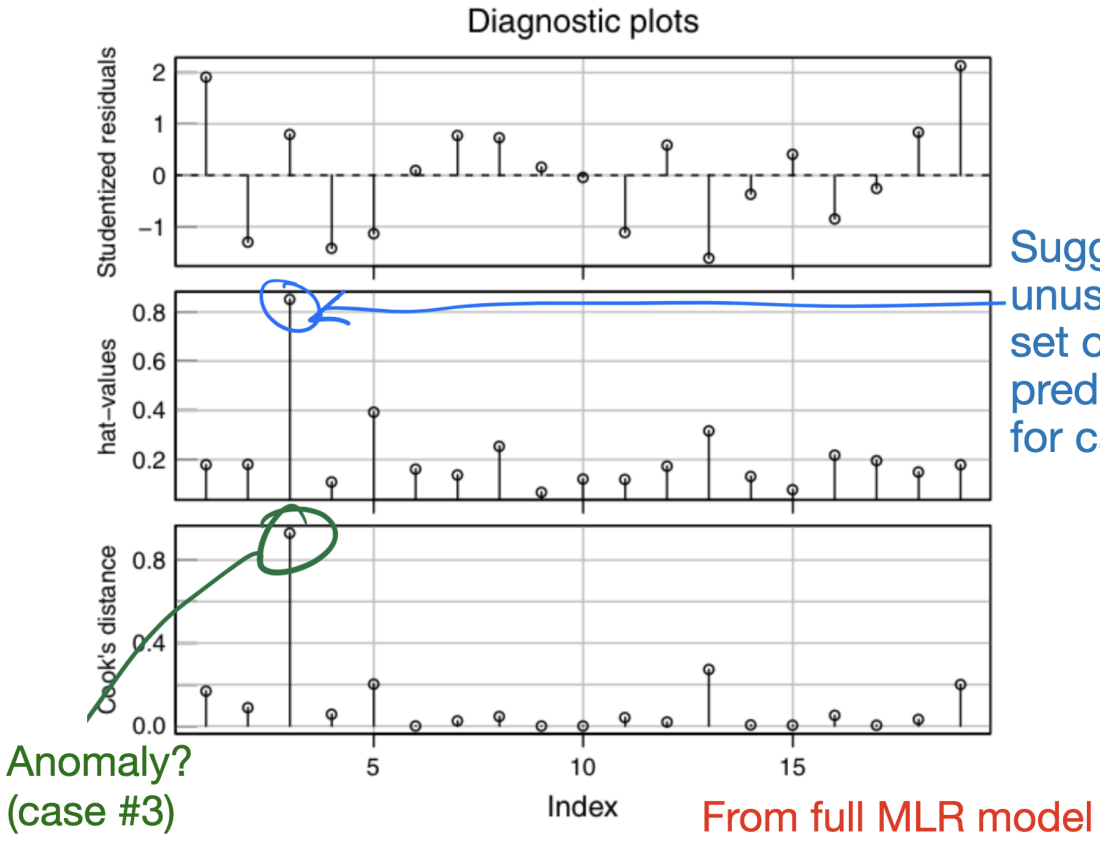


Figure 9.9 Diagnostic statistics for the rat data.

Table 9.3 Regression Summary for the Rat Data with Case 3 Deleted

	Estimate	Std. Error	t Value	Pr(>|t|)
(Intercept)	0.3114	0.2051	1.52	0.1512
BodyWt	–0.0078	0.0187	–0.42	0.6838
LiverWt	0.0090	0.0187	0.48	0.6374
Dose	1.4849	3.7131	0.40	0.6953

$\hat{\sigma} = 0.0782$ with 14 df, $R^2 = 0.0211$.

Rat number 3, with weight 190g, 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.

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:

$$\hat{e} = (I - H)Y = (I - H)(X\beta + e) = (I - H)e.$$

In scalar form, the ith residual is: $\hat{e}_i = e_i - \sum_{j=1}^{n} h_{ij}e_j$.
By CLT will nearly normal even if e_i not normally distributed. 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 $n \uparrow$ for fixed p', second term has smaller variance than first term so becomes less important.
$\therefore \hat{e}_i$ can be used to test for normality.

Normal Probability Plot

Suppose we have a sample of n numbers $z_1, z_2, \ldots, z_n$, 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 z's 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.

