Book Summary: *Regression Diagnostics: An Introduction, 2nd Edition*

Yan Zeng

Version 1.1, last updated on 2020-07-05

Abstract

Summary of Fox [19].

Contents

1 Introduction

All the problems discussed in this monograph vary in degree from trivial to catastrophic, but I view nonlinearity as intrinsically the most serious problem, because it implies that we're fitting the wrong equation to the data.

2 The Linear Regression Model: Review

2.1 The Normal Linear Regression Model

The normal linear regression model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i,\ \varepsilon_i \sim \text{NID}(0, \sigma^2).$$

where y_i is the value of the response variable for the ith of n cases; the x_{ij}'s are values of k regressors for case i; the β_j's are population regression coefficients; ε_i is the regression error for case i, with 0 expectation (mean) and constant variance σ^2; and "NID" means "normally and independently distributed".

Linearity: Because the errors have means of 0, the conditional expectation μ of the response is a linear function of the parameters and the regressors:

$$\mu_i = E[y_i|x_{i1}, \cdots, x_{ik}] = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}.$$

Constant error variance: The conditional variance of the response is the same for all cases

$$V(y_i|x_{i1}, \cdots, x_{ik}) = \sigma^2.$$

Normality: The errors are normally distributed $\varepsilon_i \sim N(0, \sigma^2)$, and hence the conditional distribution of the response is also normal $y_i \sim N(\mu_i, \sigma^2)$.

Independence: The cases are independently sampled, so that ε_i is independent of $\varepsilon_{i'}$ for $i \neq i'$, and hence y_i is independent of $y_{i'}$.

Fixed x's or x's independent of ε: The explanatory variables, and thus the regressors, are either fixed with respect to replication of the study (as would be the case in a designed experiment), or the errors are independent of the values of the x's.

The x's are not perfectly collinear: If the x's are fixed, then no x can be a perfect linear function of others; if the x's are sampled along with y, then the x's cannot be perfectly collinear in the population. If the x's are perfectly collinear, then it is impossible to separate their effects on y.

Not all the assumptions of the normal linear model are required for all purposes. For example, if the distribution of the errors is nonnormal, then the least-squares estimators of the regression coefficients may not be efficient, but they are still unbiased, and standard methods of statistical inference for constructing confidence intervals and performing hypothesis tests still produce valid results, at least in large samples.

If the x's are fixed or independent of the errors, then only the assumption of linearity is required for the least-squares coefficients to be unbiased estimators of the β's, but if the errors have different variances or are dependent, then the least-squares estimators may be inefficient and their standard errors may be substantially biased.

Finally, if the errors have different variances that are known up to a constant of proportionality, say $V(\varepsilon_i) = \sigma^2/w_i$, but the other assumptions of the model hold, then weighted least squares (WLS) regression provides efficient estimates of the β's and correct coefficient standard errors.

2.2 Least-Squares Estimation

2.3 Statistical Inference for Regression Coefficients

2.4 *The Linear Regression Model in Matrix Form

Let the column vector $\boldsymbol{x}_k$ be the T observations on variable x_k, $k = 1, \cdots, K$, and assemble these data in an $T \times K$ data matrix $\boldsymbol{X}$. In most contexts, the first column of $\boldsymbol{X}$ is assumed to be a column of 1s:

$$\boldsymbol{x}_1 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{T \times 1}$$

so that β_1 is the constant term in the model. Let $\boldsymbol{y}$ be the T observations y_1, $\cdots$, y_T, and let $\boldsymbol{\varepsilon}$ be the column vector containing the T disturbances. The **Classical Linear Regression Model** (CLRM) can be written as

$$\boldsymbol{y} = \boldsymbol{x}_1\beta_1 + \cdots + \boldsymbol{x}_K\beta_K + \boldsymbol{\varepsilon},\ \boldsymbol{x}_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iT} \end{bmatrix}_{T \times 1}$$

or in matrix form

$$\boldsymbol{y}_{T \times 1} = \boldsymbol{X}_{T \times K}\boldsymbol{\beta}_{K \times 1} + \boldsymbol{\varepsilon}_{T \times 1}.$$

Assumptions of the CLRM (Brooks [4, page 44], Greene [20, page 16-24]):

(1) **Linearity:** The model specifies a linear relationship between y and x_1, $\cdots$, x_K.

$$\boxed{\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}}$$

(2) **Full rank:** There is no exact linear relationship among any of the ndependent variables in the model. This assumption will be necessary for estimation of the parameters of the model (see formula (2.1)).

$$\boxed{\boldsymbol{X} \text{ is a } T \times K \text{ matrix with rank } K.}$$

(3) **Exogeneity of the independent variables:** $E[\varepsilon_i | x_{j1}, x_{j2}, \cdots, x_{jK}] = 0$. This states that the expected value of the disturbance at observation i in the sample is not a function of the independent variables observed at any observation, including this one. This means that the independent variables will not carry useful information for prediction of ε_i.

$$\boxed{E[\boldsymbol{\varepsilon}|\boldsymbol{X}] = \boldsymbol{0}.}$$

(4) **Homoscedasticity and nonautocorrelation:** Each disturbance, ε_i has the same finite variance, σ^2, and is uncorrelated with every other disturbance, ε_j.

$$\boxed{E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\boldsymbol{X}] = \sigma^2\boldsymbol{I}.}$$

(5) **Data generation:** The data in $(x_{j1}, x_{j2}, \cdots, x_{jK})$ may be any mixture of constants and random variables. That is, the regressors are either fixed with respect to replication of the study (as would be the case in a designed experiment), or the errors ε's are independent of the values of the x's.

$$\boxed{\boldsymbol{X} \text{ may be fixed or random.}}$$

(6) **Normal distribution:** The disturbances are normally distributed.

$$\boxed{\boldsymbol{\varepsilon}|\boldsymbol{X} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}).}$$

In order to obtain estimates of the parameters β_1, β_2, $\cdots$, β_K, the *residual sum of squares* (RSS)

$$RSS = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \sum_{t=1}^{T} \hat{\varepsilon}_t^2 = \sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{K} x_{it}\beta_i \right)^2$$

is minimised so that the coefficient estimates will be given by the *ordinary least squares (OLS) estimator*

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \cdots \\ \hat{\beta}_k \end{bmatrix} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{2.1}$$

In order to calculate the standard errors of the coefficient estimates, the variance of the errors, σ^2, is estimated by the estimator

$$s^2 = \frac{RSS}{T-K} = \frac{\sum_{t=1}^{T} \hat{\varepsilon}_t^2}{T-K} \tag{2.2}$$

where we recall K is the number of regressors including a constant. In this case, K observations are "lost" as K parameters are estimated, leaving $T-K$ degrees of freedom.

Then the parameter variance-covariance matrix is given by

$$\text{Var}(\hat{\boldsymbol{\beta}}) = s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{2.3}$$

And the coefficient standard errors are simply given by taking the square roots of each of the terms on the leading diagonal. In summary, we have (Brooks [4, page 91-92])

$$\begin{cases} \hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon} \\ s^2 = \frac{\sum_{t=1}^{T} \hat{\varepsilon}_t^2}{T-K} \\ \text{Var}(\hat{\boldsymbol{\beta}}) = s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}. \end{cases} \tag{2.4}$$

The OLS estimator is the best linear unbiased estimator (BLUE), consistent and asymptotically normally distributed (CAN), and if the disturbances are normally distributed, asymptotically efficient among all CAN estimators.

Justifying the least squares principle. In most situations, OLS remains the most popular technique for estimating regressions for the following three reasons:

• Using OLS is easier than the alternatives. Other techniques require more mathematical sophistication and more computing power.

• OLS is sensible. You can avoid positive and negative residuals canceling each other out and find a regression line that's as close as possible to the observed data points.

• OLS results have desirable characteristics. For example, when $K = 2$,

✓The regression line always passes through the sample means of Y and X, or $\overline{Y} = \hat{\beta}_0 + \hat{\beta}_1\overline{X}$ (the point $(\overline{X}, \overline{Y})$ falls on the line $y = \hat{\beta}_0 + \hat{\beta}_1 x$): by the definition of $\hat{\beta}_0$ and $\hat{\beta}_1$.

✓The mean of the estimated (predicted) Y value is equal to the mean value of the actual Y, or $\overline{\hat{Y}} = \overline{\hat{\beta}_0 + \hat{\beta}_1 X} = \hat{\beta}_0 + \hat{\beta}_1\overline{X} = \overline{Y}$.

✓The mean of the residuals is zero, or $\overline{\hat{\varepsilon}} = \overline{Y - (\hat{\beta}_0 + \hat{\beta}_1 X)} = \overline{Y} - (\hat{\beta}_0 + \hat{\beta}_1\overline{X}) = 0$.

✓The residuals are uncorrelated with the predicted Y, or $\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})\hat{\varepsilon}_i = 0$.

✓The residuals are uncorrelated with observed values of the independent variable, or $\sum_{i=1}^{n} \hat{\varepsilon}_i X_i = 0$.

The t-statistic for hypothesis testing is given by

$$\frac{\hat{\beta}_i - \text{hypothesized value}}{SE(\hat{\beta}_i)} \sim t(T-K)$$

where $SE(\hat{\beta}_i) = \sqrt{\text{Var}(\hat{\boldsymbol{\beta}})_{ii}}$, and is used to test single hypotheses. The F-test is used to test more than one coefficient simultaneously.

Under the F-test framework, two regressions are required. The *unrestricted regression* is the one in which the coefficients are freely determined by the data, and the *restricted regression* is the one in which the coefficients are restricted, i.e. the restrictions are imposed on some βs. Thus the F-test approach to hypothesis testing is also termed *restricted least squares*.

The F-test statistic for testing multiple hypotheses about the coefficient estimate is given by

$$\frac{RRSS - URSS}{URSS} \times \frac{T-K}{m} \sim F(m, T-K) \tag{2.5}$$

where $URSS$ is the residual sum of squares from unrestricted regression, $RRSS$ is the residual sum of squares from restricted regression, m is the number of restrictions[1], T is the number of observations, and K is the number of regressors in the unrestricted regression.

To see why the test centres around a comparison of the residual sums of squares from the restricted and unrestricted regressions, recall that OLS estimation involved choosing the model that minimised the residual sum of squares, with no constraints imposed. Now if, after imposing constraints on the model, a residual sum of squares results that is not much higher than the unconstrained model's residual sum of squares, it would be concluded that the restrictions were supported by the data. On the other hand, if the residual sum of squares increased considerably after the restrictions were imposed, it would be concluded that the restrictions were not supported by the data and therefore that the hypothesis should be rejected.

It can be further stated that $RRSS \geq URSS$.[2] Only under a particular set of very extreme circumstances will the residual sums of squares for the restricted and unrestricted models be exactly equal. This would be the case when the restriction was already present in the data, so that it is not really a restriction at all.

Finally, we note any hypothesis that could be tested with a t-test could also have been tested using an F-test, since

$$t^2(T-K) \sim F(1, T-K).$$

3 Examining and Transforming Regression Data

According to Fox [15], we have the following data craft playbook.

The univariate quartet: histograms, density estimation, quantile-comparison plots, boxplots.

• Histograms suffer from several problems: arbitrary origin of the bin system, arbitrary width of the bins, discontinuous, and non-adaptive bin width failing to capture details and avoid noise simultaneously.

• Density estimation: *kernel density estimator* selects the window width for the kernel estimator primarily by trial and error, and *adaptive kernel estimator* adjusts the window width so that the window is narrower where data are plentiful and wider where data are sparse.

• Quantile-comparison plots can reveal skewness and highlight the tails of distributions. They are usually used not to plot a variable directly but for derived quantities, such as residuals from a regression model.

• Boxplots are most useful as adjuncts to other displays (e.g. in the margins of a scatter plot) or for comparing several distributions. When the explanatory variable is discrete, parallel boxplots can be used to display the conditional distributions of Y .

The following rule is used to identify outliers:

• The hinge-spread (or inter-quartile range) is the difference between the hinges:

$$H\text{-spread} = H_U - H_L$$

• The "fences" are located 1.5 hinge-spread beyond the hinges:

$$F_L = H_L - 1.5 \times H\text{-spread}, \; F_U = H_U + 1.5 \times H\text{-spread},$$

[1] Informally, the number of restrictions can be seen as "the number of equality signs under the null hypothesis".

[2] Recall URSS is the shortest distance from a vector to its projection plane.

Observations beyond fences are identified as outliers.

Plotting multivariate data: It is important to understand an essential limitation of the scatterplot matrix as a device for analyzing multivariate data:

• By projecting the multidimensional point cloud onto pairs of axes, the plot focuses on the marginal relationships between the corresponding pairs of variables.

• The object of data analysis for several variables is typically to investigate partial relationships, not marginal associations.

• Y can be related marginally to a particular X even when there is no partial relationship between the two variables controlling for other X's.

• It is also possible for there to be a partial association between Y and an X but no marginal association.

• Furthermore, if the Xs themselves are nonlinearly related, then the marginal relationship between Y and a specific X can be nonlinear even when their partial relationship is linear.

Transformations: The Box-Cox family

$$X \to X^{(p)} = \frac{X^p - 1}{p}$$

The Yeo-Johnson family

$$X^{[p]} = \begin{cases} \frac{(X+1)^p - 1}{p} & X \geq 0 \\ \frac{(-X+1)^{2-p} - 1}{p} & X < 0 \end{cases}$$

Transforming skewness.

• Highly skewed distributions are difficult to examine.

• Apparently outlying values in the direction of the skew are brought in towards the main body of the data.

• Unusual values in the direction opposite to the skew can be hidden prior to transforming the data.

• Statistical methods such as least-squares regression summarize distributions using means. The mean of a skewed distribution is not a good summary of its center.

• Descending the ladder of powers to $\log X$ makes the distribution more symmetric by pulling in the right tail.

• Ascending the ladder of powers (towards X^2 and X^3) can correct a negative skew.

Transforming nonlinearity: Mosteller and Tukey's "bulging rule" for selecting a transformation.

Transforming non-constant spread: Tukey suggested graphing the log hinge-spread against the log median. The slope of the linear trend, if any, in the spread-level plot can be used to suggest a spread-stabilizing power transformation of the data: if $\log \text{spread} \approx a + b \log \text{level}$, then the corresponding spread-stabilizing transformation uses the power $p = 1 - b$.

Transforming proportions: the logit transformation.

3.1 Univariate Displays

Plotting univariate distributions: **histogram**, **boxplot**, **adaptive kernel nonparametric density estimate**, and **normal quantile-comparison plot**.

Adaptive kernel nonparametric density estimate can reveal bimodal pattern, which is not apparent in other plots.

If we fit a line to the QQ plot, then its intercept should estimate μ and its slope should estimate σ. Moreover, systematic nonlinearity in the QQ plot is indicative of departures from the reference distribution.

```
CIA <- read.table("CIA.txt", header=TRUE)
# assumes that CIA.txt is in the current directory, adjust as necessary

library(car)
library(RcmdrMisc)
```

```
par(mfrow=c(2, 2))
Hist(CIA$infant, xlab="Infant Mortality Rate per 1000", ylab="Frequency",
     col="gray", main="(a)")
Boxplot(~infant, data=CIA, main="(b)")
densityPlot(CIA$infant, from=0, normalize=TRUE,
            xlab="Infant Mortality Rate per 1000", main="(c)")
qqPlot(~infant, data=CIA, ylab="Infant Mortality Rate per 1000",
       xlab="Normal Quantiles", main="(d)",
       id=list(method=c(TRUE, rep(FALSE, 132), TRUE)), col.lines="black")
```

(a) histogram, (b) boxplot, (c) adaptive kernel nonparametric density estimate, and (d) normal quantile-comparison plot.

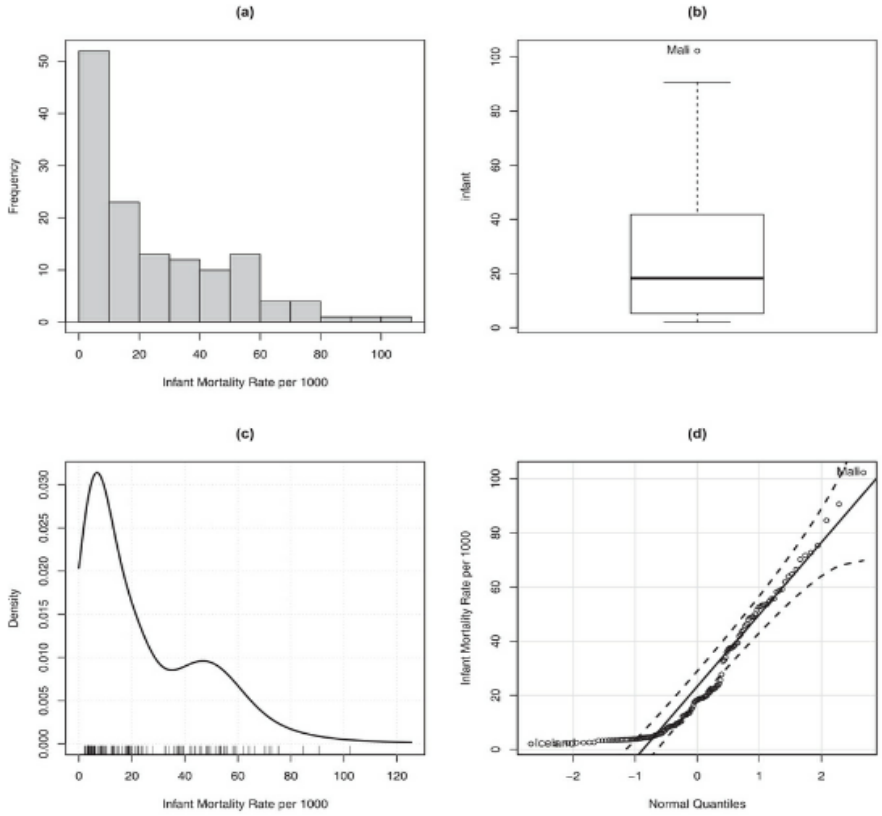


Figure 1: Univariate "quartet"

3.1.1 RcmdrMisc::Hist: Plot a Histogram

Description: This function is a wrapper for the `hist` function in the `base` package, permitting percentage scaling of the vertical axis in addition to frequency and density scaling.

