Regression Diagnostics

John Fox

McMaster University
Canada

November 2009

Pós-Graduação em Epidemiologia em Saúde Pública/Escola Nacional de Saúde Pública Programa de "APOIO A REALIZAÇÃO DE CURSOS DE CURTA DURAÇÃO COM ABRANGÊNCIA INTERNACIONAL", Vice-Presidência de Ensino, Informação e Comunicação da Fiocruz

1. Introduction

- Linear and generalized linear models make strong assumptions about the structure of data, assumptions that often do not hold in applications.
- Especially in small samples, these models can also be sensitive to unusual data; in extreme cases, the results might be determined by one or a very small number of observations.
- It is therefore important to examine data carefully, both prior to and after fitting a regression model to the data.

2. Outline

- Data craft: examining and transforming variables.
- Unusual data in linear models: outliers, leverage points, and influential observations, and what to do about them.
- Non-normality, non-constant error variance, and nonlinearity in linear models: methods of detection, transformation, and other strategies.
- Diagnostics for unusual data and nonlinearity in generalized linear models.

3. Data Craft

3.1 Goals

- To motivate the inspection and exploration of data as a necessary preliminary to statistical modeling.
- To review (quickly) familiar graphical displays (histograms, boxplots, scatterplots).
- To introduce displays that may not be familiar (nonparametric density estimates, quantile-comparison plots, scatterplots matrices, jittered scatterplots).
- To introduce the 'family' of power transformations.
- To show how power transformations can be used to correct common problems in data analysis, including skewness, nonlinearity, and non-constant spread.

- To introduce the logit transformation for proportions (time permitting).

3.2 A Preliminary Example

- Careful data analysis begins with inspection of the data, and techniques for examining and transforming data find direct application to the analysis of data using linear models.
- The data for the four plots in Figure 1, given in the table below, were cleverly contrived by Anscombe (1973) so that the least-squares regression line and all other common regression 'outputs' are identical in the four datasets.

$X_{a,b,c}$	Y_a	Y_b	Y_c	X_d	Y_d
10	8.04	9.14	7.46	8	6.58
8	6.95	8.14	6.77	8	5.76
13	7.58	8.74	12.74	8	7.71
9	8.81	8.77	7.11	8	8.84
11	8.33	9.26	7.81	8	8.47
14	9.96	8.10	8.84	8	7.04
6	7.24	6.13	6.08	8	5.25
4	4.26	3.10	5.39	19	12.50
12	10.84	9.13	8.15	8	5.56
7	4.82	7.26	6.42	8	7.91
5	5.68	4.74	5.73	8	6.89

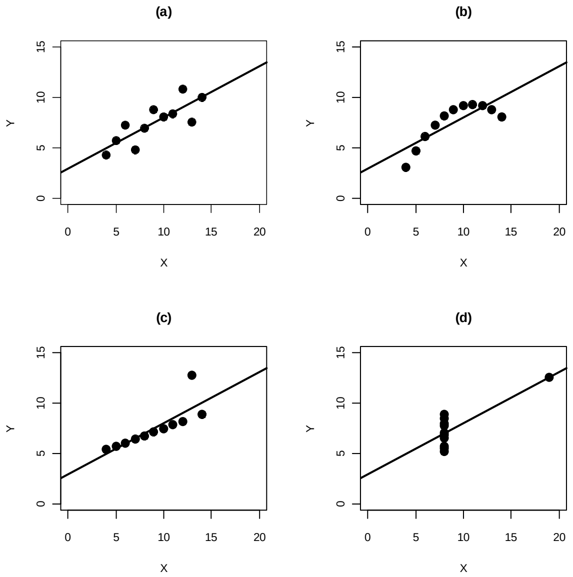

Figure 1. Anscombe's "quartet": Each data set has the same linear least--squares regression of Y on X.

▶ It is clear, however, that each graph tells a different story about the data:

- In (a), the linear regression line is a reasonable descriptive summary of the tendency of Y to increase with X.
- In Figure (b), the linear regression fails to capture the clearly curvilinear relationship between the two variables; we would do much better to fit a quadratic function here, $Y = a + bX + cX^2$.
- In Figure (c), there is a perfect linear relationship between Y and X for all but one outlying data point. The least-squares line is pulled strongly towards the outlier, distorting the relationship between the two variables for the rest of the data. When we encounter an outlier in real data we should look for an explanation.
- Finally, in (d), the values of X are invariant (all are equal to 8), with the exception of one point (which has an X-value of 19); the least-squares line would be undefined but for this point. We are usually uncomfortable having the result of a data analysis depend so centrally on a single influential observation.

- Only in this fourth dataset is the problem immediately apparent from inspecting the numbers.