Example for illustration only (particular dataset not of interest here)

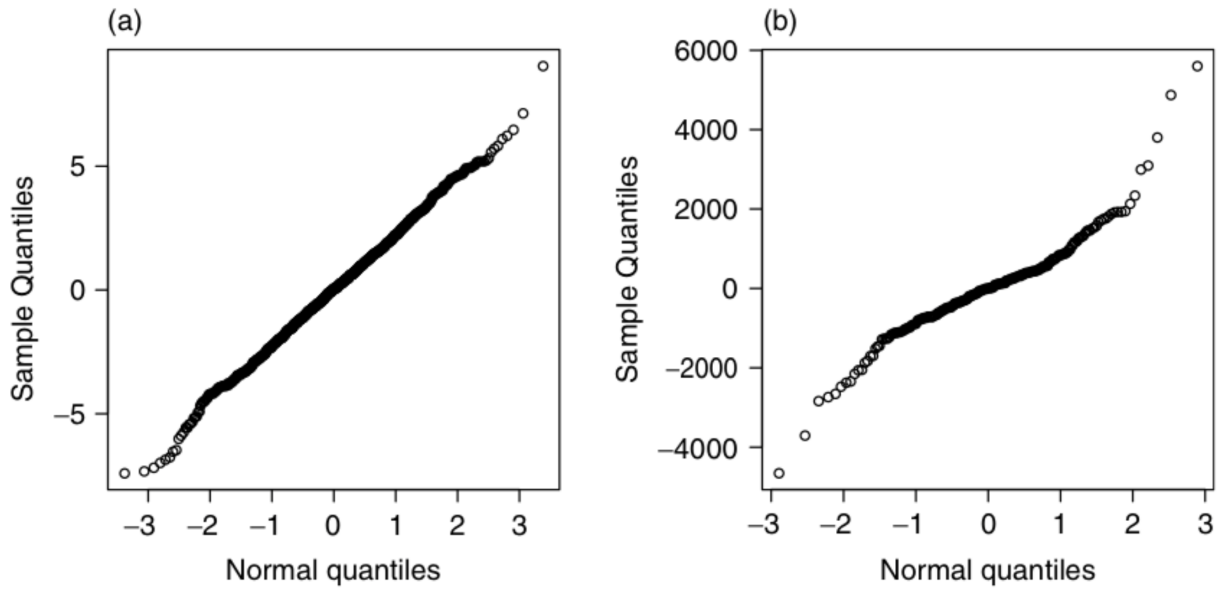


Figure 9.11 Normal probability plots of residuals for (a) the heights data and (b) the transactions data.

Detection of Multicollinearity

Source: Yibi Huang, University of Chicago, 2022
(from Chatterjee and Hadi (2005), *Regression Analysis by Example*, Wiley, New Jersey)

Predictors of an MLR Model Cannot Be Linearly Dependent

MLR requires predictors to be **linearly independent**, i.e., no predictor can be expressed as a linear combination of others.
- **Example:** $X_1 = \#$ of undergrads, $X_2 = \#$ of grads, $X_3 = \#$ of students, then X_1, X_2, X_3 are linearly dependent since $X_1 + X_2 = X_3$.
No unique LS estimates for coefficients if predictors are linearly dependent.
Example: If $X_1 + X_2 = X_3$, then in the model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

$$Y = \beta_0 + (\beta_1 + \beta_3) X_1 + (\beta_2 + \beta_3) X_2 + \epsilon$$

The coefficients $(\beta_0, \beta_1, \beta_2, \beta_3)$ and $(\beta_0, \beta_1', \beta_2', \beta_3')$ give identical predictions for Y if: $\beta_1' = \beta_1 + \beta_3, \quad \beta_2' = \beta_2 + \beta_3, \quad \beta_3' = 0.$

The Problem of Multicollinearity (MC)

- **Multicollinearity (MC)** means predictors in an MLR model have a **close-to-exact linear relationship**, i.e., predictors are nearly linearly dependent.
- When MC problem exists, the LS estimates for β_j exist but would have **large variability**.

- Recall: We interpret $\hat{\beta}_j$ as the mean response change when X_j increases by one unit **holding all other predictors fixed**.
  If predictors are strongly correlated, it may be impossible to alter X_j while holding other predictors fixed.

To examine the existence (or lack) of equal educational opportunities in public educational institutions, the following variables were measured for 70 schools selected at random in 1965:

- **ACHV:** Student achievement index (higher values are better).
- **FAM:** Faculty credentials index.
- **PEER:** The influence of their peer group in the school.
- **SCHOOL:** School facility/resource index.