3.1.2 car::Boxplot: Boxplots with Point Identification

Description: `Boxplot` is a wrapper for the standard `R boxplot` function, providing point identification, axis labels, and a formula interface for boxplots without a grouping variable.

3.1.3 car::densityPlot: Nonparametric Density Estimates

Description: `densityPlot` contructs and graphs nonparametric density estimates, possibly conditioned on a factor, using the standard `R density` function or by default `adaptiveKernel`, which computes an adaptive kernel density estimate. `depan` provides the Epanechnikov kernel and `dbiwt` provides the biweight kernel.

3.1.4 car::qqPlot: Quantile-Comparison Plot

Description: Plots empirical quantiles of a variable, or of studentized residuals from a linear model, against theoretical quantiles of a comparison distribution. Includes options not available in the `qqnorm` function.

Details: Draws theoretical quantile-comparison plots for variables and for studentized residuals from a linear model. A comparison line is drawn on the plot either through the quartiles of the two distributions, or by robust regression.

Any distribution for which quantile and density functions exist in R (with prefixes q and d, respectively) may be used. When plotting a vector, the confidence envelope is based on the SEs of the order statistics of an independent random sample from the comparison distribution (see Fox, 2016). Studentized residuals from linear models are plotted against the appropriate t-distribution with a point-wise confidence envelope computed by default by a parametric bootstrap, as described by Atkinson [1]. The function `qqp` is an abbreviation for `qqPlot`.

3.2 Transformations for Symmetry

The *family of power transformations*, $x' = x^\lambda$ $(x > 0)$ is often effective in making the distribution of a numeric variable x more nearly normal, or at least more symmetric. For $\lambda = 0$, we use the $\log_{10}$ transformation: $x' = \ln(x)$, as if $\log_{10}$ were the 0th power. A modified family of power transformations is called the **Box-Cox transformation**:

$$x' = t_{BC}(x, \lambda) = \begin{cases} x^{(\lambda)} = \frac{x^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log_{10}(x) & \lambda = 0. \end{cases}$$

Tukey [26] characterized the family of power transformations as the *ladder of powers and roots*: Descending the ladder of powers and roots from $\lambda = 1$ (no transformation) toward $\lambda = 1/2$ (square root), $\lambda = 0$ ($\log_{10}$), and $\lambda = -1$ (inverse) increasingly spreads out the small values of x relative to the large values. Conversely, ascending the ladder of powers toward $\lambda = 2$ (square) and $\lambda = 3$ (cube) spreads out the large values relative to the small values. This observation explains why power transformations can make skewed data more symmetric.

Whenever the ratio of the largest to the smallest values of a strictly positive variable exceeds 10 (an order of magnitude), it's worth considering the log transformation, which often works remarkably well.

```
symbox(~infant, data=CIA, xlab=expression("Powers,"~lambda), ylab="",
       powers = c(-1, -0.5, 0, 0.33, 0.5, 1))
mtext(2, 1, text=expression(t[BC]("Infant Mortality",~lambda)))
```

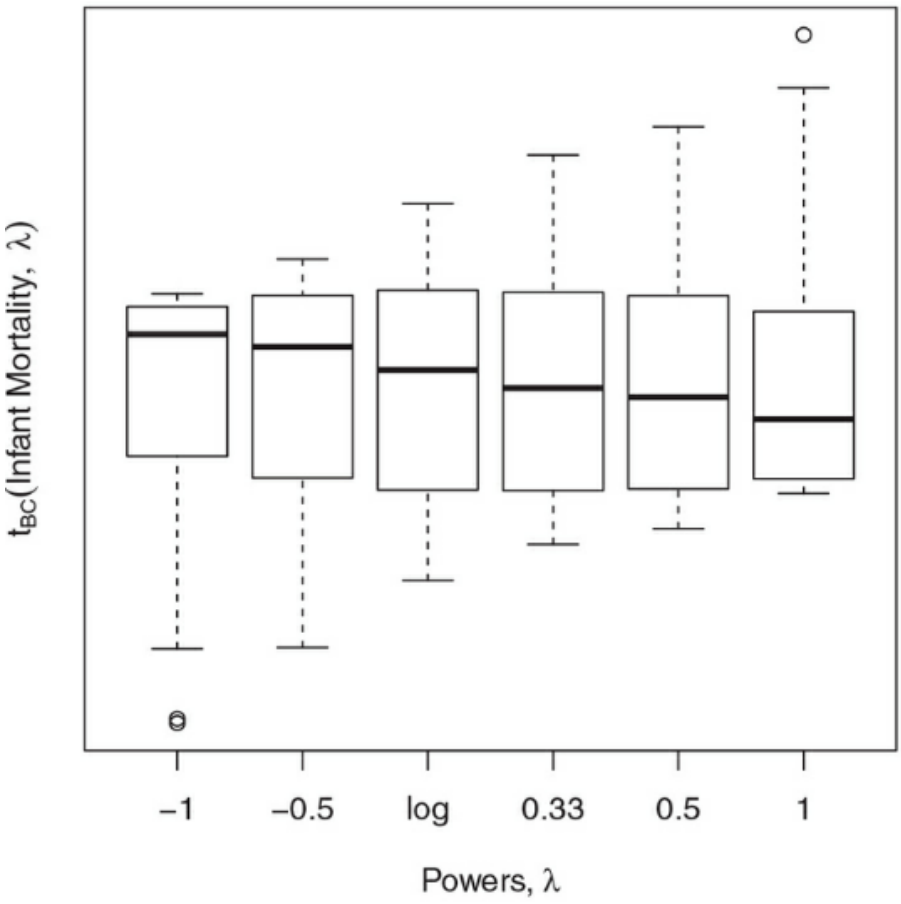


Figure 2: Box plots for various transformations

3.2.1 car::symbox: Boxplots for transformations to symmetry

Description: `symbox` first transforms x to each of a series of selected powers, with each transformation standardized to mean 0 and standard deviation 1. The results are then displayed side-by-side in boxplots, permiting a visual assessment of which power makes the distribution reasonably symmetric.

3.3 *Selecting Transformations Analytically

When Box and Cox [2] introduced their family of modified power transformations, they did so in the context of the regression model

$$t_{BC}(y_i, \lambda) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

where the transformation $t_{BC}(y_i, \lambda)$ of the response is selected to make the errors ε as nearly normally distributed as possible, and where the transformation parameter λ is formally estimated along with the regression coefficients, β_0, β_1, $\cdots$, β_k, and the variance σ^2 of the errors. Their estimation method is similar to *maximum likelihood (ML)* estimation.

We introduce the closely related idea of transforming a variable (or set of variables) to make its *unconditional* distribution as close to normal (or multivariate normal) as possible. Suppose that $t_{BC}(x, \lambda) \sim N(\mu, \sigma^2)$ for a suitable choice of λ. To estimate λ, we find the values of λ, μ, and σ^2 that maximize the normal log-likelihood

$$\ln L(\lambda, \mu, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}[t_{BC}(x_i, \lambda) - \mu]^2$$

The standard errors of the ML estimate $\hat{\lambda}$ can be obtained in the usual manner from the second derivative of $\ln L$ with respect to λ, allowing us to construct Wald-based confidence intervals and tests for λ. Similarly, comparing the values of $\ln L$ at $\lambda = \hat{\lambda}$ and $\lambda = 1$ (i.e. no transformation) provides a *likelihood-ratio test* of the null hypothesis $H_0 : \lambda = 1$.

This approach extends to transforming several variables $x = (x_1, \cdots, x_p)^T$ toward multivariate normality, with

$$\log L(\lambda, \mu, \Sigma) = -\frac{n}{2}\log(2\pi) - \log(\sqrt{\det \Sigma}) - \frac{1}{2}\sum_{i=1}^{n}[t_{BC}(x_i, \lambda) - \mu]^T \Sigma^{-1}[t_{BC}(x_i, \lambda) - \mu]$$

where $\lambda = (\lambda_1, \cdots, \lambda_p)^T$ is the vector of transformation parameters, $\mu = (\mu_1, \cdots, \mu_p)^T$ is the mean vector, and $\Sigma_{p\times p}$ is the variance-covariance matrix of the transformed data,

$$t_{BC}(x, \lambda) = [tBC(x_1, \lambda_1), \cdots, t_{BC}(x_p, \lambda_p)]^T$$

and $\det \Sigma$ is the determinant of Σ.

In the CIA World Factbook data set, transforming the four variables (infant mortality, GDP per capita, the Gini coefficient, and health spending) toward multivariate normality produces the estimated power transformations that make the distribution of the variables more symmetric and the pairwise regressions more nearly linear.

```
# Estimation of transformation of infant mortality

S(pt <- powerTransform(infant ~ 1, data=CIA))
pt$lambda # estimated lambda
sqrt(pt$invHess) # SE

# Fig. 3.6: scatterplot matrix for CIA data

scatterplotMatrix(~infant + gdp + gini + health, data=CIA,
                  var.labels=c("Infant Mortality", "GDP per Capita",
                               "Gini Coefficient", "Health Spending"),
                  smooth=list(smoother=loessLine, var=FALSE, lwd.smooth=3),
                  col="black")

# Fig. 3.7: scatterplot matrix for transformed CIA data

scatterplotMatrix(~log(infant) + basicPower(gdp, 0.2) + log(gini) +
                      log(health), data=CIA,
                  var.labels=c(expression(log("Infant Mortality")),
                               expression("GDP per Capita"^0.2),
                               expression(log("Gini Coefficient")),
                               expression(log("Health Spending"))),
                  smooth=list(smoother=loessLine, var=FALSE, lwd.smooth=3),
                  col="black")

# Table 3.2: estimates of transformation parameters

S(pt4 <- powerTransform(cbind(infant, gdp, gini, health) ~ 1, data=CIA))
```

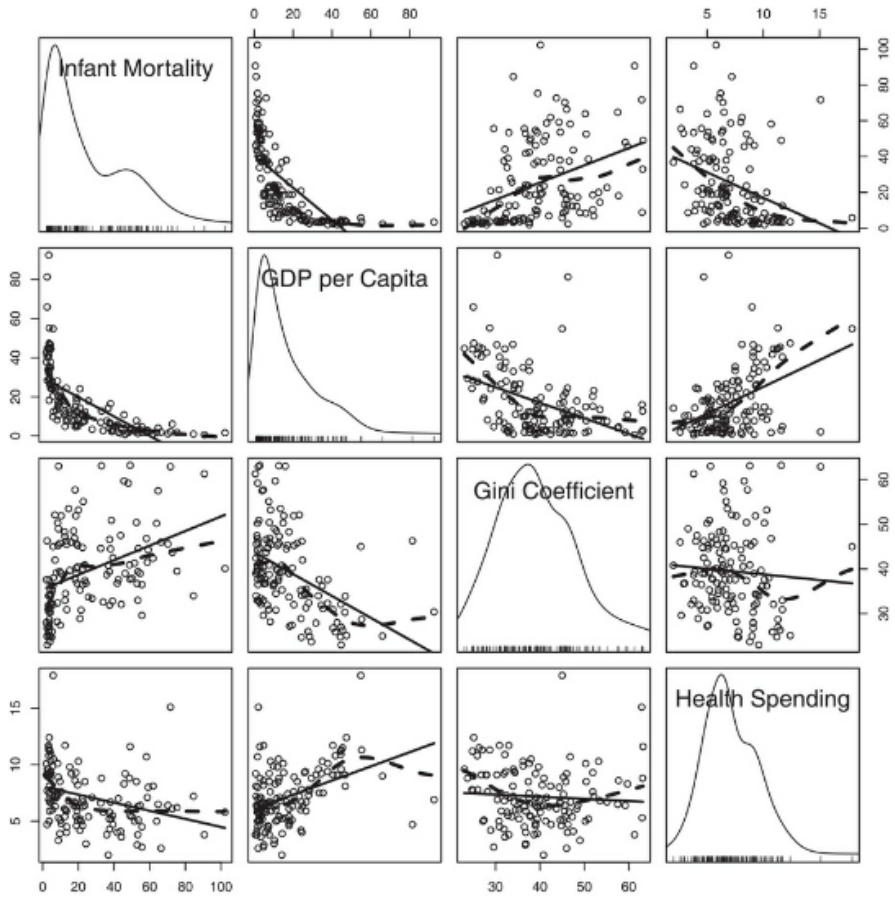


Figure 3: Scatterplot matrix for the CIA World Factbook data: untransformed

Table 3.2 *CIA World Factbook*

Variable (x_j)	$\widehat{\lambda}_j$	*95% Confidence Interval for* λ_j
Infant mortality	−0.0084	(−0.1449, 0.1282)
GDP per capita	0.2139	(0.1074, 0.3204)
Gini coefficient	−0.2399	(−0.8724, 0.3926)
Health spending	0.3057	(−0.0411, 0.6526)

Figure 4: Table of transformation parameters

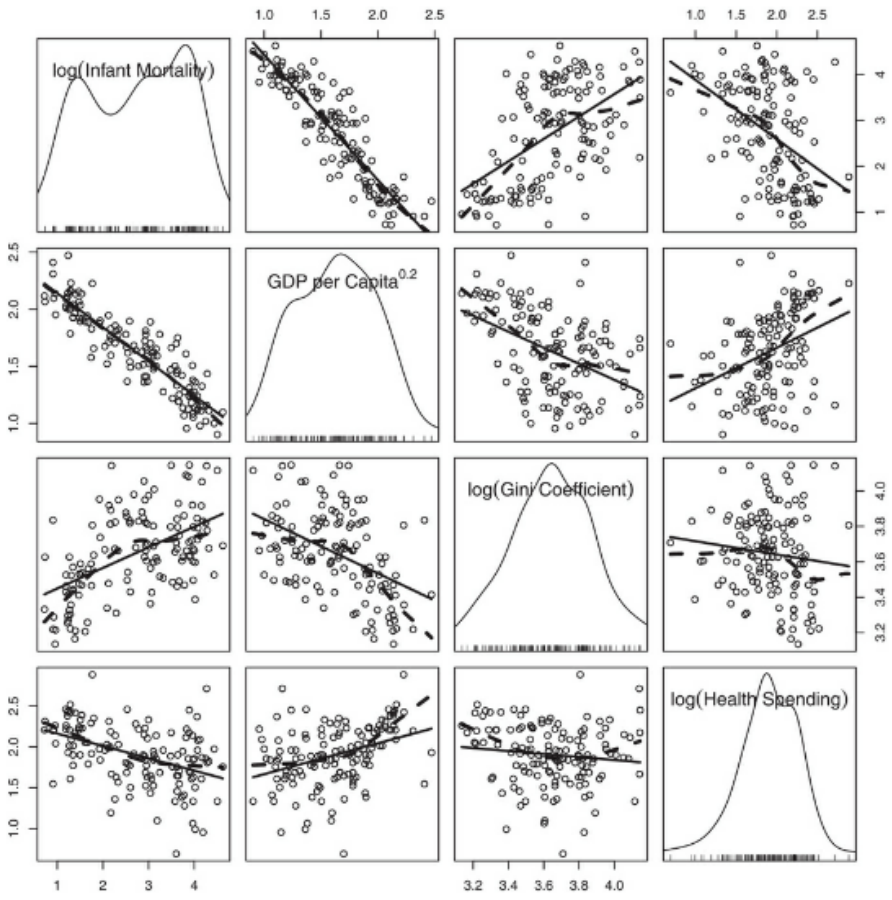


Figure 5: Scatterplot matrix for the CIA World Factbook data: transformed

3.3.1 car::powerTransform: Finding univariate or multivariate power transformations

Description: `powerTransform` uses the maximum likelihood-like approach of Box and Cox [2] to select a transformatiion of a univariate or multivariate response for normality, linearity and/or constant variance. Available families of transformations are the default Box-Cox power family and two additioal families that are modifications of the Box-Cox family that allow for (a few) negative responses. The summary method automatically computes two or three likelihood ratio type tests concerning the transformation powers.