3.3 Univariate Displays

3.3.1 Histograms

▶ Figure 2 shows two *histograms* for the distribution of infant morality rate per 1000 live births for 193 nations of the world (using 1998 data from the UN).

- The range of infant mortality is dissected into equal-width class intervals (called 'bins'); the number of observations falling into each interval is counted; and these frequency counts are displayed in a bar graph.
- Both histograms use bins of width 10 they differ in that the bins in (a) start at 0, while those in (b) start at -5. The two histograms are more similar than different but they do give slightly different impressions of the shape of the distribution.

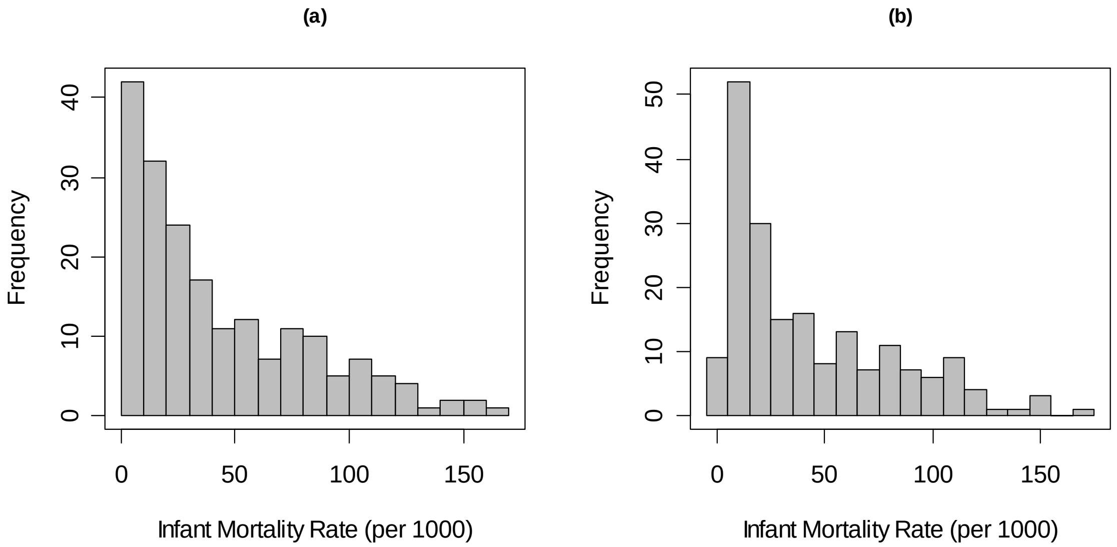

Figure 2. Two histograms with the same bin width but different origins for infant mortality in the United Nations data.

▶ Histograms are very useful graphs, but they suffer from several problems:

- The visual impression of the data conveyed by a histogram can depend upon the arbitrary origin of the bin system.
- Because the bin system dissects the range of the variable into class intervals, the histogram is discontinuous (i.e., rough) even if, as in the case of infant mortality, the variable is continuous.
- The form of the histogram depends upon the arbitrary width of the bins.
- If we use bins that are narrow enough to capture detail where data are plentiful — usually near the center of the distribution — then they may be too narrow to avoid ‘noise’ where data are sparse — usually in the tails of the distribution.

3.3.2 Density Estimation

- ▶ *Nonparametric density estimation* addresses the deficiencies of traditional histograms by averaging and smoothing.
- ▶ The *kernel density estimator* continuously moves a window of fixed width across the data, calculating a locally weighted average of the number of observations falling in the window — a kind of running proportion.
  - • The smoothed plot is scaled so that it encloses an area of one.
  - • Selecting the window width for the kernel estimator is primarily a matter of trial and error — we want a value small enough to reveal detail but large enough to suppress random noise.
  - • The adaptive kernel estimator is similar, except that the window width is adjusted so that the window is narrower where data are plentiful and wider where data are sparse.
- ▶ An example is shown in Figure 3.

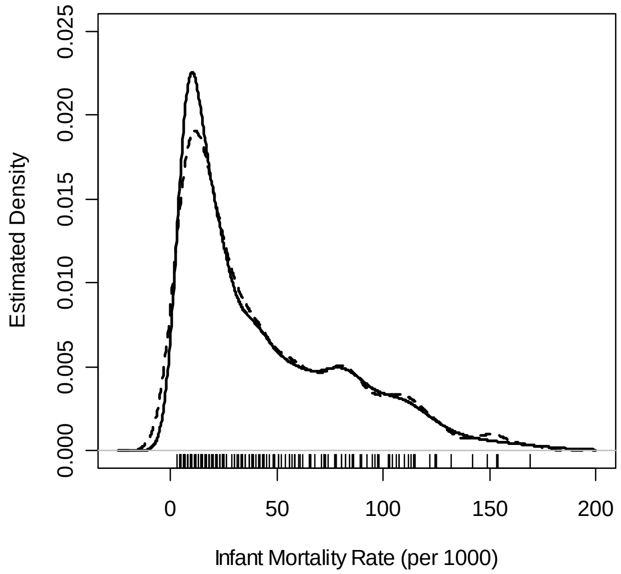

Figure 3. Kernel (broken line) and adaptive-kernel (solid line) density estimators for the distribution infant mortality. A “one-dimensional scatterplot” (or “rug plot”) of the observations is shown at the bottom.