Goal: To identify important determinants of student achievement.

$$\text{Model:} \quad \text{ACHV} = \beta_0 + \beta_1 \cdot \text{FAM} + \beta_2 \cdot \text{PEER} + \beta_3 \cdot \text{SCHOOL} + \epsilon.$$

```
summary(lm(ACHV ~ FAM + PEER + SCHOOL, data=EEO))
```

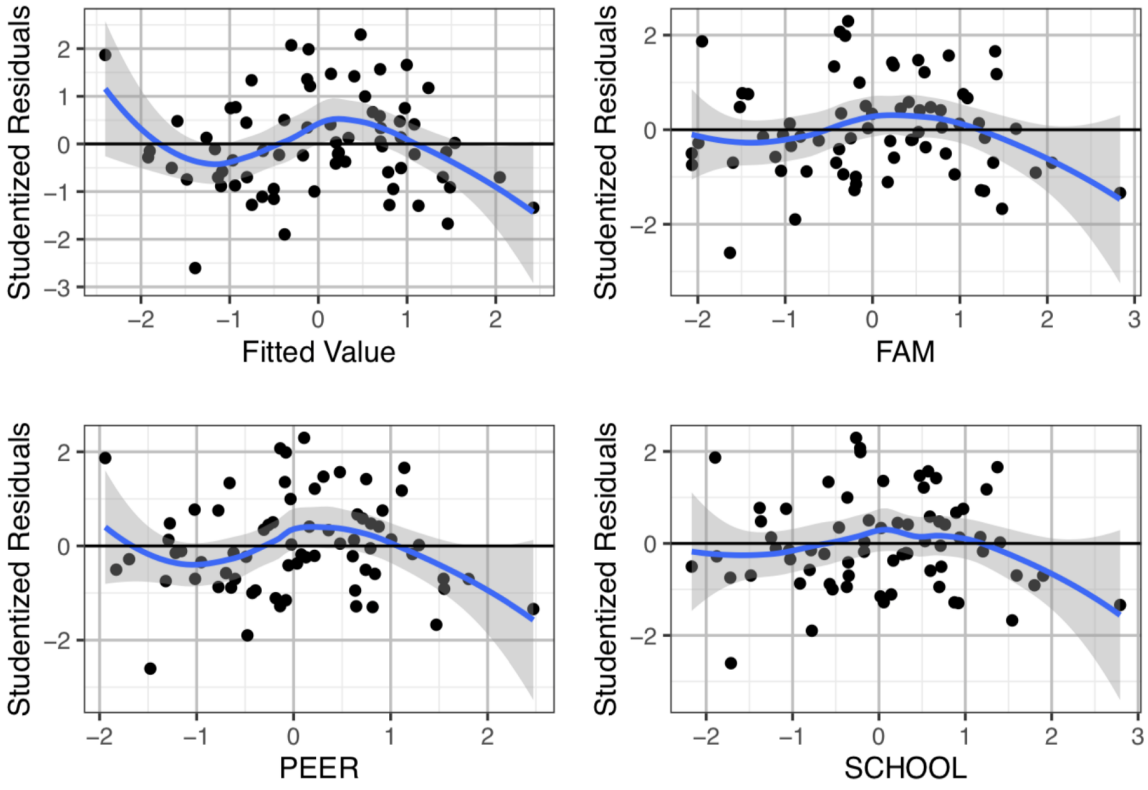
	Estimate	**Std. Error**	**t value**	**Pr($>\|t\|$)**	
(Intercept)	-0.06996	0.25064	-0.279	0.781	
FAM	1.10126	1.41056	0.781	0.438	Residual standard er-
PEER	2.32206	1.48129	1.568	0.122	
SCHOOL	-2.28100	2.22045	-1.027	0.308	