3.3.2 car::S: Modified Functions for Summarizing Linear, Generalized Linear, and Some Other Models

Description: **car** package replacements for the `summary (S)` and `confint (Confint)` functions for `lm`, `glm`, `multinom`, and `polr` objects, with additional arguments but the same defaults as the original functions. The `Confint` method for "`polr`" objects profiles the likelihood to get confidence intervals for the regression parameters but uses Wald intervals for the thresholds. Default methods that call the standard `R summary` and `confint` functions are provided for the `S` and `Confint` generics, so the **car** functions should be safe to use in general. The default method for `Confint` also assumes that there is an appropriate `coef` method. For briefer model summaries, see `brief`.

3.3.3 car::scatterplotMatrix: Scatterplot Matrices

Description: This function provides a convenient interface to the `pairs` function to produce enhanced scatterplot matrices, including univariate displays on the diagonal and a variety of fitted lines, smoothers,

variance functions, and concentration ellipsoids. `spm` is an abbreviation for `scatterplotMatrix`.

3.3.4 car::bcPower, car::basicPower: Box-Cox, Box-Cox with Negatives Allowed, Yeo-Johnson and Basic Power Transformations

Description: Transform the elements of a vector or columns of a matrix using, the Box-Cox, Box-Cox with negatives allowed, Yeo-Johnson, or simple power transformations.

3.4 Transforming Data with Zero and Negative Values

A simple way to apply power transformations to data with 0 or negative values is first to add a positive constant (called a *start* by Tuckey [26]) to the data that is sufficiently large to make all values positive.

A related idea is to modify the Box-Cox power family to accommodate 0 or negative values. For example, Hawkins and Weisberg [22] define the family of modified powers $t_{HW}(x, \lambda, \gamma)$ by applying the Box-Cox transformation $t_{BC}(z, \lambda)$ to the variable $z = \frac{1}{2}(x + \sqrt{x^2 + \gamma^2})$. When $\gamma = 0$, $t_{HW}(x, \lambda, 0) = t_{BC}(x, \lambda)$. Like an ordinary start, γ for the Hawkins–Weisberg family can be selected arbitrarily, or it can be estimated formally along with λ by adapting Box and Cox's method.

Finally, power transformations are effective in modifying the shape of a strictly positive variable x only when the ratio of the largest to the smallest x-values in the data is sufficiently large. If this ratio is small, then power transformations are nearly linear. In this case, we can increase the ratio $x_{\max}/x_{\min}$ by adding a negative start to the data.

3.5 Transformations for Linearity

The linear regression model can accommodate nonlinear relationships through devices such as polynomial regression and regression splines (both of which are discussed in Chapter 6), but in some circumstances, we can render a nonlinear relationship linear by transforming the response variable y, an explanatory variable x, or both. In particular, if the relationship between y and x is monotone (i.e., strictly increasing or strictly decreasing) and simple (in the sense that its direction of curvature doesn't change), then transformation of x or y might be a viable strategy.

Mosteller and Tukey [24] introduced the *bulging rule* to guide trial-and-error selection of power transformations of x, y, or both to linearize simple monotone nonlinear relationships:

- When the bulge points *left*, transform x down the ladder of powers and roots, using a transformation such as $\log(x)$.
- When the bulge points *right*, transform x up the ladder of powers and roots, using a transformation such as x^2.
- When the bulge points *down*, transform y down the ladder of powers and roots, using a transformation such as $\log(y)$.
- When the bulge points *up*, transform y up the ladder of powers and roots, using a transformation such as y^2.

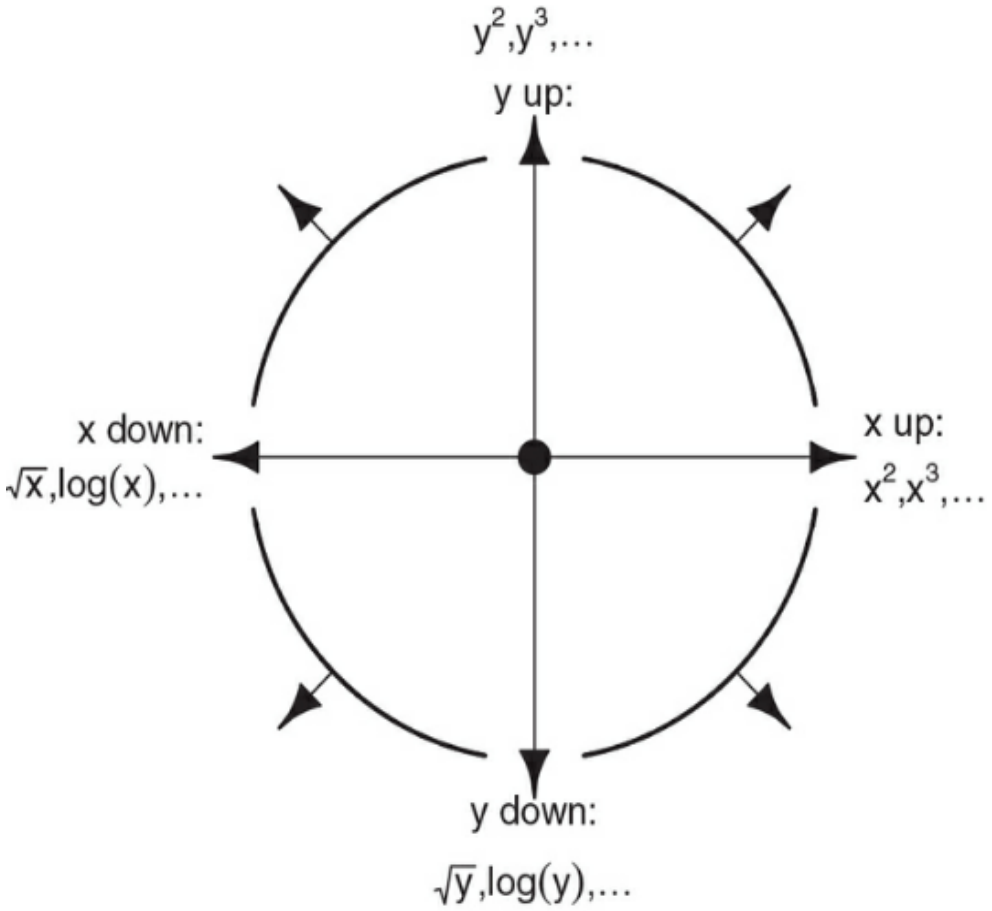


Figure 6: Bulging rule of power transformation

```
scatterplot(infant ~ gdp, data=CIA, smooth=list(smoother=loessLine, var=FALSE,
                                                lwd.smooth=3), col="black",
            regLine=list(lwd=3),
            xlab="GDP per Capita ($1000s)",
            ylab="Infant Mortality Rate (per 1000)")
```

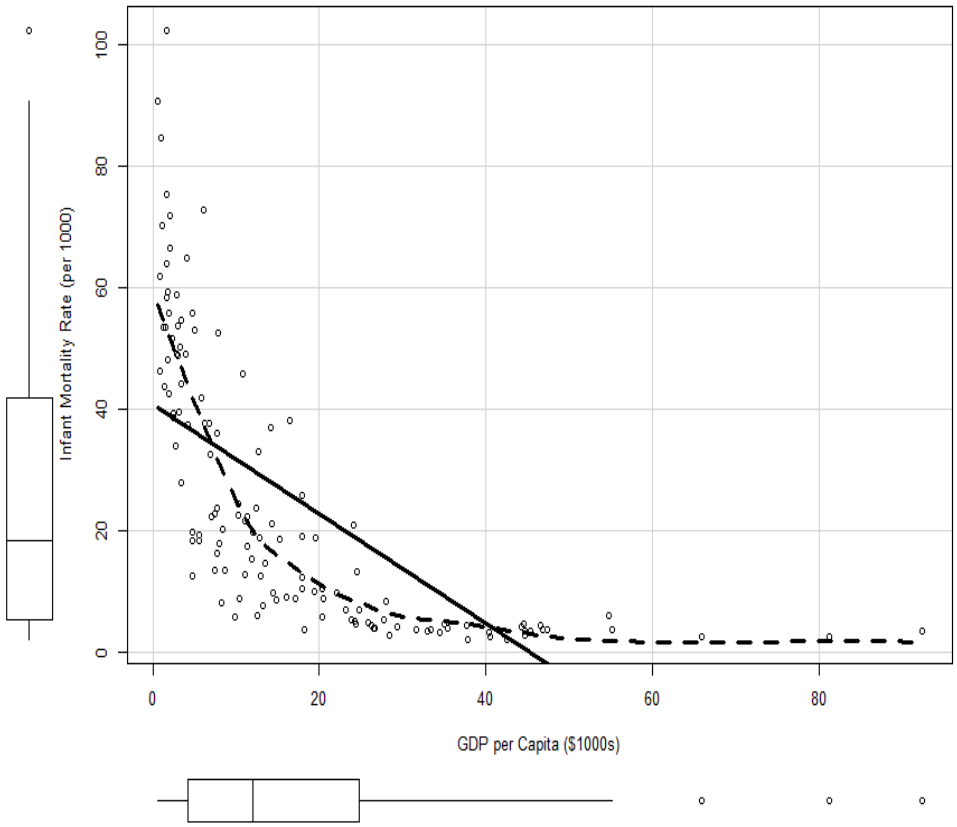


Figure 7: Enhanced scatterplot

3.5.1 car::scatterplot: Enhanced Scatterplots with Marginal Boxplots, Point Marking, ...

Description: This function uses basic R graphics to draw a two-dimensional scatterplot, with options to allow for plot enhancements that are often helpful with regression problems. Enhancements include adding marginal boxplots, estimated mean and variance functions using either parametric or nonparametric methods, point identification, jittering, setting characteristics of points and lines like color, size and symbol, marking points and fitting lines conditional on a grouping variable, and other enhancements. `sp` is an abbreviation for `scatterplot`.

3.6 *How the Loess Smoother Works

3.7 Transforming Nonconstant Variations

A simple setting for examining conditional variation is a categorical explanatory variable that divides the data into groups.

```
par(par)
par(mfrow=c(1, 1))

# Fig. 3.13: Boxplots of GDP vs region

Boxplot(gdp ~ region, data=CIA, id=list(location="lr"),
        ylab="GDP per Capita ($1000s)", xlab="Region of the World")
```

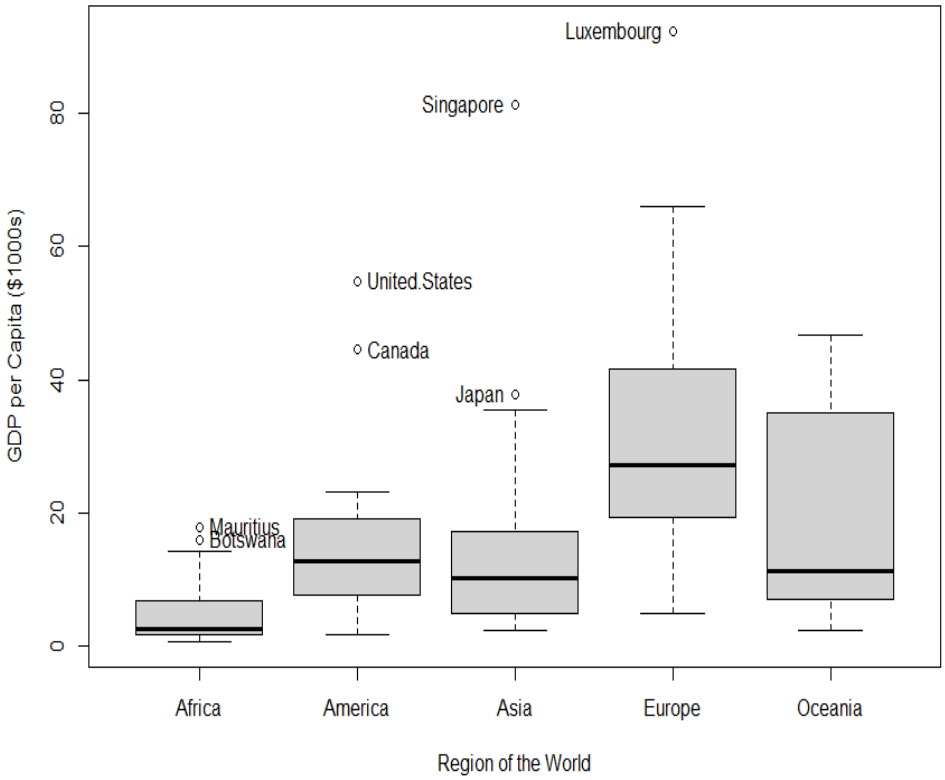


Figure 8: Boxplots by category

Tukey [26] suggested plotting the log of the IQR (a measure of variation or "spread") against the log of the median (a measure of center or "level"), to create a spread-level plot. Tukey explained further that the slope b of the line fit to the spread-level plot can be used for a variation-stabilizing power transformation, selecting the power $\lambda \approx 1 - b$.

```
# reorder levels of region
CIA$region <- factor(CIA$region, levels=c("Europe", "America",
                                          "Oceania", "Asia", "Africa"))

spreadLevelPlot(gdp ~ region, data=CIA, main="",
                ylab="Inter-Quartile Range", col.lines="black")
```

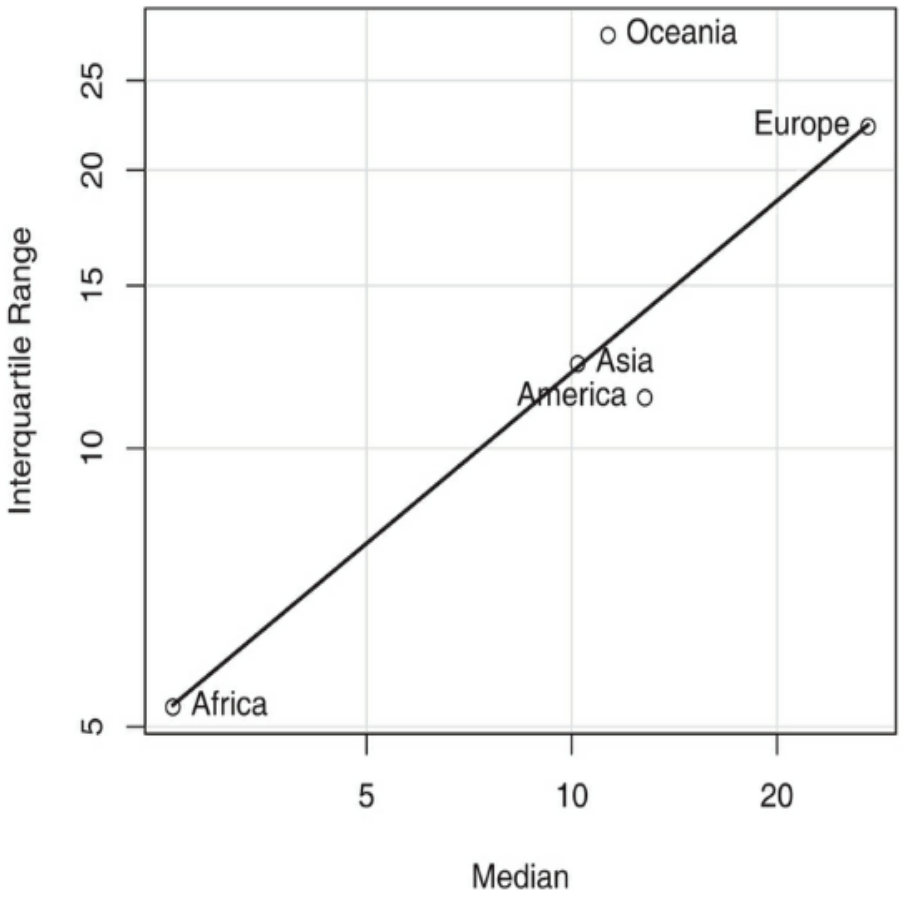


Figure 9: Spread-level plot for transforming nonconstant variations

3.7.1 car::Boxplot: Boxplots with Point Identification

Description: `Boxplot` is a wrapper for the standard `R boxplot` function, providing point identification, axis labels, and a formula interface for boxplots without a grouping variable.

3.7.2 car::spreadLevelPlot: Spread-Level Plots

Description: Creates plots for examining the possible dependence of spread on level, or an extension of these plots to the studentized residuals from linear models.

3.8 Interpreting Results When Variables Are Transformed

Simplicity of interpretation isn't, however, an argument for grossly distorting the data. It is nevertheless true that we usually pay a cost in interpretability when we transform variables that have familiar scales. In most instances, simple interpretations of regression coefficients are unavailable for models with transformed variables or complex interactions. We can always resort to graphical representation of a regression model, however. One framework for doing so, *predictor effect plots*, is described in the context of nonlinearity diagnostics in Chapter 6.