3.3.3 Quantile-Comparison Plots

- ▶ Quantile-comparison plots are useful for comparing an empirical sample distribution with a theoretical distribution, such as the normal distribution. A strength of the display is that it does not require the use of arbitrary bins or windows.
- ▶ Let $P(x)$ represent the theoretical cumulative distribution function (CDF) to which we wish to compare the data; that is, $\Pr(X \leq x) = P(x)$.
  - • A simple (but flawed) procedure is to calculate the empirical cumulative distribution function (ECDF) for the observed data, which is simply the proportion of data below each x:

$$\widehat{P}(x) = \frac{\overset{n}{\underset{i=1}{\#}} (X_i \leq x)}{n}$$

  - • As illustrated in Figure 4, however, the ECDF is a ‘stair-step’ function, while the CDF is typically smooth, making the comparison difficult.

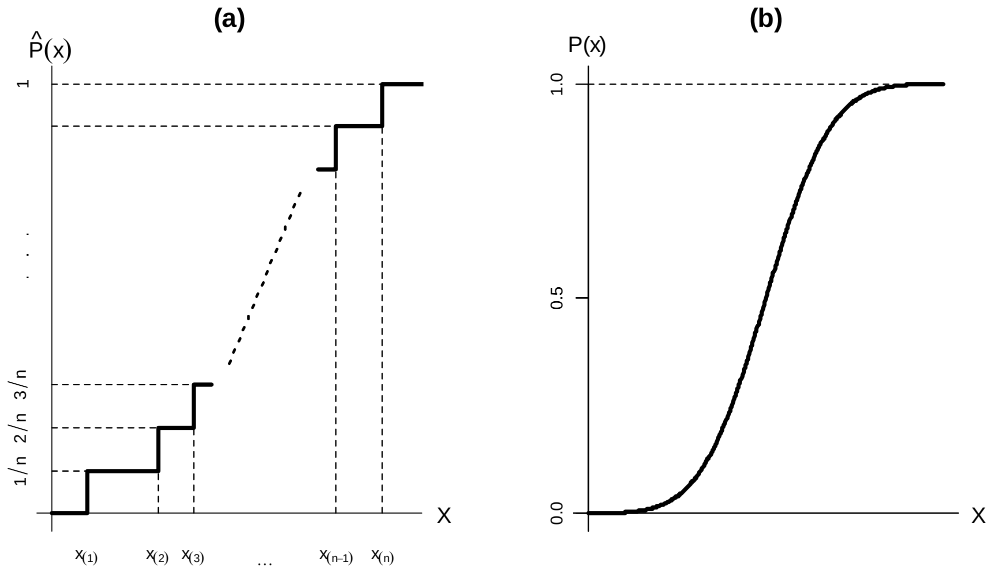

Figure 4. (a) Typical ECDF; (b) typical CDF.

▶ The quantile-comparison plot avoids this problem by never constructing the ECDF explicitly:

1. Order the data values from smallest to largest, denoted $X_{(1)}, X_{(2)}, ..., X_{(n)}$. The $X_{(i)}$ are called the *order statistics* of the sample.
2. By convention, the cumulative proportion of the data 'below' $X_{(i)}$ is given by
$$P_i = \frac{i - \frac{1}{2}}{n}$$
(or a similar formula).
3. Use the inverse of the CDF (the *quantile function*) to find the value z_i corresponding to the cumulative probability P_i; that is,
$$z_i = P^{-1}\left(\frac{i - \frac{1}{2}}{n}\right)$$
4. Plot the z_i as horizontal coordinates against the $X_{(i)}$ as vertical coordinates.

- If X is sampled from the distribution P, then $X_{(i)} \approx z_i$.
- If the distributions are identical except for location, then $X_{(i)} \approx \mu + z_i$.
- If the distributions are identical except for scale, then $X_{(i)} \approx \sigma z_i$.
- If the distributions differ both in location and scale but have the same shape, then $X_{(i)} \approx \mu + \sigma z_i$.