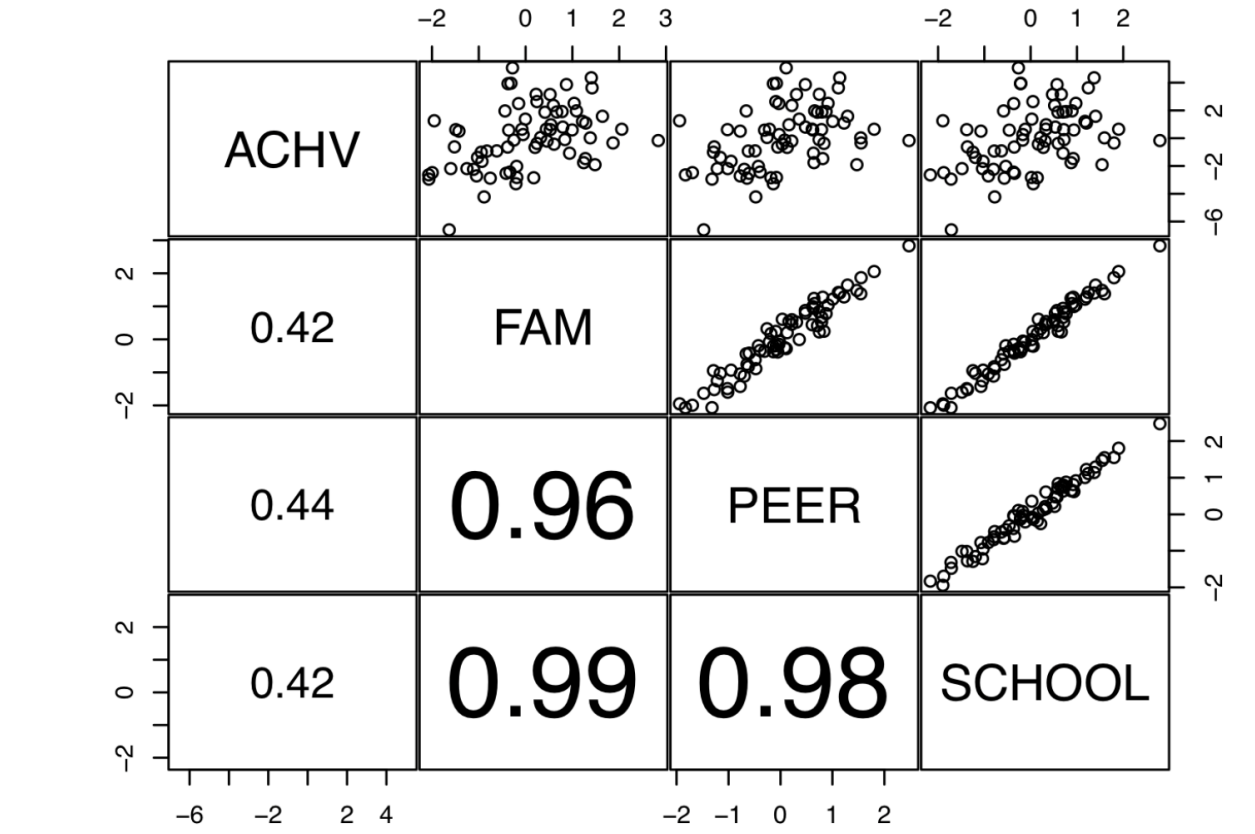
ror: 2.07 on 66 degrees of freedom.
Multiple R-squared: 0.2063, Adjusted R-squared: 0.1702.
F-statistic: 5.717 on 3 and 66 DF, p-value: 0.001535.

- None of the 3 predictors is significant, but the overall model F-statistic is significant (i.e., at least one of the 3 predictors is significant).

- The coefficient of **SCHOOL** is **negative**! Normally, we expect higher student achievement if schools have more resources.

All Residual Plots Look Fine





All 3 Predictors Are Highly Correlated!

- Correlations between the 3 predictors are extremely high!
- Knowing anyone of 3 predictors, we can predict the other 2 very accurately.
- So it's almost like we have only one predictor. Only 1 of the 3 predictors is really needed.
- However, multicollinearity prevents us from identifying the important predictors.

Significance of Individual Predictors

```
summary(lm(ACHV ~ FAM, data=EEO))$coef
```

	Estimate	**Std. Error**	**t value**	**Pr($>\|t\|$)**
Intercept	-0.02427	0.2486	-0.09761	0.922526
FAM	0.88010	0.2310	3.81036	0.000301

```
summary(lm(ACHV ~ PEER, data=EEO))$coef
```

	Estimate	**Std. Error**	**t value**	**Pr(¿—t—)**
Intercept	-0.03087	0.2460	-0.1255	0.900525
PEER	1.08090	0.2676	4.0387	0.000139

```
summary(lm(ACHV ~ SCHOOL, data=EEO))$coef
```

	Estimate	**Std. Error**	**t value**	**Pr(¿—t—)**
Intercept	-0.01043	0.2487	-0.04194	0.9666696
SCHOOL	0.92834	0.2446	3.79540	0.0003163

Observations:

- The multiple R^2 values for the three models are 0.1759, 0.1935, and 0.1748 respectively.
- These values are close to the multiple R^2 of 0.2063 for the full model including all three predictors.