4 Unusual Data: Outliers, Leverage, and Influence

A *regression outlier* is a case whose response value is unusual *given* the value of the explanatory variable. A *univariate outlier* is a value of y or x that is *unconditionally* unusual.

$$\text{influence on coefficients} = \text{leverage} \times \text{outlyingness}$$

4.1 Measuring Leverage: Hat Values

The *hat value* h_i are a common measure of leverage in least squares regression. These values are so named because it is possible to express the fitted values $\hat{Y}_j$ in terms of the observed values Y_i:

$$\hat{Y}_j = \sum_{i=1}^{n} h_{ij} Y_i.$$

Thus, the weight h_{ij} captures the contribution of observation Y_i to the fitted value $\hat{Y}_j$: if h_{ij} is large, then the ith observation can have a substantial impact on the jth fitted value.

Properties of the hat values:

- $h_{ii} = \sum_{j=1}^{n} h_{ij}^2$, and so the hat value $h_i \equiv h_{ii}$ summarizes the potential influence (the leverage) of Y_i on all of the fitted values.
- $1/n \leq h_i \leq 1$.
- The average hat value is $\bar{h} = (k+1)/n$.
- In simple regression analysis, the hat values measure distance from the mean of X:

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{j=1}^{n}(X_j - \bar{X})^2}$$

- In multiple regression, h_i measures distance from the centroid of the X's, taking into account the correlational and variational structure of the X's. Multivariate outliers in the X-space are thus high leverage observations. The response-variable values are not at all involved in determining leverage.

4.1.1 car::leveragePlots: Regression Leverage Plots

Description: These functions display a generalization, due to Sall [25] and Cook and Weisberg [11], of added-variable plots to multiple-df terms in a linear model. When a term has just 1 df, the leverage plot is a rescaled version of the usual added-variable (partial-regression) plot.

Details: The function intended for direct use is `leveragePlots`. The model can contain factors and interactions. A leverage plot can be drawn for each term in the model, including the constant. `leveragePlot.glm` is a dummy function, which generates an error message.

4.2 Detecting Outliers: Studentized Residuals

Even if the errors ε_i have equal variance (as assumed in the general linear model), the residuals e_i do not: $V(e_i) = \sigma_\varepsilon^2(1 - h_i)$. Therefore, high leverage observations tend to have small residuals. Intuitively, this is because these observations can coerce the regression surface to be close to them.

Although we can form a *standardized residual* by calculating $e_i' = e_i/(s\sqrt{1-h_i})$, this measure suffers from the defect that the numerator and the denominator are not independent, preventing e_i' from following a t-distribution: when $|e_i|$ is large, $s = \sqrt{\sum_i e_i^2/(n-k-1)}$, which contains e_i^2, tends to be large as well.

Suppose we refit the regression model after removing the ith case, obtaining a *deleted estimate* $s_{(-i)}$ of σ_ε that is based on the remaining $n-1$ observations. Then the *studentized residual*

$$e_i^* = \frac{e_i}{s_{(-i)}\sqrt{1-h_i}}$$

has independent numerator and denominator and follows a t-distribution with $n-k-2$ degrees of freedom.

An alternative, but equivalent and illuminating, procedure for finding the studentized residuals employs the *mean-shift outlier* model:

$$y_{i'} = \beta_0 + \beta_1 x_{i'1} + \cdots + \beta_k x_{i'k} + \gamma^{(i)} d_{i'}^{(i)} + \varepsilon_{i'}$$

where $d^{(i)}$ is a dummy variable set to 1 for case i and 0 for all other cases. It would be natural to specify the above equation if before examing the data we suspected that case i differed from the others. Then, to test $H_0 : \gamma^{(i)} = 0$, we would find $t_i = \frac{\hat{\gamma}^{(i)}}{se(\hat{\gamma}^{(i)})}$, which is distributed as t_{n-k-2} under H_0 and which (it turns out) is the studentized residual e_i^*.

In most applications we want to look for *any* outliers that may occur in the data: we can in effect refit the mean-shift model n times, producing studentized residuals e_1^*, e_2^*, $\cdots$, e_n^*. Usually, our interest then focuses on the largest absolute e_i^*, denoted as $e_{\max}^*$. Because we have picked the biggest of n test statistics, it is not legitimate simply to use t_{n-k-2} to find a p-value for $e_{\max}^*$. Instead, we perform a *Bonferroni adjustment* to the p-value: let $p' = P(t_{n-k-2} > e_{\max}^*)$. Then the Bonferroni p-value for testing the statistical significance of $e_{\max}^*$ is $p = 2np'$.

4.2.1 car::outlierTest: Bonferroni Outlier Test

Description: Reports the Bonferroni p-values for testing each observation in turn to be a mean-shift outlier, based Studentized residuals in linear (t-tests), generalized linear models (normal tests), and linear mixed models.

Details: For a linear model, p-values reported use the t distribution with degrees of freedom one less than the residual df for the model. For a generalized linear model, p-values are based on the standard-normal distribution. The Bonferroni adjustment multiplies the usual two-sided p-value by the number of observations. The `lm` method works for `glm` objects. To show all of the observations set `cutoff=Inf` and `n.max=Inf`.

4.3 Measuring Influence: Cook's Distance and Other Case Deletion Diagnostics

Cook's statistic D is calculated as

$$D_i = \frac{(e_i')^2}{k+1} \cdot \frac{h_i}{1-h_i}$$

where the first term is a measure of discrepancy (recall $e_i' = e_i/(s_e\sqrt{1-h_i})$ is the *standardized residual*), and the second term is a measure of leverage. We look for values of D_i that are substantially larger than the rest.

Because all of the deletion statistics depend on the hat-values and residuals, a graphical alternative is to plot the studentized residual e_i^* against the hat-value h_i and to look for observations for which both are big.

In developing measures of influence in regression, we have focused on changes in the regression coefficients. Other regression outputs may be examined as well, including the standard errors of the regression coefficients and consequently the size of individual confidence intervals for the coefficients, the size of the joint ellipsoidal confidence region for the regression coefficients, and the degree of collinearity among the x's. All these measures depend on the leverages and the residuals, however. For an extended discussion of various measures of influence, see Chatterjee and Hadi [6], Chapters 4 and 5.

4.3.1 car::influencePlot: Regression Influence Plot

Description: This function creates a "bubble" plot of Studentized residuals versus hat values, with the areas of the circles representing the observations proportional to the value Cook's distance. Vertical reference lines are drawn at twice and three times the average hat value, horizontal reference lines at -2, 0, and 2 on the Studentized-residual scale.

4.3.2 stats::plot.lm: Plot Diagnostics for an `lm` Object

Description: Six plots (selectable by `which`) are currently available: a plot of residuals against fitted values, a Scale-Location plot of `sqrt(|residuals|)` against fitted values, a Normal Q-Q plot, a plot of Cook's distances versus row labels, a plot of residuals against leverages, and a plot of Cook's distances against leverage/(1-leverage). By default, the first three and 5 are provided.

4.4 Numerical Cutoffs for Noteworthy Case Diagnostics

It is generally more effective to examine the distributions of these quantities directly to locate unusual values. Cutoffs for a diagnostic statistic may be derived from statistical theory, or they may result from examination of the sample distribution of the statistic. Large samples have the ability to absorb discrepant data without changing the results substantially, but it is still often of interest to identify relatively influential points, even if no observation has strong absolute influence.

The cutoffs presented below are derived from statistical theory.

Hat-values exceeding about twice the average $\bar{h} = (k+1)/n$ are noteworthy. In small samples, using $2 \times \bar{h}$ tends to nominate too many points for examination, and $3 \times \bar{h}$ can be used instead.

Studentized residuals: under ideal conditions, about 5% of studentized residuals are outside the range $|e_i^*| \leq 2$; it is therefore reasonable to draw attention to observations outside this range.

Standardized change in regression coefficients: $d_{ij}^* = \frac{\hat{\beta}_j - \hat{\beta}_{j(-i)}}{s_{(-i)}(\hat{\beta}_j)} > 2/\sqrt{n}$.

Measures of influence: for Cook's D, the cutoff is $D_i > \frac{4}{n-k-1}$. For DFFITS $= e_i^* \sqrt{\frac{h_i}{1-h_i}}$, the cutoff is $2\sqrt{\frac{k+1}{n-k-1}}$.

4.5 Jointly Influential Cases: Added-Variable Plots

Influential subsets or multiple outliers can often be identified by applying single-observation diagnostics, such as Cook's D and studentized residuals, sequentially.

It can be important to refit the model after deleting each point, because the presence of a single influential value can dramatically affect the fit at other points, and the sequential approach is not always successful.

An attractive alternative is to employ graphical methods, and a particularly useful influence graph (**the "grand champion" of regression graphs**) is the *added-variable plot* (also called a *partial-regression plot* or a *partial-regression leverage plot*).

Let $y_i^{(1)}$ represent the residuals from the least-squares regression of y on all the x's with the exception x_1, that is, the residuals from the fitted model

$$y_i = \hat{\beta}_0^{(1)} + \hat{\beta}_2^{(1)} x_{i2} + \cdots + \hat{\beta}_k^{(1)} x_{ik} + y_i^{(1)},\ i = 1, \cdots, n$$

Likewise, $x_i^{(1)}$ are residuals from the least-squares regression of x_1 on the other x's:

$$x_{1i} = c_0^{(1)} + c_2^{(1)} x_{i2} + \cdots + c_k^{(1)} x_{ik} + x_i^{(1)}$$

The notation emphasizes the interpretation of the residuals $y^{(1)}$ and $x^{(1)}$ as the parts of y and x_1 that remain when the linear dependence of these variables on $x_2, \cdots, x_k$ is removed. The **added-variable plot for** x_1 is the scatter plot of $y^{(1)}$ versus $x^{(1)}$, and it has the following very interesting properties.

- The slope of the least-squares simple-regression line of $y^{(1)}$ on $x^{(1)}$ is the same as the least-squares slope $\hat{\beta}_1$ for x_1 in the full multiple regression.
- The residuals from this simple regression are the same as the residuals e_i from the full regression; that is, $y_i^{(1)} = \hat{\beta}_1 x_i^{(1)} + \varepsilon_i$.
- Consequently, the standard deviation of the residuals in the added-variable plot is s from the multiple regression (if we use the residual degrees of freedom, $n - k - 1$ from the multiple regression to compute s).

• The standard error of $\hat{\beta}_1$ in the multiple regression is then

$$SE(\hat{\beta}_1) = \frac{s}{\sqrt{\sum \left(x_i^{(1)}\right)^2}}$$

• Because the $x_i^{(1)}$ are residuals, they are less variable than the explanatory variable x_1 if x_1 is correlated with the other x's. The added-variable plot therefore shows how collinearity can degrade the precision of estimation by decreasing the conditional variation of an explanatory variable.

The collection of added-variable plots for $x_1, \cdots, x_k$ in effect converts the graph for the multiple regression, the natural scatter plot for which has $k + 1$ dimensions and consequently cannot be drawn when $k > 2$ or 3, into a sequence of two-dimensional scatter plots.

In the context of the current chapter, plotting $y^{(j)}$ against $x^{(j)}$ permits us to examine leverage and influence of the cases on $\hat{\beta}_j$. By the same token, added-variable plots usefully display leverage and influence on the coefficients for all kinds of regressors, including dummy regressors and interaction regressors.

4.5.1 car::avPlots: Added-Variable Plots

Description: These functions construct added-variable, also called partial-regression, plots for linear and generalized linear models.

Details: The function intended for direct use is `avPlots` (for which `avp` is an abbreviation).

4.6 Should Unusual Data Be Discarded?

It is important to investigate why an observation is unusual.

Outliers or influential data may motivate model respecification. 1) For example, the pattern of outlying data may suggest the introduction of additional explanatory variables. 2) In some instances, transformation of the response variable or of an explanatory variable may draw apparent outliers towards the rest of the data, by rendering the error distribution more symmetric or by eliminating nonlinearity.

Except in clear-cut cases, we are justifiably reluctant to delete observations or to respecify the model to accommodate unusual data. Because robust regressions methods assign zero or very small weight to highly discrepant data, the result is generally not very different from careful application of least squares, and, indeed, robust regression weights can be used to identify outliers.

4.7 *Unusual Data: Details

4.7.1 Hat Values and the Hat Matrix

$$\widehat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

Here, $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the *hat* matrix. The hat matrix is symmetric and idempotent. The diagonal entries of the hat matrix $h_i = h_{ii}$ are called *hat values*.

4.7.2 The Distribution of the Least-Squares Residuals

$$\mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}} = (\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon}$$

So

$$E[\mathbf{e}] = 0,\ V(\mathbf{e}) = (\mathbf{I} - \mathbf{H})V(\boldsymbol{\varepsilon})(\mathbf{I} - \mathbf{H})' = \sigma^2(\mathbf{I} - \mathbf{H})$$

The residuals, therefore, are correlated and usually have unequal variances, even though the errors are, by assumption, independent with equal variances.

4.7.3 Case Deletion Diagnostics

4.7.4 The Added-Variable Plot

The first two properties of added-variable plots can be explained intuitively by the geometry of projection.

5 Nonnormality and Nonconstant Error Variance

The assumption that the errors all have 0 means is equivalent to assuming that the functional form of the model is correct, an assumption that we'll address in the next chapter on nonlinearity.

The assumption of normally distributed errors is almost always arbitrary and made for mathematical convenience. Nevertheless, the central limit theorem ensures that under very broad conditions, inference based on the least-squares regression coefficients is approximately valid in all but small samples. **Why, then, should we be concerned about nonnormal errors**?

First, although the validity of inferences for least-squares estimation is robust, p-values for tests and the coverage of confidence intervals are approximately correct in large samples even when the assumption of normality is violated, the method is not robust in efficiency: The least-squares estimator is maximally efficient (has smallest sampling variance) among unbiased estimators when the errors are normal. For some types of error distributions, however, particularly those with heavy tails, the efficiency of least-squares estimation decreases markedly. To a large extent, heavy-tailed error distributions are problematic because they give rise to outliers.

Second, highly skewed error distributions, apart from their propensity to generate outliers in the direction of the skew, compromise the interpretation of the least-squares fit. This fit is, after all, a conditional mean (of y given the x's), and the mean is not a good measure of the center of a highly skewed distribution. Consequently, we may prefer to transform the data to produce a symmetric error distribution.

Finally, a multimodal error distribution suggests the omission from the model of one or more discrete explanatory variables that divide the data naturally into groups. An examination of the distribution of the residuals may therefore motivate elaboration of the regression model.

The consequences of **seriously violating the assumption of constant error variance** are somewhat different: The least-squares coefficients are still unbiased estimators of the population regression coefficients if the assumptions of (1) linearity and (2) independence of the x's and the errors hold, but statistical inference may be compromised, with distorted p-values for hypothesis tests and confidence intervals that don't have the stated coverage. For these negative consequences to occur, nonconstant error variance must typically be severe (see, e.g., Fox [16] Section 12.2.4 and Exercise 12.5).

5.1 Detecting and Correcting Nonnormality

The regression residuals are the key to the behavior of the errors, but, as explained in Chapter 4, the distribution of the residuals e_i isn't as simple as the distribution of the errors ε_i: Even if the errors are independent with equal variances, the residuals are correlated with generally different variances. The studentized residuals e_i^*, though correlated, have equal variances and are distributed as t with degrees of freedom $n-k-2$ if the assumptions of the regression model are correct. One way, therefore, to address the assumption of normality is to compare the distribution of the studentized residuals with t_{n-k-2} in a quantile comparison plot.

The quantile comparison plot is effective in displaying the tail behavior of the residuals. Other univariate graphical displays, such as histograms and density estimates, effectively supplement the quantile comparison plot.

However, where we drew QQ plots for independently sampled data, the studentized residuals from a least-squares fit are correlated. Assessing their sampling variation is therefore more difficult. Atkinson [1] suggested constructing a pointwise confidence envelope by a *parametric bootstrap*.

5.1.1 MASS::studres: Extract Studentized Residuals from a Linear Model

Description: The Studentized residuals. Like standardized residuals, these are normalized to unit variance, but the Studentized version is fitted ignoring the current data point. (They are sometimes called jackknifed residuals).

5.2 *Selecting a Normalizing Transformation Analytically

Atkinson [1] introduced a constructed-variable diagnostic plot based on an approximate score test for the transformation parameter in the Box—Cox regression model.

5.3 Detecting and Dealing With Nonconstant Error Variance

It is common for error variance to increase as the expectation of Y grows larger, or there may be a systematic relationship between error variance and a particular X. The former situation can often be detected by plotting residuals against fitted values; the latter by plotting residuals against each X.

Because the residuals have unequal variances even when the variance of the errors is constant, it is preferable to plot studentized residuals against fitted values. It often helps to plot $|e_i^*|$ or $(e_i^*)^2$ against $\hat{Y}$.

It is also possible to adapt Tukey's spread-level plot (as long as all of the fitted values are positive), graphing log absolute studentized residuals against log fitted values.

There are alternatives to transformation for dealing with non-constant error variance. 1) Weighted-least-squares (WLS) regression, for example, can be used, down-weighting observations that have high variance; 2) it is also possible to correct the estimated standard errors of the ordinary least squares (OLS) estimates for non-constant spread.