5. It is often helpful to place a comparison line on the plot to facilitate the perception of departures from linearity.
6. We expect some departure from linearity because of sampling variation; it therefore assists interpretation to display the expected degree of sampling error in the plot. The standard error of the order statistic $X_{(i)}$ is
$$\text{SE}(X_{(i)}) = \frac{\widehat{\sigma}}{p(z_i)}\sqrt{\frac{P_i(1 - P_i)}{n}}$$
where $p(z)$ is the probability-density function corresponding to the CDF $P(z)$.

- The values along the fitted line are given by $\widehat{X}_{(i)} = \widehat{\mu} + \widehat{\sigma} z_i$.
- An approximate 95 percent confidence 'envelope' around the fitted line is therefore

$$\widehat{X}_{(i)} \pm 2 \times \text{SE}(X_{(i)})$$

▶ Figure 5 display normal quantile-comparison plots for several illustrative distributions:

(a) A sample of $n = 100$ observations from a normal distribution with mean $\mu = 50$ and standard deviation $\sigma = 10$.

(b) A sample of $n = 100$ observations from the highly positively skewed χ^2 distribution with two degrees of freedom.

(c) A sample of $n = 100$ observations from the very-heavy-tailed t distribution with two degrees of freedom.

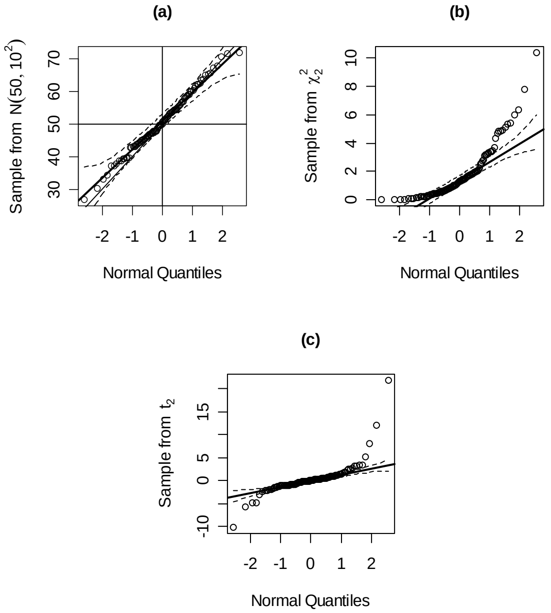

Figure 5. Normal quantile comparison plots for samples of size $n = 100$ drawn from three distributions.

- A normal quantile-comparison plot for the infant-mortality data appears in Figure 6.
  - The positive skew of the distribution is readily apparent.
  - The multi-modal character of the data, however, is not easily discerned in this display:
- Quantile-comparison plots highlight the tails of distributions.
  - This is important, because the behavior of the tails is often problematic for standard estimation methods like least-squares, but it is useful to supplement quantile-comparison plots with other displays.
- Quantile-comparison plots are usually used not to plot a variable directly but for derived quantities, such as residuals from a regression model.

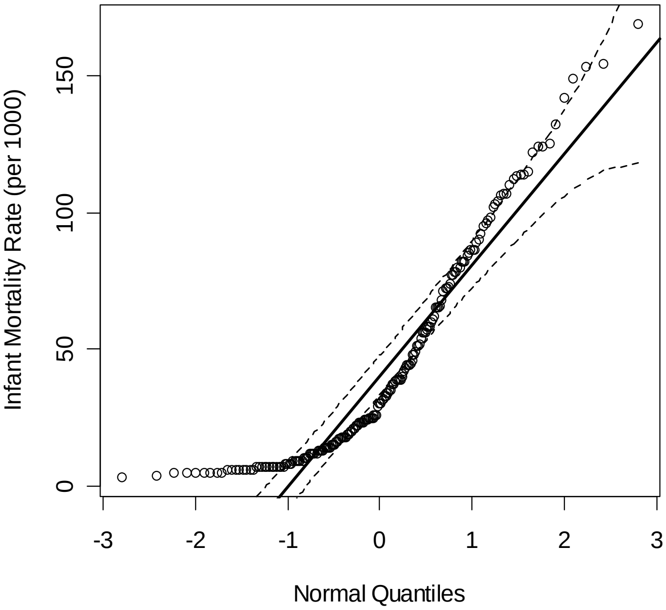

Figure 6. Normal quantile-comparison plot for infant mortality.

3.3.4 Boxplots

- Boxplots (due to John Tukey) present summary information on center, spread, skewness, and outliers.
- An illustrative boxplot, for the infant-mortality data, appears in Figure 7.
- This plot is constructed according to these conventions:

1. A scale is laid off to accommodate the extremes of the data.
2. The central box is drawn between the hinges, which are simply defined quartiles, and therefore encompasses the middle half of the data. The line in the central box represents the median.