ANOVA Comparisons

```
lmFPS = lm(ACHV ~ FAM+PEER+SCHOOL, data=EEO)
lmF = lm(ACHV ~ FAM, data=EEO)
lmP = lm(ACHV ~ PEER, data=EEO)
lmS = lm(ACHV ~ SCHOOL, data=EEO)
anova(lmF, lmFPS)
```

	Res.Df	RSS	Df	Sum of Sq	F	$Pr(>F)$
1	68	294				
2	68	283	2	10.8	1.26	0.29

```
anova(lmP, lmFPS)
```

	Res.Df	RSS	Df	Sum of Sq	F	$Pr(>F)$
1	68	287				
2	68	283	2	4.56	0.53	0.59

```
anova(lmS, lmFPS)
```

	Res.Df	RSS	Df	Sum of Sq	F	$Pr(>F)$
1	68	294				
2	68	283	2	11.2	1.31	0.28

Observations:

- All three single-predictor models fit the data nearly as well as the model including all 3 predictors.
- It is difficult to identify which predictor is more important.

Summary: Effects of Multicollinear Data

- Trouble in identifying important predictors.
- Estimates are very sensitive to what other variables exist in the model.
  - The estimated coefficient of **SCHOOL** changes from -2.281 to 0.928 when **FAM** and **PEER** are removed from the model.

```
lm(ACHV ~ FAM+PEER+SCHOOL, data=EEO)$coef
```

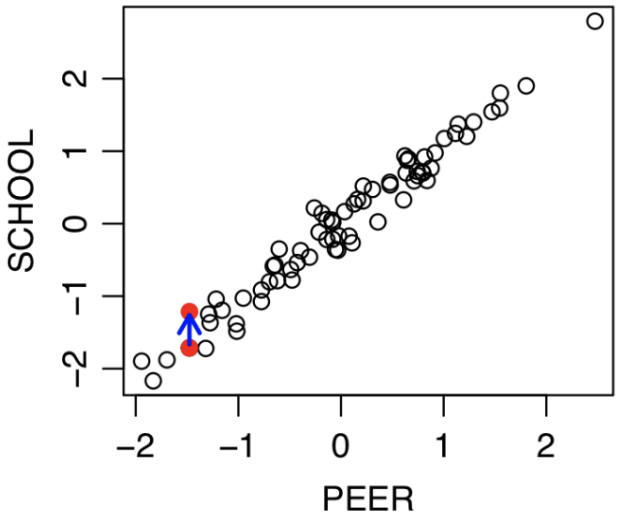
Intercept	FAM	PEER	SCHOOL
-0.06996	1.10126	2.32206	-2.28100

```
lm(ACHV ~ SCHOOL, data=EEO)$coef
```

Intercept	SCHOOL
-0.01043	0.92834

Conclusion: Estimated coefficients are very sensitive to small changes in data.

If the value of SCHOOL of the 28th school is decreased by 0.5 from -1.713 to -2.213, observe the estimates for β's are changed drastically!




```
EEO.new = EEO
EEO.new$SCHOOL[28] = EEO$SCHOOL[28] - 0.5
lm(ACHV ~ FAM+PEER+SCHOOL, data=EEO.new)$coef
```

Intercept	FAM	PEER	SCHOOL
-0.02081	-0.28845	0.94995	0.41394

Original Estimates:

Intercept	FAM	PEER	SCHOOL
-0.06996	1.10126	2.32206	-2.28100

Why Multicollinearity Makes Predictors Insignificant

Model:

$$\text{ACHV} = \beta_0 + \beta_1 \cdot \text{FAM} + \beta_2 \cdot \text{PEER} + \beta_3 \cdot \text{SCHOOL} + \epsilon$$

Steps:

1. Regress **ACHV** on **PEER** and **SCHOOL**.
2. Regress **FAM** on **PEER** and **SCHOOL**.
3. Fit a simple linear regression model (SLR) using the residuals from Step 1 as the response and the residuals from Step 2 as the predictor.