Non-constant error variance is a serious problem only when it is relatively extreme.

5.3.1 car::ncvTest: Score Test for Non-Constant Error Variance

Description: Computes a score test of the hypothesis of constant error variance against the alternative that the error variance changes with the level of the response (fitted values), or with a linear combination of predictors.

Details: This test is often called the Breusch-Pagan test [3]; it was independently suggested with some extension by Cook and Weisberg [10]. `ncvTest.glm` is a dummy function to generate an error when a `glm` model is used.

The Breusch-Pagan test. This test assumes that heteroskedasticity may be a linear function of all the independent variables in the model: $\varepsilon_i^2 = \alpha_0 + \alpha_1 X_{i1} + \cdots + \alpha_p X_{ip} + u_i$. The values for ε_i^2 aren't known in practice, so the $\hat{\varepsilon}_i^2$ are calculated from the residuals and used as proxies for ε_i^2. Generally, the BP test is based on the estimation of $\hat{\varepsilon}_i^2 = \alpha_0 + \alpha_1 X_{i1} + \cdots + \alpha_p X_{ip} + u_i$. Alternatively, a BP test can be performed by estimating $\hat{\varepsilon}_i^2 = \delta_0 + \delta_1 \hat{Y}_i$, where $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip}$. Here's how to perform a BP test:

1. Estimate your model, $Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i$, using OLS.
2. Obtain the predicted Y values ($\hat{Y}_i$) after estimating the model.
3. Estimate the auxiliary regression, $\hat{\varepsilon}_i^2 = \delta_0 + \delta_1 \hat{Y}_i$, using OLS.
4. Retain the R-squared value $R^2_{\hat{\varepsilon}^2}$, from this auxiliary regression.
5. Calculate the F-statistic, $F = \frac{\frac{R^2_{\hat{\varepsilon}^2}}{1}}{\frac{(1-R^2_{\hat{\varepsilon}^2})}{n-2}}$, or the chi-squared statistic, $\chi^2 = nR^2_{\hat{\varepsilon}^2}$. If either of these test statistics is significant, then you have evidence of heteroskedasticity.

5.3.2 car::spreadLevelPlot: Spread-Level Plots

Description: Creates plots for examining the possible dependence of spread on level, or an extension of these plots to the studentized residuals from linear models.

Details: Except for linear models, computes the statistics for, and plots, a Tukey spread-level plot of log(hinge-spread) vs. log(median) for the groups; fits a line to the plot; and calculates a spread-stabilizing transformation from the slope of the line. For linear models, plots log(abs(studentized residuals) vs. log(fitted values). Point labeling was added in November, 2016. The function `slp` is an abbreviation for `spreadLevelPlot`.

5.4 Testing for Nonconstant Error Variance

See Zeng [28]: The **Goldfeld-Quant test**, the **White's general test**, the **Breusch-Pagan test**, and the **Park test**.

5.5 Robust Coefficient Standard Errors

WLS estimation. GLM. Quantile regression. Huber-White standard errors. See Zeng [28] for some of the details.

5.6 Bootstrapping

See Efron and Tibshirani [14] and Davison and Hinkley [13] for extensive treatments of bootstrapping, or Fox [16], Chapter 21, for a briefer exposition.

5.7 Weighted Least Squares

See Zeng [28] for an introduction.

5.8 *Robust Standard Errors and Weighted Least Squares: Details

6 Nonlinearity

Nonlinearity (i.e., fitting the wrong equation to the data) is the most fundamental of the problems discussed in this monograph. Component-plus-residual plots are the primary graphical device for diagnosing nonlinearity. As in the case of nonconstant error variance, therefore, it is necessary to focus on particular patterns of departure from linearity.

6.1 Component-Plus-Residual Plots

Although it is useful in multiple regression to plot y against each x (e.g., the row pertaining to y in the scatter plot matrix for y and the x's), these plots do not tell the whole story–and can be misleading–because our interest centers on the partial relationship between y and each x, controlling for the other x's, not on the marginal relationship between y and a single x. Residual-based plots are consequently more relevant in this context.

Plotting residuals or studentized residuals against each x, perhaps augmented by a loess smooth, is helpful in detecting departures from linearity. However, residual plots cannot distinguish between monotone and nonmonotone nonlinearity. The distinction is important, because monotone nonlinearity frequently can be corrected by simple transformations of the variables. *Component-plus-residual plots*, also called *partial-residual plots*, are often an effective tool for this purpose.

Define the *partial residuals* for the jth regressor as

$$e_i^{(j)} = b_j x_{ij} + e_i$$

Then plot $e^{(j)}$ versus x_j. A loess smooth may help in interpreting the plot.

In the context of multiple regression, we generally prefer to transform an x rather than y, unless we see a common pattern of nonlinearity in the partial relationships of y to several x's: Transforming y changes

the shape of its relationship to all of the x's simultaneously, and also changes the shape of the distribution of the residuals.

An alternative to plotting against a transformed x, or against a partial fit, is to plot partial residuals against the original untransformed variable but to show the partial fit as a curve on the graph.

6.1.1 car::crPlots: Component+Residual (Partial Residual) Plots

Description: These functions construct component+residual plots, also called partial-residual plots, for linear and generalized linear models.

Details: The function intended for direct use is `crPlots`, for which `crp` is an abbreviation. The model cannot contain interactions, but can contain factors. Parallel boxplots of the partial residuals are drawn for the levels of a factor.

6.1.2 car::ceresPlots: Ceres Plots

Description: These functions draw Ceres plots for linear and generalized linear models.

Details: Ceres plots are a generalization of component+residual (partial residual) plots that are less prone to leakage of nonlinearity among the predictors. The function intended for direct use is `ceresPlots`. The model cannot contain interactions, but can contain factors. Factors may be present in the model, but Ceres plots cannot be drawn for them.

6.2 When Are Component-Plus-Residual Plots Accurate?

Cook [7] explored the circumstances under which component-plus-residual plots accurately visualize the unknown partial regression function $f(x_1)$ in the model

$$y_i = \beta_0 + f(x_{i1}) + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

where $E[\varepsilon_i] = 0$. The partial regression function $f(x_1)$ isn't necessarily linear, but the other explanatory variables enter the model linearly. We fit the *working model*

$$y_i = \beta_0' + \beta_1' x_{i1} + \beta_2' x_{i2} + \cdots + \beta_k' x_{ik} + \varepsilon_i'$$

in which x_1 enters the model linearly along with the other x's. The partial residuals for the working model estimate

$$\varepsilon_i^{(1)} = \beta_1' x_{i1} + \varepsilon_i'$$

rather than $f(x_{i1}) + \varepsilon_i$. We hope that any nonlinear part of $f(x_1)$ is captured in the residuals from the working model ε_i'. Cook [7] showed that $\varepsilon_i^{(1)} = f(x_{i1}) + \varepsilon_i$ either if the partial regression function is linear after all or if the other x's are each linearly related to x_1. We can then legitimately smooth the scatter plot of the partial residuals versus x_1 to estimate $f(x_1)$.

The takeaway message is that there's an advantage in having linearly related x's, a goal that's promoted, for example, by transforming the x's toward multivariate normality. In practice, it's only strongly nonlinearly related x's that seriously threaten the validity of component-plus-residuals plots.

6.3 More Robust Component-Plus-Residuals Plots

Although **augmented component-plus-residual plots** and **CERES plots** can produce accurate nonlinearity diagnostics when conventional component-plus-residual plots break down, the latter generally work well,

6.4 Component-Plus-Residual Plots for Interactions

Fox and Weisberg [17] described a framework for component-plus-residual plots that' s general enough to accommodate not only nonlinear terms in a linear model, such as polynomials and transformations of xs, but also interactions of arbitrary complexity.

Even without partial residuals, predictor effect plots are useful for visualizing complex regression models, such as models with transformed explanatory variables, polynomial regressors, regression splines (described later in this chapter), and interactions.

6.5 Marginal Model Plots

Recall that in a regression with a response variable y and k numeric explanatory variables $x_1, \cdots, x_k$, the natural graph of the data is a $(k+1)$-dimensional scatter plot of y against the x's. Having estimated a parametric regression model, we could add the fitted regression surface to the graph and compare it with a multidimensional nonparametric regression fit to determine whether the model adequately represents the conditional mean of y as a function of the x's.

The information in the imagined $(k+1)$-dimensional scatter plot is also contained in the infinite set of two-dimensional scatter plots of y against all linear combinations $a_1x_1 + \cdots + a_kx_k$ of the x's. In each such two-dimensional plot, we can smooth the points in the graph and compare that with a smooth of the fitted values $\hat{y}$ computed from the regression model. This turns out to be a very general result that applies to many kinds of regression models.

For a linear regression model, it is also of interest to check the assumption of constant conditional variance of y given the x's (discussed more generally in Chapter 5), and this can be accomplished by smoothing the two sets of residuals from the conditional mean smoothers of the data and the fitted values in a marginal plot. By extension, marginal plots of the data and the fitted values against any variable should show similar conditional means and variation if the model is correct. Following Cook and Weisberg [12], we normally construct marginal model plots only for numeric explanatory variables and for fitted values.

6.6 Testing for Nonlinearity

A general approach to testing for nonlinearity–or, more expressively, if the model isn't linear in an explanatory variable, *lack of fit*–is to specify another, larger model that can capture a more general partial relationship of the response to an explanatory variable. The two models can then be compared by a likelihood-ratio F-test.

Discrete Data. Because it partitions the data into groups, a discrete X (or combination of X's) facilitates straightforward tests of nonlinearity and non-constant error variance.

1) The incremental F-test for nonlinearity of a discrete explanatory variable, say X_1, compares the following general model

$$Y_i = \alpha + \gamma_1 D_{i1} + \cdots + \gamma_{m-1} D_{i,m-1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

with the model specifying a linear effect of X_1,

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

where $D_1, \cdots, D_{m-1}$ are dummy regressors constructed to represent the m categories of X_1.

2) Levene's test for non-constant error variance. See Fox [15] §5.6.2 for details.

6.7 Modeling Nonlinear Relationships With Regression Splines

We've thus far described three strategies for dealing with nonlinearity: (1) in the case of simple monotone nonlinearity, transforming the response or, more commonly in the context of multiple regression, an explanatory variable to straighten their relationship; (2) in the case of nonmonotone nonlinearity, fitting a low-degree polynomial function of an explanatory variable, such as a quadratic or a cubic; (3) simply treating a discrete numeric explanatory variable as a factor to model any pattern of nonlinearity.

Although these strategies suffice for the examples in this chapter, transformations work only for simple, monotone nonlinearity; low-degree polynomials can only accommodate relationships of certain shapes, and they can be disconcertingly nonlocal; and treating numeric explanatory variables as factors produces complicated models that may overfit the data and can be hard to describe succinctly.

Nonparametric regression methods such as loess are sensitive to local characteristics of the data and can fit smooth relationships of arbitrary shape, but nonparametric regression entails the high overhead of abandoning the linear regression model. As it turns out, we can often produce results very similar to nonparametric regression by using constrained piecewise polynomials called regression splines, which can be included as regressors in a linear model and consequently are fully parametric.

6.8 *Transforming Explanatory Variables Analytically

Box and Tidwell [5] introduced a similar regression model in which power transformations λ_1, λ_2, $\cdots$, λ_k, of the explanatory variables, are estimated as parameters:

$$y_i = \beta_0 + \beta_1 x_{i1}^{\lambda_1} + \beta_2 x_{i2}^{\lambda_2} + \cdots + \beta_k x_{ik}^{\lambda_k} + \varepsilon_i$$

where all the x_{ij}'s are positive. The regression coefficients β_0, β_1, β_2, $\cdots$, β_k are typically estimated *after* and conditional on the transformations, because the β's don't really have meaning until the λ's are selected.

For many other techniques, see Fox [15], §5.7.

7 Collinearity

Collinearity is different from the other problems discussed in this monograph in two related respects: (1) Except in exceptional circumstances (explained below), collinearity is fundamentally a problem with the data rather than with the specification of the regression model. (2) As a consequence, there is usually no satisfactory solution for a true collinearity problem.

A strong, but less than perfect, linear relationship among the x's causes the least-squares regression coefficients to be unstable.

7.1 Collinearity and Variance Inflation

The estimated variance of the least squares regression coefficient β_j is

$$V(\hat{\beta}_j) = \frac{s^2}{(n-1)s_j^2} \frac{1}{1-R_j^2}$$

where R_j^2 is the squared multiple correlation from the regression of x_j on the other x's. The impact of collinearity on the precision of estimation is captured by $\frac{1}{1-R_j^2}$, called the *variance inflation factor* (VIF).

As a rule of thumb, VIFs greater than 10 signal a highly likely multicollinearity problem, and VIFs between 5 and 10 signal a somewhat likely multicollinearity issue. Remember to check also other evidence of multicollinearity (insignificant t-statistics, sensitive or nonsensical coefficient estimates, and nonsensical coefficient signs and values). A high VIF is only an indicator of potential multicollinearity, but it may not result in a large variance for the estimator if the variance of the independent variable is also large.

In general, imprecise regression estimates in social research are more frequently the product of large error variance (i.e., weak relationships between the response and the explanatory variables), relatively small samples, and homogeneous samples (i.e., explanatory variables with little variation) than of serious collinearity.

Because the precision of estimation of β_j is most naturally expressed as the width of the confidence interval for this parameter, and because the width of the confidence interval is proportional to the standard error of $\hat{\beta}_j$, we recommend examining the square root of the VIF in preference to the VIF itself. Table 7.1 reveals that linear relationships among the x's must be very strong before collinearity seriously degrades the precision of estimation:

7.1.1 car::vif: Variance Inflation Factors

Description: Calculates variance-inflation and generalized variance-inflation factors for linear, generalized linear, and other models.

7.2 Visualizing Collinearity

There are two ways to visualize the impact of collinearity on estimation: (1) examining data and confidence ellipses and (2) comparing marginal scatter plots with corresponding added-variable plots. These aren't really diagnostic graphs but rather graphs for further understanding how collinearity affects the precision of estimation.

7.3 Generalized Variance Inflation

The VIF is only a sensible measure for terms in a model that are represented by a single parameter. Examples of multiple-parameter terms are sets of dummy-variable coefficients for a factor with more than two levels, and polynomial or regression-spline coefficients for a numeric explanatory variable.

The correlations among a set of dummy regressors generally depend on which level of the factor is selected as the baseline level, but the fit of the model to the data and the intrinsic meaning of the model don't change with this essentially arbitrary choice of regressors. Similarly, we can generally reduce correlations among polynomial regressors by expressing a numeric explanatory variable x as deviations from its mean.

The *generalized variance inflation factors* (GVIFs) for two coefficients (e.g., for two dummy regressors) is interpretable as the increase in the squared area of the joint-confidence ellipse for the two corresponding parameters relative to the area of this ellipse for otherwise similar data in which the two regressors are unrelated to the other regressors in the model.

Fox and Monette recommend taking the $2p$th root of the GVIF, effectively reducing it to a linear measure of imprecision, and analogous to taking the square root of the VIF in the one-coefficient case.

7.4 Dealing With Collinearity

Because collinearity is fundamentally a problem with the data and not (typically) with the model, there generally isn't a satisfactory solution to the problem. There are, however, several strategies that have been suggested for estimating regression models in the presence of collinearity. None are general solutions for the problem.

- Model re-specification.
- Variable selection ("machine learning").
- Regularization (ridge regression and lasso regression).
- Prior information about β's.

7.5 *Collinearity: Some Details

7.5.1 The Joint Confidence Ellipse and the Data Ellipse

7.5.2 Computing Generalized Variance Inflation

8 Diagnostics for Generalized Linear Models

Many of the unusual-data diagnostics, nonlinearity diagnostics, and collinearity diagnostics introduced in previous chapters for the linear regression model fit by least squares can be straightforwardly extended to GLMs.

8.1 Generalized Linear Models: Review

8.2 Detecting Unusual Data in GLMs

8.3 An Illustration: Mroz's Data on Women's Labor Force Participation

8.4 Nonlinearity Diagnostics for GLMs

8.5 Diagnosing Collinearity in GLMs

8.6 Quasi-Likelihood Estimation of GLMs

8.7 *GLMs: Further Background

8.8 Iteratively Weighted Least Squares

9 Concluding Remarks

• Examine your data carefully *before* specifying a regression model for the data; correct errors in the data and anticipate potential problems.

• If you have a moderately large (or larger) data set in which the individual cases aren't of intrinsic and direct interest, consider randomly dividing the data into two parts, one to be used to explore the data and specify a tentative statistical model for them and the other to be used to validate the model.

• Be especially concerned about unusual data in small data sets, but don't ignore the issue in larger data sets: errors in data and unanticipated interesting characteristics.

• In smaller data sets, a nice overview of unusual data is provided by a Cook's D bubble plot of studentized residuals versus hat values, and added-variable plots are generally of interest even in larger data sets.

• Generally attend to the shape of the conditional distribution of the response before checking the functional specification of the model.