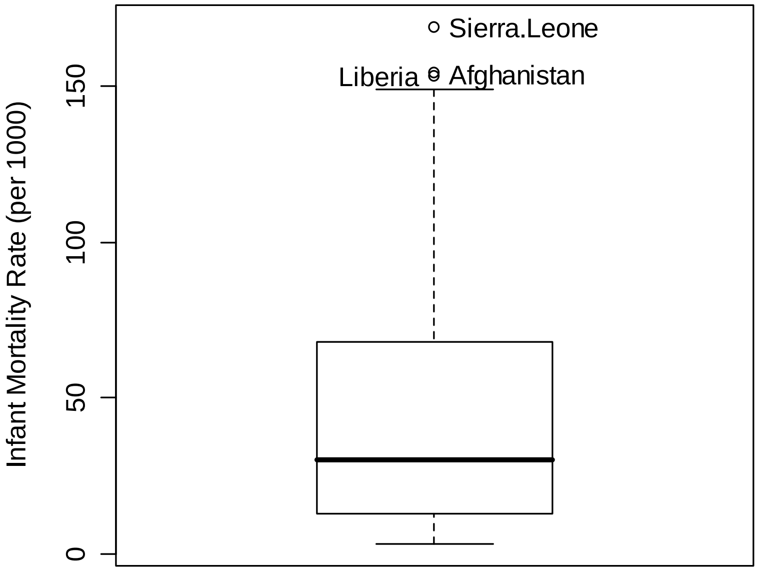

Figure 7. Boxplot of infant mortality.

3. The following rule is used to identify outliers, which are shown individually in the boxplot:
   - 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-spreads beyond the hinges:
   $$\begin{aligned} \mathsf{F}_L &= H_L - 1.5 \times H\text{-spread} \\ \mathsf{F}_U &= H_U + 1.5 \times H\text{-spread} \end{aligned}$$
   Observations beyond fences are identified as outliers. The fences themselves are not shown in the display. (Points beyond $\pm 3 \times H$-spread are extreme outliers.)
   - The ‘whisker’ growing from each end of the central box extends either to the extreme observation on its side of the distribution (as at the low end of the infant-mortality data) or to the most extreme non-outlying observation, called the ‘adjacent value’ (as at the high end of the infant-mortality distribution).

▶ The boxplot of the infant-mortality distribution clearly reveals the skewness of the distribution:
   - The lower whisker is shorter than the upper whisker; and there are outlying observations at the upper end of the distribution, but not at the lower end.
   - The median is closer to the lower hinge than to the upper hinge.
   - The apparent multi-modality of the infant-mortality data is not represented in the boxplot.

▶ Boxplots are most useful as adjuncts to other displays (e.g., in the margins of a scatterplot) or for comparing several distributions.

3.4 Plotting Bivariate Data

▶ The scatterplot — a direct geometric representation of observations on two quantitative variables (generically, Y and X)— is the most useful of all statistical graphs. Scatterplots are familiar, so I will limit this presentation to a few points (see Figure 8):

- It is convenient to work in a computing environment that permits the interactive identification of observations in a scatterplot.
- Since relationships between variables in many disciplines are often weak, scatterplots can be dominated visually by 'noise.' It often helps to enhance the plot with a non-parametric regression of Y on X.
- Scatterplots in which one or both variables are highly skewed are difficult to examine because the bulk of the data congregate in a small part of the display. It often helps to 'correct' substantial skews prior to examining the relationship between Y and X.
- Scatterplots in which the variables are discrete are difficult to examine.

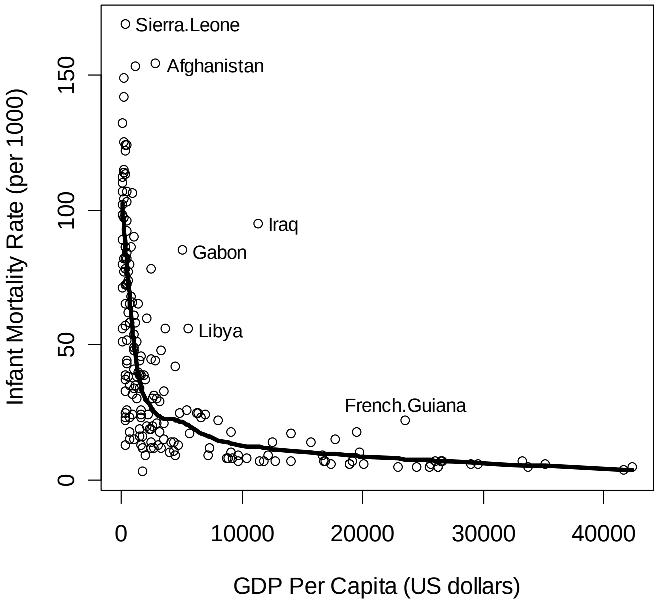

Figure 8. Scatterplot of infant morality by GDP per capita, for the UN data. The solid line is for a lowess smooth with a span of .5.