Insight:

- Recall in SLR:
$$s.e.(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum_i (x_i - \bar{x})^2}}$$
- When **FAM** is highly collinear with **PEER** and **SCHOOL**, the residuals in Step 2 are nearly 0, resulting in $SD \approx 0$.
- This makes $s.e.(\hat{\beta}_1)$ huge $\Rightarrow$ small t-value $\Rightarrow$ insignificant predictor.

Looking for Multicollinearity

- Multicollinearity is...
  - Associated with unstable coefficient estimates.
  - A result of linear relationships between predictors.
  - Not due to model mis-specification.
- We do not worry about fixing multicollinearity until the model diagnostics of other assumptions are satisfactory.
  - Some indications of multicollinearity arise during the process of adding and removing variables and altering or removing observations.

Signs of Multicollinearity

While finding a good model, look for instability in estimated $\hat{\beta}_j$:

- Large changes in some $\hat{\beta}_j$ when a variable is added or deleted.
- Large changes in some $\hat{\beta}_j$ when a data point is altered or dropped.

Once the model fit is good, look for:

- Signs of some $\hat{\beta}_j$ do not conform to prior expectations.
- Coefficients or variables that are expected to be important have large standard errors (small t values).

Examples:

- Standard errors for all predictors are high in the EEO Data, resulting in small t-values. We would expect all three predictors to be important.
- t-value for **DOPROD** was small and negative, when we would expect it to be positive and important.

Multicollinearity and Correlation

- Of course, the pairwise scatterplot is helpful for detecting multicollinearity.
- Unfortunately, it only helps to detect linear relationships between pairs of variables.
- There may be a higher-level relationship even with no pairwise correlations.

Variance Inflation Factor

The variance inflation factor for a predictor X_j in the MLR model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

is defined to be:

$$\text{VIF}_j = \frac{1}{1 - R_j^2}, \quad j = 1, \ldots, p,$$

where R_j^2 is the multiple R^2 for regressing X_j on the other predictors in the model.

- If X_j is orthogonal (has 0 correlation) to all other predictors, then $R_j^2 = 0$ and $\text{VIF}_j = 1$.
- Increasing values of VIF_j indicate departure from orthogonality toward multicollinearity.
- A rule of thumb: VIFs > 10 suggest multicollinearity.

Interpreting VIF

- In simple linear regression, $\text{Var}(\hat{\beta}_1) = \sigma^2 / \sum_{i=1}^n (x_i - \bar{x})^2$.
- In multiple linear regression, the situation is complicated by the existence of correlation among the predictors. **This is what VIF measures.**
- In fact, one can show that:

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2} \cdot \text{VIF}_j.$$

- VIF_j indicates the proportional increase in $\text{Var}(\hat{\beta}_j)$ due to its collinearity with other predictors, relative to the orthogonal case.

VIF in R

The `vif()` function in the `car` library can calculate the VIF_j for us.

```
library(car)
vif(lm(ACHV ~ FAM + PEER + SCHOOL, data=EEO))
```

Output:

```
FAM     PEER   SCHOOL
37.58   30.21   83.16
```

Observe the VIF for **SCHOOL** in the model above is:

$$\text{VIF}_{\text{SCHOOL}} = \frac{1}{1-R^2} = \frac{1}{1-0.987974} \approx 83.155,$$

where $R^2 = 0.987974$ is the multiple R^2 of regressing **SCHOOL** on the other two predictors **FAM** and **PEER**.

```
summary(lm(SCHOOL ~ FAM + PEER, data=EEO))$r.squared
[1] 0.987974
```

Variance Inflation Due to Multicollinearity

Observe how much the **Std. Error** for **SCHOOL** is inflated when **FAM** and **PEER** are included.

```
summary(lm(ACHV ~ SCHOOL, data=EEO))$coef
```

Estimate	Std. Error	t value	Pr(¿—t—)
0.92834	0.2446	3.79540	0.0003163

```
summary(lm(ACHV ~ FAM + PEER + SCHOOL, data=EEO))$coef
```

Estimate	**Std. Error**	**t value**	**Pr(¿—t—)**
-2.28100	2.2204	-1.0273	0.3080

The ratio of two standard errors of **SCHOOL** in the two models is:

$$\frac{2.2204481}{0.244597} \approx 9.08, \quad \text{close to } \sqrt{\text{VIF}_{\text{SCHOOL}}} = \sqrt{83.33}.$$