• Plotting (studentized) residuals against fitted values is a useful diagnostic for detecting nonconstant error variance.

• Nonlinearity, in the general sense of lack of fit, should always be a concern. Component-plus-residual plots and their various extensions are the go-to diagnostics for detecting lack of fit.

• Don't be quick to blame collinearity.

9.1 Complementary Reading

Weisberg [27] and Fox [16] on a variety of diagnostics. Cook and Weisberg [9] and Cook [8] on regression diagnostics and associated topics. McCullagh and Nelder [23] on generalized linear models.

A Quick-R: Regression Diagnostics

Tutorial source: https://www.statmethods.net/stats/rdiagnostics.html

```
# Assume that we are fitting a multiple linear regression
# on the MTCARS data
library(car)
fit <- lm(mpg~disp+hp+wt+drat, data=mtcars)

#---------- Outliers

# Assessing Outliers
```

```
outlierTest(fit) # Bonferonni p-value for most extreme obs
qqPlot(fit, main="QQ Plot") #qq plot for studentized resid
leveragePlots(fit) # leverage plots
```

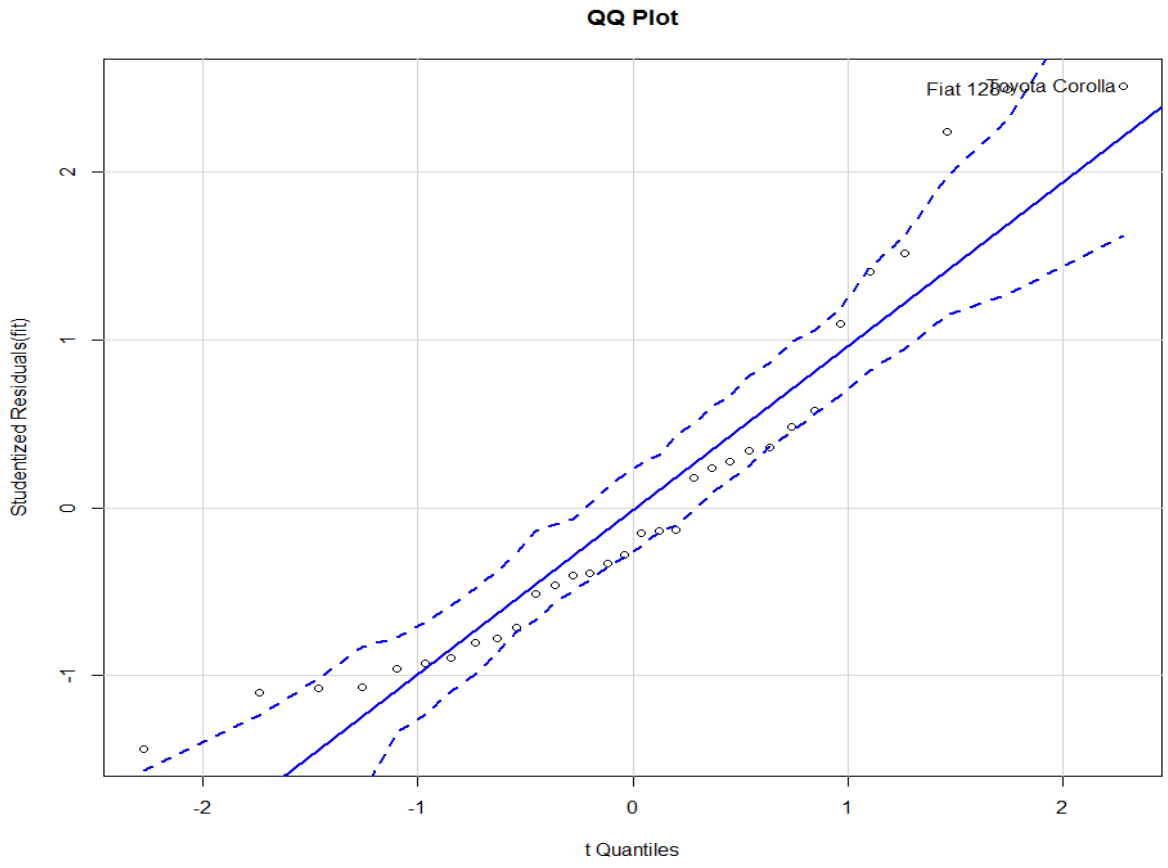


Figure 10: QQ plot for studentized residuals

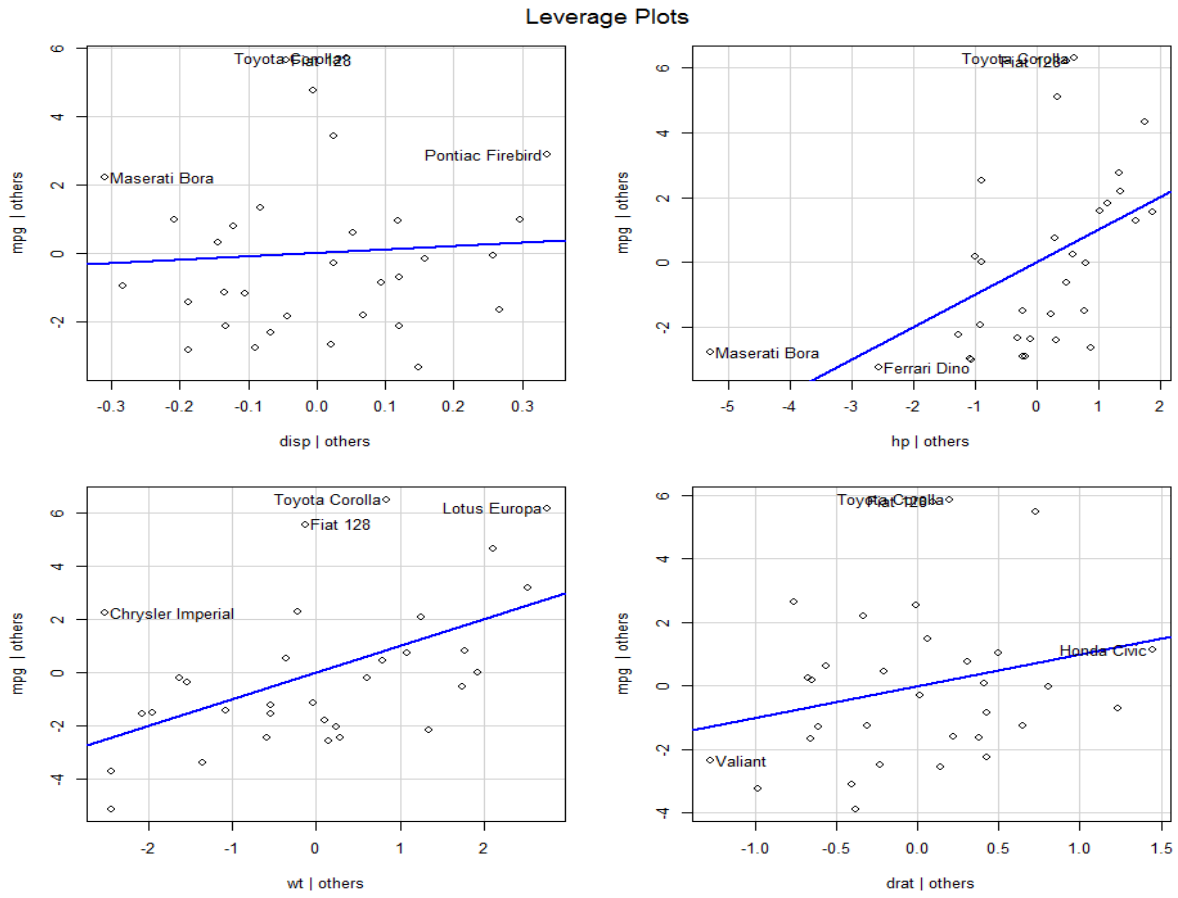
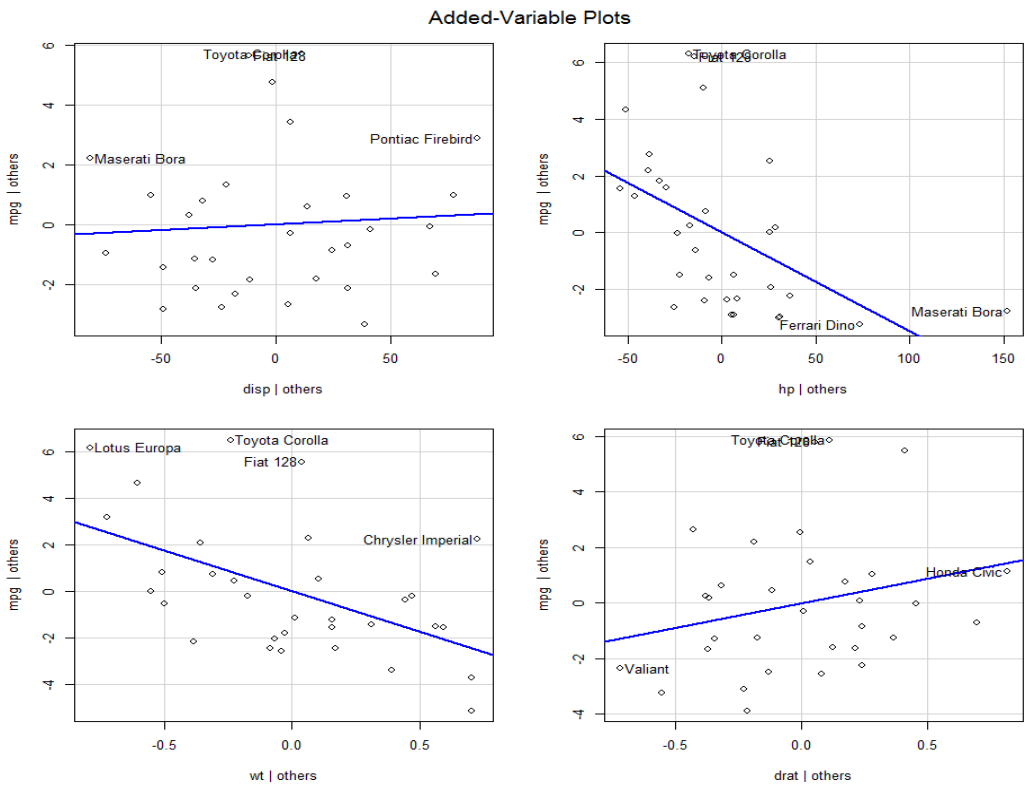
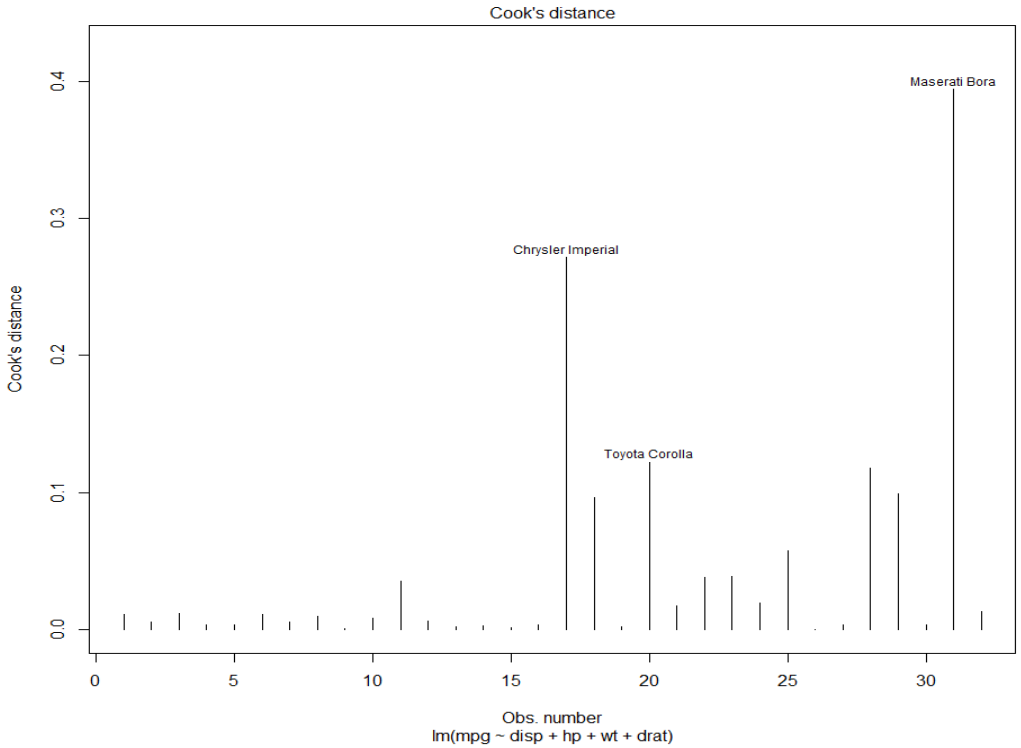


Figure 11: Generalization of added-variable plots as leverage plots

```
#----------- Influential Observations

# Influential Observations
# added variable plots
avPlots(fit)
# Cook's D plot
# identify D values > 4/(n-k-1)
cutoff <- 4/((nrow(mtcars)-length(fit$coefficients)-2))
plot(fit, which=4, cook.levels=cutoff)
# Influence Plot
influencePlot(fit, id.method="identify", main="Influence Plot",
              sub="Circle size is proportial to Cook's Distance" )
```


Added-Variable Plots
mpg | others
disp | others
Toyota Corolla
Fiat 128
Pontiac Firebird
Maserati Bora
hp | others
Ferrari Dino
wt | others
Lotus Europa
Chrysler Imperial
drat | others
Honda Civic
Valiant


Cook's distance
Cook's distance
Maserati Bora
Chrysler Imperial
Toyota Corolla
Obs. number
lm(mpg ~ disp + hp + wt + drat)

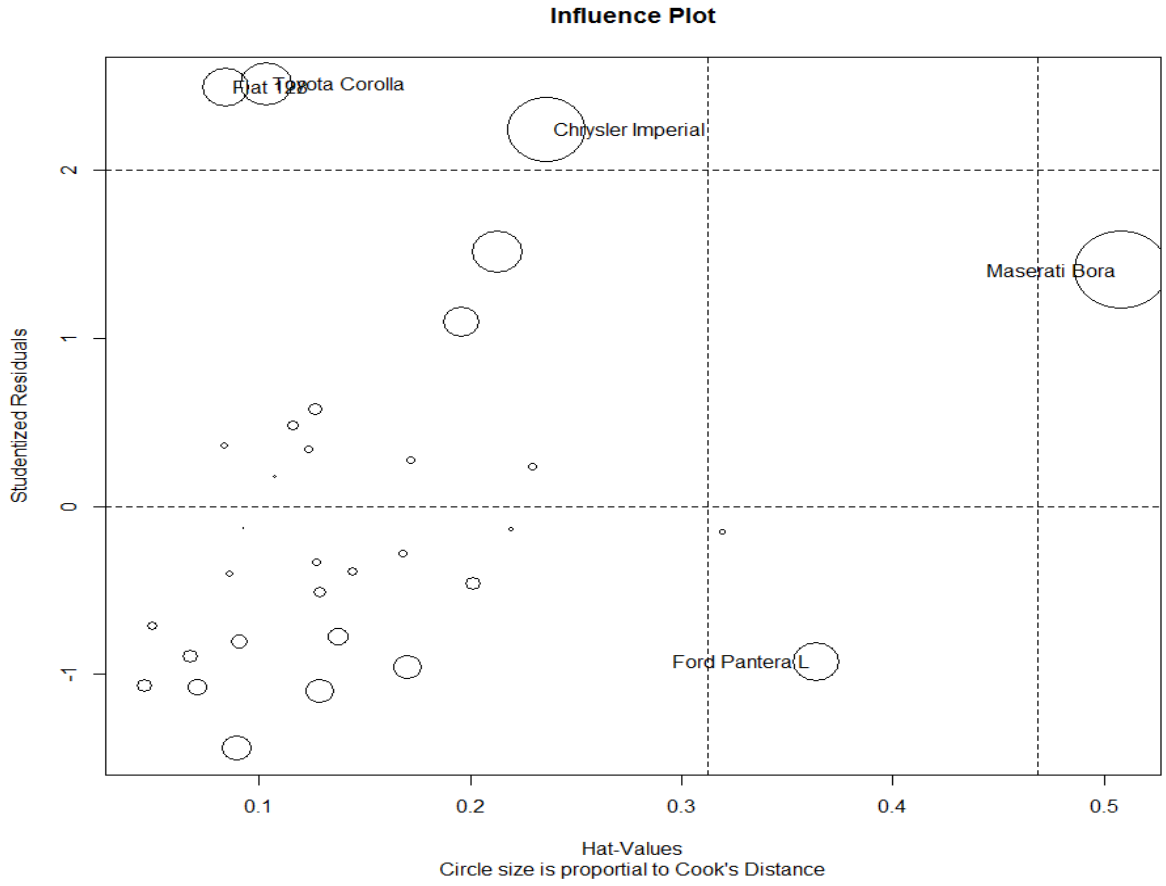


Figure 12: Residuals vs leverages with Cook's distance as bubble size

```
#----------- Non-normality

# Normality of Residuals
# qq plot for studentized resid
qqPlot(fit, main="QQ Plot")
# distribution of studentized residuals
library(MASS)
sresid <- studres(fit)
hist(sresid, freq=FALSE,
   main="Distribution of Studentized Residuals")
xfit<-seq(min(sresid),max(sresid),length=40)
yfit<-dnorm(xfit)
lines(xfit, yfit)
```