- An extreme instance of this phenomenon is shown in Figure 9, which plots scores on a ten-item vocabulary test included in NORC's General Social Survey against years of education.
  - One solution — particularly useful when only X is discrete — is to focus on the conditional distribution of Y for each value of X.
  - Boxplots, for example, can be employed to represent the conditional distributions.
  - Another solution is to separate overlapping points by adding a small random quantity to the discrete scores. For example, I have added a uniform random variable on the interval $[-0.4, +0.4]$ to each of vocabulary and education.

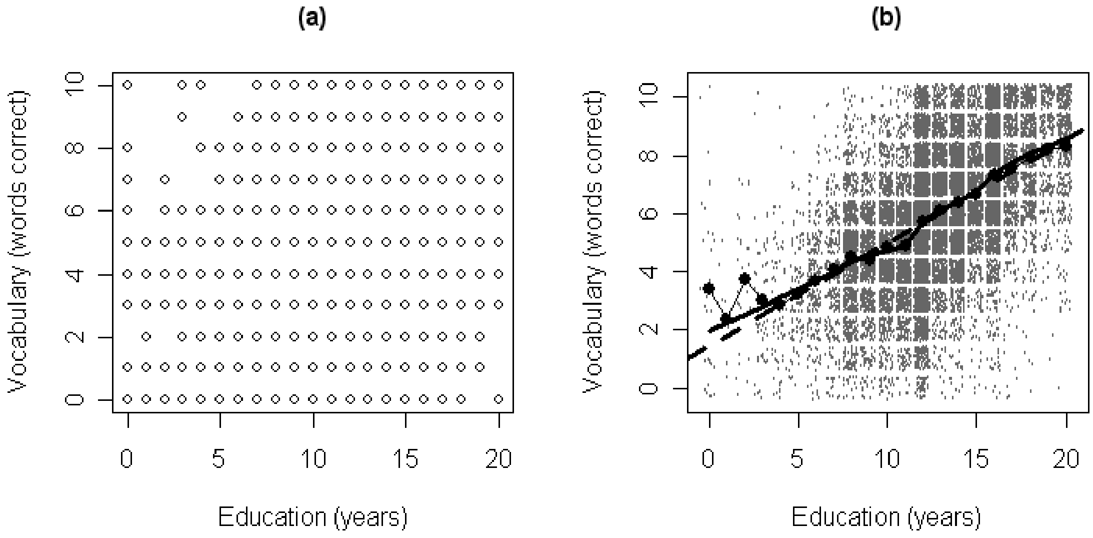

Figure 9. Vocabulary score by education: (a) original scatterplot; (b) jittered, with the least-squares lines, lowess line (for span = .2), and conditional means.

- ▶ As mentioned, when the explanatory variable is discrete, parallel boxplots can be used to display the conditional distributions of Y.
  - One common case occurs when the explanatory variable is a qualitative/categorical variable.
  - An example is shown in Figure 10, using data collected by Michael Ornstein (1976) on interlocking directorates among the 248 largest Canadian firms.
    - The response variable in this graph is the number of interlocking directorships and executive positions maintained by each firm with others in the group of 248.
    - The explanatory variable is the nation in which the corporation is controlled, coded as Canada, United Kingdom, United States, and other foreign.
    - It is relatively difficult to discern detail in this display: first, because the conditional distributions of interlocks are positively skewed; and, second, because there is an association between level and spread.

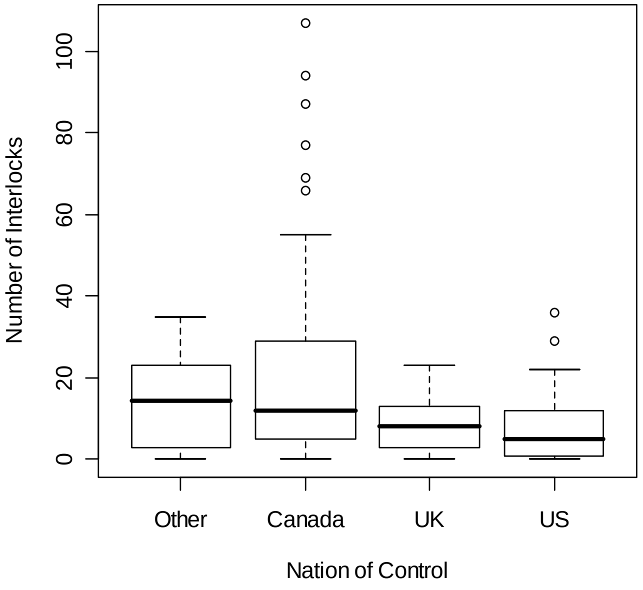

Figure 10. Parallel boxplots of number of interlocks by nation of control, for Ornstein's interlocking-directorate data.

3.5 Plotting Multivariate Data

▶ Because paper and computer screens are two-dimensional, graphical display of multivariate data is intrinsically difficult.

- Multivariate displays for quantitative data often project the higher-dimensional ‘point cloud’ of the data onto a two-dimensional space.
- The essential trick of effective multidimensional display is to select projections that reveal important characteristics of the data.
- In certain circumstances, projections can be selected on the basis of a statistical model fit to the data or on the basis of explicitly stated criteria.

▶ A simple approach to multivariate data, which does not require a statistical model, is to examine bivariate scatterplots for all pairs of variables.

- Arraying these plots in a ‘scatterplot matrix’ produces a graphical analog to the correlation matrix.

- Figure 11 shows an illustrative scatterplot matrix, for data from Duncan (1961) on the prestige, education, and income levels of 45 U.S. occupations.
- 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.

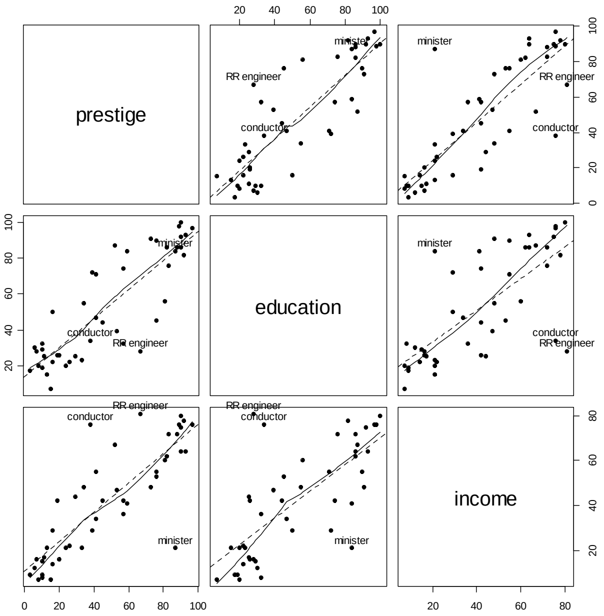

Figure 11. Scatterplot matrix for prestige, income, and education in Duncan's occupational prestige data.

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

- Despite this intrinsic limitation, scatterplot matrices often uncover interesting features of the data, and this is indeed the case here, where the display reveals three unusual observations: *Ministers*, *railroad conductors,* and *railroad engineers*.

▶ Information about a categorical third variable may be entered on a bivariate scatterplot by coding the plotting symbols. The most effective codes use different colors to represent categories, but degrees of fill, distinguishable shapes, and distinguishable letters can also be effective. (See, e.g., Figure 12, which uses data from Caroline Davis (1990) on weight and reported weight of regular exercisers.)

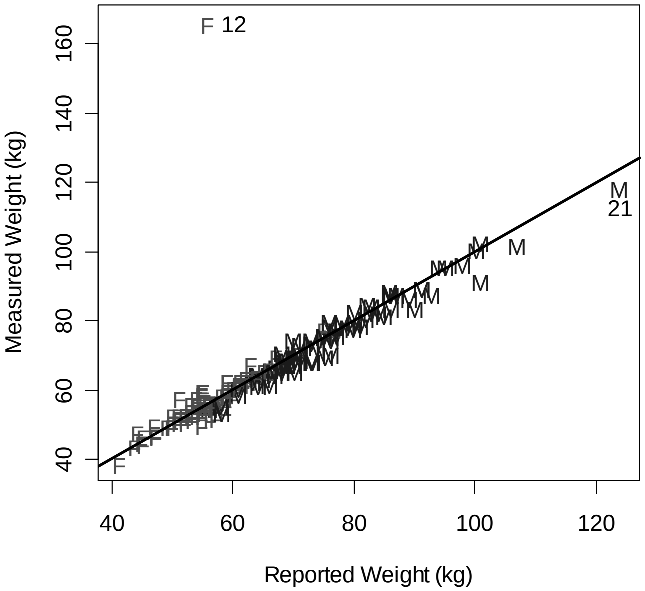

Figure 12. Measured by reported weight for 183 men (M) and women (F) engaged in regular exercise.

▶ Another useful multivariate display, directly applicable only to three variables at a time, is the three-dimensional scatterplot.