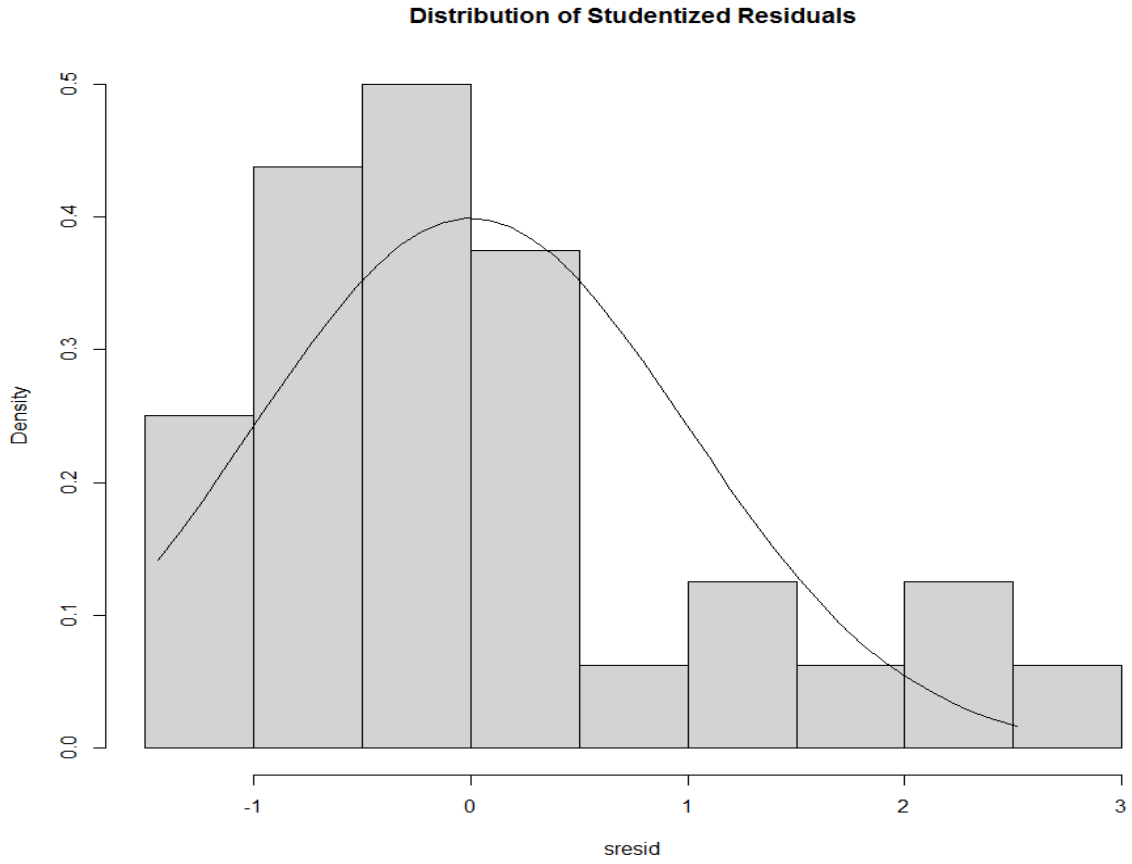


Figure 13: Nonnormality of studentized residuals

```
#----------- Non-constant Error Variance

# Evaluate homoscedasticity
# non-constant error variance test
ncvTest(fit)
# plot studentized residuals vs. fitted values
spreadLevelPlot(fit)
```

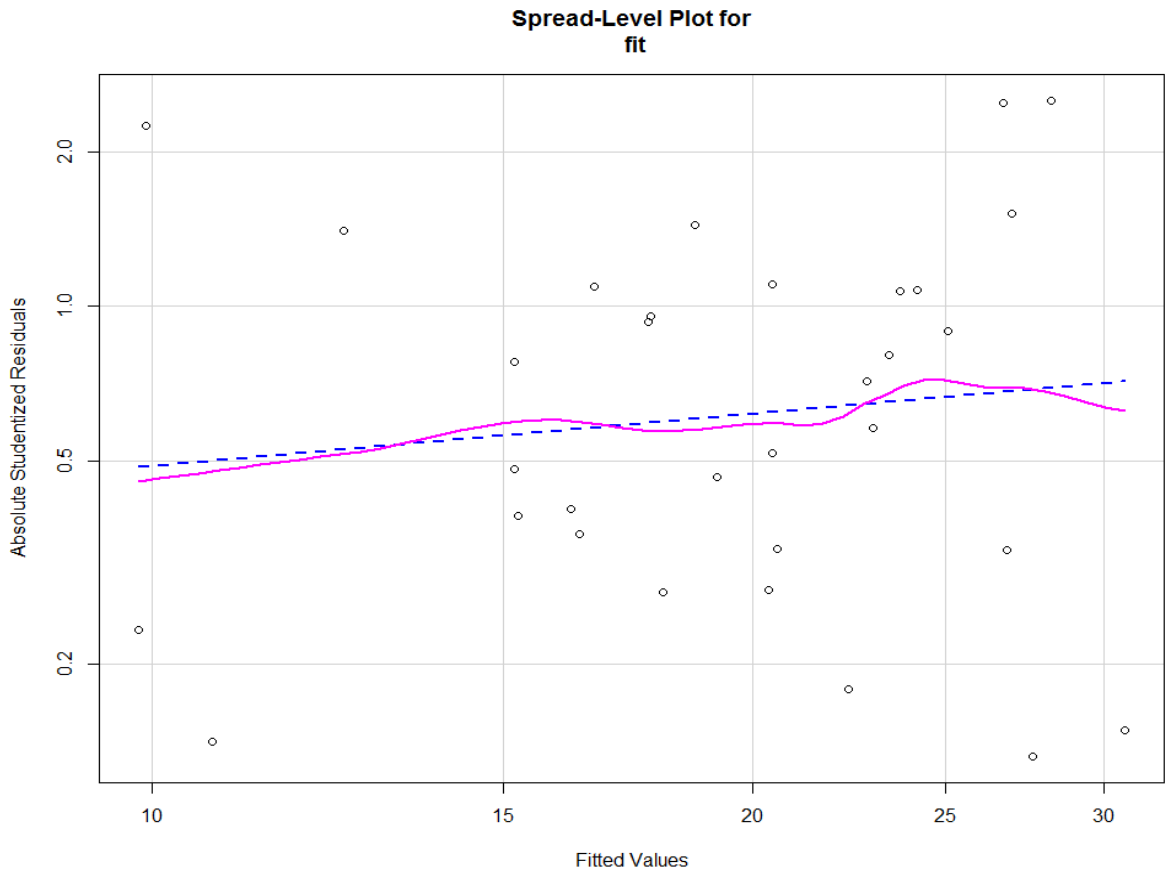


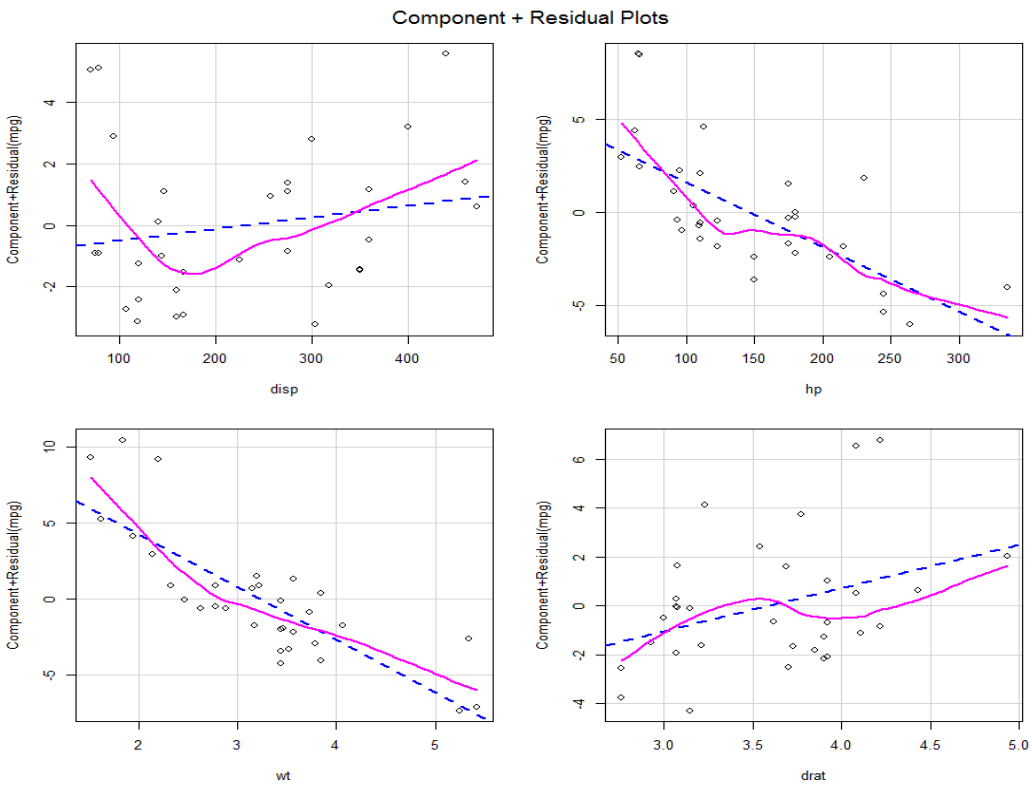
Figure 14: log(abs(studentized residuals)) vs. log(fitted values)

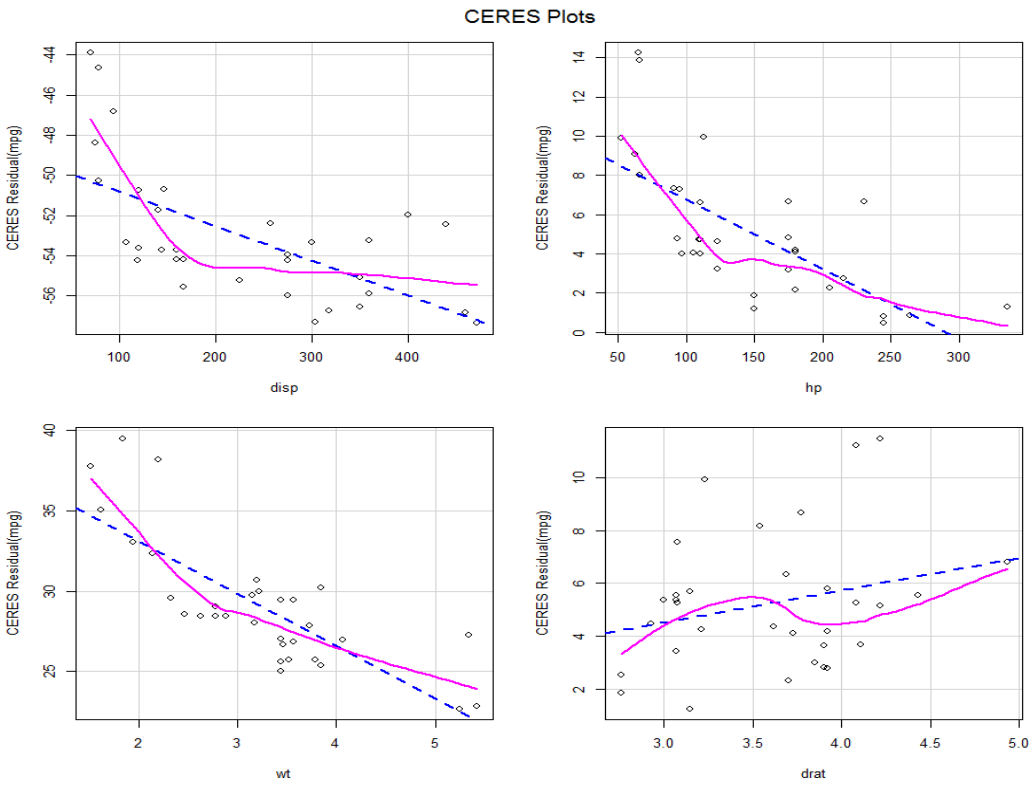
```
#----------- Multi-collinearity

# Evaluate Collinearity
vif(fit) # variance inflation factors
sqrt(vif(fit)) > 2 # problem?

#-----------  Nonlinearity

# Evaluate Nonlinearity
# component + residual plot
crPlots(fit)
# Ceres plots
ceresPlots(fit)
```


Component + Residual Plots
Component+Residual(mpg)
disp
hp
wt
drat


CERES Plots
CERES Residual(mpg)
disp
hp
wt
drat

```
#----------- Non-independence of Errors

# Test for Autocorrelated Errors
durbinWatsonTest(fit)

#----------- Additional Diagnostic Help

# Global test of model assumptions
library(gvlma)
gvmodel <- gvlma(fit)
summary(gvmodel)
```

```
Call:
lm(formula = mpg ~ disp + hp + wt + drat, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5077 -1.9052 -0.5057  0.9821  5.6883

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 29.148738   6.293588   4.631  8.2e-05 ***
disp         0.003815   0.010805   0.353  0.72675
hp          -0.034784   0.011597  -2.999  0.00576 **
wt          -3.479668   1.078371  -3.227  0.00327 **
drat         1.768049   1.319779   1.340  0.19153
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.602 on 27 degrees of freedom
Multiple R-squared:  0.8376,    Adjusted R-squared:  0.8136
F-statistic: 34.82 on 4 and 27 DF,  p-value: 2.704e-10


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = fit)

                     Value  p-value                   Decision
Global Stat        13.93816 0.007495 Assumptions NOT satisfied!
Skewness            4.31310 0.037820 Assumptions NOT satisfied!
Kurtosis            0.01378 0.906542    Assumptions acceptable.
Link Function       8.71658 0.003153 Assumptions NOT satisfied!
Heteroscedasticity  0.89470 0.344207    Assumptions acceptable.
```

Figure 15: Global validation of linear model assumptions

A.1 Global Validation of Linear Model Assumptions

gvlma::gvlma.

Description: Top-level function for Global Validation of Linear Models Assumptions.

Details: `gvlma` is the top-level function to create a `gvlma` object for assessment of linear models assumptions.

B R code for Chapter 3

```
# Regression Diagnostics, 2nd Edition
# John Fox
# last modified: 2019-06-18

# R script for Ch. 3

# Fig. 3.1: univariate displays
```

```
CIA <- read.table("CIA.txt", header=TRUE)
# assumes that CIA.txt is in the current directory, adjust as necessary

library(car)
library(RcmdrMisc)

par(mfrow=c(2, 2))
Hist(CIA$infant, xlab="Infant Mortality Rate per 1000", ylab="Frequency",
     col="gray", main="(a)")
Boxplot(~infant, data=CIA, main="(b)")
densityPlot(CIA$infant, from=0, normalize=TRUE,
            xlab="Infant Mortality Rate per 1000", main="(c)")
qqPlot(~infant, data=CIA, ylab="Infant Mortality Rate per 1000",
       xlab="Normal Quantiles", main="(d)",
       id=list(method=c(TRUE, rep(FALSE, 132), TRUE)), col.lines="black")

# Fig. 3.2: QQ plots

par <- par(mar=c(5.1, 5.1, 4.1, 2.1))

set.seed(123) # for reproducibility
x.norm <- rnorm(50, 100, 15)
par(fig=c(0, .5, .5, 1)) # top-left panel
qqPlot(x.norm, id=FALSE, col.lines="black", xlab="Quantiles of N(0, 1)",
       ylab="Sample Drawn from N(100, 15)", main="(a)")

set.seed(321)
x.t <- rt(50, df=2)
par(fig=c(.5, 1, .5, 1)) # top-right panel
par(new=TRUE)
qqPlot(x.t, id=FALSE, col.lines="black", xlab="Quantiles of N(0, 1)",
       ylab=expression("Sample Drawn from"~~t[2]), main="(b)")

set.seed(666)
x.chisq <- rchisq(50, df=2)
par(fig=c(.25, .75, 0, .5)) # bottom panel
par(new=TRUE)
qqPlot(x.chisq, id=FALSE, col.lines="black", xlab="Quantiles of N(0, 1)",
       ylab=expression("Sample Drawn from"~~chi[2]^2), main="(c)")

par(par) # restore graphical parameters
par(mfrow=c(1, 1))

# Fig. 3.3: Box-Cox family

n <- 500
x <- seq(0.1, 3, length=n)
x1 <- bcPower(x, 1)
x0.5 <- bcPower(x, 0.5)
x0 <- bcPower(x, 0)
xm0.5 <- bcPower(x, -0.5)
xm1 <- bcPower(x, -1)
```

```
x2 <- bcPower(x, 2)
x3 <- bcPower(x, 3)
xlim <- range(c(x1, x0.5, x0, xm0.5, xm1, x2, x3))

plot(range(x)+ c(-0.6, 0.5), c(-5, 10), type="n", xlab="", ylab="", las=1)
usr <- par("usr")
text(usr[2], usr[3] - 1, label="x", xpd=TRUE)
text(usr[1] - 0.2, usr[4] + 0.75, label=expression(t[BC](x, lambda)), xpd=TRUE)
lines(x, x1, lwd=2)
text(x[n]+0.0625, x1[n], labels=expression(lambda == 1), adj=c(0, 0.2))
lines(x, x2, lwd=2)
text(x[n]+0.0625, x2[n], labels=expression(lambda == 2), adj=c(0, 0.2))
lines(x, x3, lwd=2)
text(x[n]+0.0625, x3[n], labels=expression(lambda == 3), adj=c(0, 0.2))
lines(x, x0.5, lwd=2)
text(x[1]-0.025, x0.5[1], labels=expression(lambda == 0.5), adj=c(1, 0.3))
lines(x, x0, lwd=2)
text(x[1]-0.025, x0[1], labels=expression(lambda == 0), adj=c(1, 0.3))
lines(x, xm0.5, lwd=2)
text(x[1]-0.025, xm0.5[1], labels=expression(lambda == -0.5), adj=c(1, 0.3))
lines(x=c(1, 1), y=c(usr[3], 0), lty=2)
lines(x=c(usr[1], 1), y=c(0, 0), lty=2)

# Fig. 3.4: symmetry boxplots

symbox(~infant, data=CIA, xlab=expression("Powers,"~lambda), ylab="",
       powers = c(-1, -0.5, 0, 0.33, 0.5, 1))
mtext(2, 1, text=expression(t[BC]("Infant Mortality",~lambda)))

# Fig. 3.5: log-transformed infant mortality

densityPlot(~log(infant), data=CIA, adjust=0.75, xlab="log(Infant Mortality)")
basicPowerAxis(0, side="above", at=c(1, 5, 10, 20, 50, 100),
               axis.title="Infant Mortality Rate per 1000")

# Estimation of transformation of infant mortality

S(pt <- powerTransform(infant ~ 1, data=CIA))
pt$lambda # estimated lambda
sqrt(pt$invHess) # SE

# Fig. 3.6: scatterplot matrix for CIA data

scatterplotMatrix(~infant + gdp + gini + health, data=CIA,
                  var.labels=c("Infant Mortality", "GDP per Capita",
                               "Gini Coefficient", "Health Spending"),
                  smooth=list(smoother=loessLine, var=FALSE, lwd.smooth=3),
                  col="black")

# Fig. 3.7: scatterplot matrix for transformed CIA data

scatterplotMatrix(~log(infant) + basicPower(gdp, 0.2) + log(gini) +
                      log(health), data=CIA,
```

```
                  var.labels=c(expression(log("Infant Mortality")),
                               expression("GDP per Capita"^0.2),
                               expression(log("Gini Coefficient")),
                               expression(log("Health Spending"))),
                  smooth=list(smoother=loessLine, var=FALSE, lwd.smooth=3),
                  col="black")

# Table 3.2: estimates of transformation parameters

S(pt4 <- powerTransform(cbind(infant, gdp, gini, health) ~ 1, data=CIA))

# Fig. 3.8: patterns of nonlinearity

par <- par(mar=c(2, 3, 3, 2))

par(fig=c(0, .5, .5, 1)) # top-left panel
x <- seq(0, 1, length=200)
Ey <- rev(1 - x^2)
y <- Ey + 0.1*rnorm(200)
plot(x, y, axes=FALSE, frame=TRUE, main="(a) monotone, simple",
     cex.main=1, xlab="", ylab="", col="darkgray", cex=0.75)
lines(x, Ey, lwd=2)
mtext("x", side=1, adj=1)
mtext("y ", side=2, at=max(y), las=1)

par(fig=c(.5, 1, .5, 1)) # top-right panel
par(new=TRUE)
x <- seq(0.02, 0.99, length=200)
Ey <- log(x/(1 - x))
y <- Ey + 0.5*rnorm(200)
plot (x, y, axes=FALSE, frame=TRUE, main="(b) monotone, not simple",
      cex.main=1, xlab="", ylab="", col="darkgray", cex=0.75)
lines(x, Ey, lwd=2)
mtext("x", side=1, adj=1)
mtext("y ", side=2, at=max(y), las=1)

par(fig=c(.25, .75, 0, .5)) # bottom panel
par(new=TRUE)
x <- seq(0.2, 1, length=200)
Ey <- (x - 0.5)^2
y <- Ey + 0.04*rnorm(200)
plot(x, y, axes=FALSE, frame=TRUE, main="(c) non-monotone, simple",
     cex.main=1, xlab="", ylab="", col="darkgray", cex=0.75)
lines(x, Ey, lwd=2)
mtext("x", side=1, adj=1)
mtext("y ", side=2, at=max(y), las=1)

par(par)

# Fig. 3.9: bulging rule

library(plotrix)
library(sfsmisc)
```

```
library(MASS)

par <- par(mar=c(3, 5, 3, 5), mfrow=c(1,1))
eqscplot(c(-1, 1), c(-1, 1), axes=FALSE, ann=FALSE, type="n")
points(0, 0, cex=2, pch=16)

p.arrows(0, 0, 0.9, 0, fill="black")
p.arrows(0, 0, -0.9, 0, fill="black")
p.arrows(0, 0, 0, 0.9, fill="black")
p.arrows(0, 0, 0, -0.9, fill="black")

draw.arc(0, 0, radius=0.8, lwd=2, deg1=10, deg2=80, n=500)
draw.arc(0, 0, radius=0.8, lwd=2, deg1=100, deg2=170, n=500)
draw.arc(0, 0, radius=0.8, lwd=2, deg1=190, deg2=260, n=500)
draw.arc(0, 0, radius=0.8, lwd=2, deg1=280, deg2=350, n=500)

text(0.925, 0.075, labels="x up:", xpd=TRUE, adj=0)
text(0.925, -0.075, labels=expression(paste(x^2, ",", x^3, ",", ldots)),
     xpd=TRUE, adj=0)
text(-0.925, 0.075, labels="x down:", xpd=TRUE, adj=1)
text(-0.925, -0.075, labels=expression(paste(sqrt(x), ",", log(x), ",",
                                             ldots)), xpd=TRUE, adj=1)
text(0, 0.945, labels="y up:", xpd=TRUE, adj=0.5)
text(0, 1.125, labels=expression(paste(y^2, ",", y^3, ",", ldots)),
     xpd=TRUE, adj=0.5)
text(0, -0.945, labels="y down:", xpd=TRUE, adj=0.5)
text(0, -1.125, labels=expression(paste(sqrt(y), ",", log(y), ",", ldots)),
     xpd=TRUE, adj=0.5)

z <- sqrt(0.64/2)
p.arrows(z, z, z + 0.15, z + 0.15, fill="black")
p.arrows(z, -z, z + 0.15, -z - 0.15, fill="black")
p.arrows(-z, z, -z - 0.15, z + 0.15, fill="black")
p.arrows(-z, -z, -z - 0.15, -z - 0.15, fill="black")

par(par)

# Fig. 3.10: scatterplot of infant mortality vs GDP

scatterplot(infant ~ gdp, data=CIA, smooth=list(smoother=loessLine, var=FALSE,
                                                lwd.smooth=3), col="black",
            regLine=list(lwd=3),
            xlab="GDP per Capita ($1000s)",
            ylab="Infant Mortality Rate (per 1000)")

# Fig. 3.11: scatterplot of log(infant mortality) vs log(GDP)

par <- par(oma=c(0, 0, 0, 2))
scatterplot(log(infant) ~ log(gdp), data=CIA,
            smooth=list(smoother=loessLine, var=FALSE, lwd.smooth=3),
            col="black",
            regLine=list(lwd=3),
            xlab="log(GDP per Capita)", ylab="log(Infant Mortality)",
```

```
            reset.par=FALSE)
basicPowerAxis(0, side="right", at=c(1, 5, 10, 20, 50, 100),
               axis.title="Infant Mortality Rate per 1000")
basicPowerAxis(0, side="above", at=c(1, 2, 5, 10, 20, 50),
               axis.title="GDP per Capita ($1000s)")
par(par)

# regression of log(infant mortality) on log(GDP)

S(lm(log(infant) ~ log(gdp), data=CIA))

# Fig. 3.12: how loess works

par <- par(mfrow=c(2, 2), las=1)   # 2 x 2 array of graphs

D <- CIA[, c("gdp", "infant")]
gdp <- D$gdp
infant <- D$infant
ord <- order(gdp)   # sort data by gdp
gdp <- gdp[ord]
infant <- infant[ord]

x0 <- gdp[96]           # focal x = x_(96) (Latvia)
dist <- abs(gdp - x0)   # distance from focal x
h <- sort(dist)[89]     # bandwidth for span of 2/3 (where n = 134)
pick <- dist <= h       # observations within window

plot(gdp, infant, xlab="GDP per Capita ($1000s)",
     ylab="Infant Mortality Rate per 1000",
     type="n", main="(a) Observations Within the Window\nspan = 2/3")
points(gdp[pick], infant[pick], col="black")
points(gdp[!pick], infant[!pick], col="gray")
points(gdp[96], infant[96], pch=16, cex=1.5, col="black") # focal point
abline(v=x0, col="black")    # at focal x
abline(v=c(x0 - h, x0 + h), lty=2, col="black")  # window
text(x0, par("usr")[4] + 5, expression(x[(96)]), xpd=TRUE, col="black")

plot(range(gdp), c(0,1), xlab="GDP per Capita ($1000s)",
     ylab="Tricube Kernel Weight",
     type="n", main="(b) Tricube Weights")
abline(v=x0, col="black")
abline(v=c(x0 - h, x0 + h), lty=2, col="black")

tricube <- function(x, x0, h) {
    z <- abs(x - x0)/h
    ifelse(z < 1, (1 - z^3)^3, 0)
}
tc <- function(x) tricube(x, x0, h) # to use with curve
curve(tc, min(gdp), max(gdp), n=1000, lwd=2, add=TRUE, col="black")
points(gdp[pick], tricube(gdp, x0, h)[pick], pch=16, col="black")
abline(h=c(0, 1), col="gray")

plot(gdp, infant, xlab="GDP per Capita ($1000s)",
```

```
     ylab="Infant Mortality Rate per 1000",
     type="n", main="(c) Local Quadratic Regression")
points(gdp[pick], infant[pick], col="black")
points(gdp[!pick], infant[!pick], col="gray")
abline(v=x0, col="black")
abline(v=c(x0 - h, x0 + h), lty=2, col="black")
mod <- lm(infant ~ poly(gdp, 2), weights=tricube(gdp, x0, h))
new <- data.frame(gdp=seq(x0 - h, x0 + h, length=51))
fit <- predict(mod, newdata=new)
# local regression line
lines(seq(x0 - h, x0 + h, length=51), fit, lwd=2, col="black")
points(x0, fit[26], pch=16, cex=2, col="black")
text(x0, par("usr")[4] + 5, expression(x[(96)]), xpd=TRUE, col="black")
text(x0 + 1, fit[26] + 2.5, expression(hat(y)[(96)]), adj=c(0, 0), col="black")

plot(gdp, infant, xlab="GDP per Capita  ($1000s)",
     ylab="Infant Mortality Rate (per 1000)",
     main="(d) Complete Local-Quadratic Estimate", col=gray(0.25))
yhat <- numeric(length(gdp))
for (i in 1:length(gdp)){   # kernel estimate at each x
    x0 <- gdp[i]
    dist <- abs(gdp - x0)
    h <- sort(dist)[89]
    mod <- update(mod, weights=tricube(gdp, x0, h))
    yhat[i] <- predict(mod, newdata=data.frame(gdp=x0))
}
lines(gdp, yhat, lwd=2, col="black")

par(par)
par(mfrow=c(1, 1))

# Fig. 3.13: Boxplots of GDP vs region

Boxplot(gdp ~ region, data=CIA, id=list(location="lr"),
        ylab="GDP per Capita ($1000s)", xlab="Region of the World")

# Fig. 3.14: spread-level plot for GDP by region

# reorder levels of region
CIA$region <- factor(CIA$region, levels=c("Europe", "America",
                                          "Oceania", "Asia", "Africa"))

spreadLevelPlot(gdp ~ region, data=CIA, main="",
                ylab="Inter-Quartile Range", col.lines="black")

# Fig. 3.15: boxplots for transformed GDP by region

par <- par(mar=c(5.1, 4.1, 3.1, 5.5), mfrow=c(1, 2))

Boxplot(gdp^(1/3) ~ region, data=CIA, id=list(location="lr"),
        ylab=expression(GDP^{1/3}), xlab="Region of the World", main="(a)")
basicPowerAxis(1/3, side="right", at=c(1, 2, 5, 10, 20, 50, 100),
               axis.title="GDP per Capita ($1000s)")
```

```
Boxplot(log(gdp) ~ region, data=CIA, id=list(location="lr"),
        ylab="log(GDP)", xlab="Region of the World", main="(b)")
basicPowerAxis(0, side="right", at=c(1, 2, 5, 10, 20, 50, 100),
               axis.title="GDP per Capita ($1000s)")

par(par)
```

References

[1] Atkinson, A. C. (1985). *Plots, transformations, and regression: An introduction to graphical methods of diagnostic regression analysis*. Oxford, England: Clarendon. 10, 24, 25

[2] Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society Series B (Methodological)*, 26(2), 211–252. 11, 14

[3] Breusch, T. S. and Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47, 1287–1294. 25

[4] Chris Brooks. *Introductory econometrics for finance*, 2ed.. New York, Cambridge University Press, 2008. 5, 6

[5] Box, G. E. P. and Tidwell, P. W. (1962). Transformation of the independent variables. *Technometrics*, 4, 531–550. 29

[6] Chatterjee, S. and Hadi, A. S. (1988). *Sensitivity analysis in linear regression*. New York, NY: Wiley 21

[7] Cook, R. D. (1993). Exploring partial residual plots. *Technometrics*, 35, 351–362. 27

[8] Cook, R. D. (1998). *Regression graphic: Ideas for studying regressions through graphics*. New York, NY:Wiley. 31

[9] Cook, R. D. and Weisberg, S. (1982). *Residuals and influence in regression*. New York, NY:CRC Press. 31

[10] Cook, R. D. and Weisberg, S. (1983). Diagnostics for heteroscedasticity in regression. *Biometrika*, 70, 1–10. 25

[11] Cook, R. D. and Weisberg, S. (1991). Added Variable Plots in Linear Regression. In Stahel, W. and Weisberg, S. (eds.), *Directions in Robust Statistics and Diagnostics*. Springer, 47-60. 20

[12] Cook, R. D. and Weisberg, S. (1997). Graphics for assessing the adequacy of regression models. *Journal of the American Statistical Association*, 92, 490–499. 28

[13] Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge, England: Cambridge University Press. 26

[14] Efron, B. and Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York, NY: Chapman & Hall. 26

[15] John Fox. *Regression Diagnostics*, November 2009, FIOCRUZ Rio de Janeiro - Brasil. 7, 28, 29

[16] Fox, J. (2016). *Applied regression analysis and genearalized linear models*, 3rd ed.. Thousand Oaks, CA:Sage. 24, 26, 31

[17] Fox, J. and Weisberg, S. (2018). Visualizing fit and lack of fit in complex regression models with predictor effect plots and partial residuals. *Journal of Statistical Software*, 87(9), 1–27. 28

[18] Fox, J. & Weisberg, S. (2019). *An R companion to applied regression*, 3rd ed.. Thousand Oaks, CA:Sage.

[19] John Fox. *Regression Diagnostics*, 2ed.. Sage Publications, 2020. 1

[20] William H. Greene. *Econometric analysis*, 7ed.. Prentice Hall, 2012. 5

[21] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York, NY: Springer.

[22] Hawkins, D. M. and Weisberg, S. (2017). Combining the Box−Cox power and generalised log transformations to accommodate negative responses in linear and mixed-effects linear models. *South African Statistics Journal*, 51, 317–328. 15

[23] McCullagh, P. & Nelder, J. A. (1989). *Generalized linear models*, 2nd ed.. London, England:CRC press. 31

[24] Mosteller, F. and Tukey, J. W. (1977). *Data analysis and regression: A second course in statistics*. Reading, MA: Addison-Wesley. 15

[25] Sall, J. (1990). Leverage plots for general linear hypotheses. *American Statistician* 44, 308–315. 20

[26] Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA:Addison-Wesley. 10, 15, 18

[27] Weisberg, S. (2014). *Applied linear regression*, 4th ed.. Hoboken, NJ:Wiley. 31

[28] Yan Zeng. *Classical Linear Rgression Model: Assumptions and Diagnostic Tests*. Unpublished manuscript. 26