- This display is an illusion produced by modern statistical software, since the graph really represents a projection of a three-dimensional scatterplot onto a two-dimensional computer screen.
- Nevertheless, motion (e.g., rotation) and the ability to interact with the display — sometimes combined with the effective use of perspective, color, depth-cueing, fitted surfaces, and other visual devices — can produce a vivid impression of directly examining a three-dimensional space.

3.6 Transformations: The Family of Powers and Roots

▶ 'Classical' statistical models make strong assumptions about the structure of data, assumptions which often fail to hold in practice.

- One solution is to abandon classical methods.
- Another solution is to transform the data so that they conform more closely to the assumptions.
- As well, transformations can often assist in the examination of data in the absence of a statistical model.

▶ A particularly useful group of transformations is the 'family' of powers and roots:

$$X \rightarrow X^p$$

- If p is negative, then the transformation is an inverse power: $X^{-1} = 1/X$, and $X^{-2} = 1/X^2$.
- If p is a fraction, then the transformation represents a root: $X^{1/3} = \sqrt[3]{X}$ and $X^{-1/2} = 1/\sqrt{X}$.

▶ It is sometimes convenient to define the family of power transformations in a slightly more complex manner (called the *Box-Cox family*):

$$X \rightarrow X^{(p)} \equiv \frac{X^p - 1}{p}$$

▶ Since $X^{(p)}$ is a linear function of X^p, the two transformations have the same essential effect on the data, but, as is apparent in Figure 13, $X^{(p)}$ reveals the essential unity of the family of powers and roots:

- Dividing by p preserves the direction of X, which otherwise would be reversed when p is negative:

X	X^{-1}	$\frac{X^{-1}-1}{-1}$
1	1	0
2	$1/2$	$1/2$
3	$1/3$	$2/3$
4	$1/4$	$3/4$

- The transformations $X^{(p)}$ are 'matched' above $X = 1$ both in level and slope.

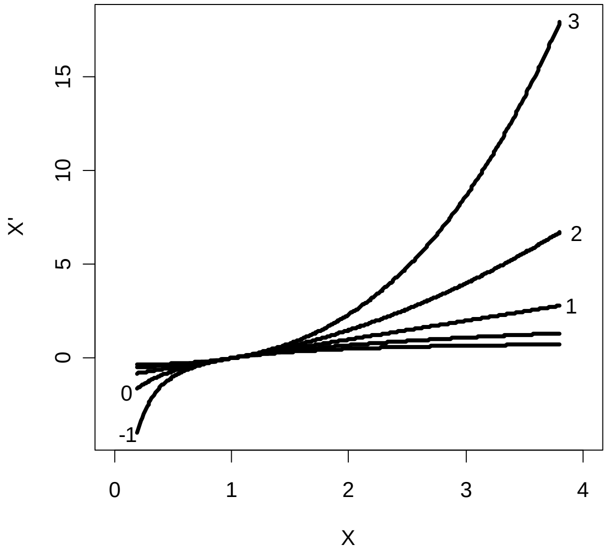

Figure 13. The Box-Cox famlily of modified power transformations, $X^{(p)} = (X^p - 1)/p$, for values of $p = -1, 0, 1, 2, 3$. When $p = 0$, $X^{(p)} = \log_e X$.

- The power transformation X^0 is useless, but the very useful $\log$ transformation is a kind of ‘zeroth’ power:

$$\lim_{p \to 0} \frac{X^p - 1}{p} = \log_e X$$

where $e \approx 2.718$ is the base of the natural logarithms. Thus, we will take $X^{(0)} \equiv \log(X)$.

  - It is generally more convenient to use logs to the base 10 or base 2, which are more easily interpreted than logs to the base e.
  - Changing bases is equivalent to multiplying by a constant.

- Review of logs:
  - logs are exponents: $\log_b x = y$ (“the log of x to the base b is y”) means that $b^y = x$.
  - Some examples:

$$
\begin{array}{rcc}
\log_{10} 100 = 2 & \Longleftrightarrow & 10^2 = 100 \\
\log_{10} 0.01 = -2 & \Longleftrightarrow & 10^{-2} = \frac{1}{10^2} = 0.01 \\
\log_{10} 10 = 1 & \Longleftrightarrow & 10^1 = 10 \\
\log_2 8 = 3 & \Longleftrightarrow & 2^3 = 8 \\
\log_2 \left(\frac{1}{8}\right) = -3 & \Longleftrightarrow & 2^{-3} = \frac{1}{2^3} = \frac{1}{8} \\
\log_b 1 = 0 & \Longleftrightarrow & b^0 = 1
\end{array}
$$

- Descending the ‘ladder’ of powers and roots from $p = 1$ (i.e., no transformation) towards $X^{(-1)}$ compresses the large values of X and spreads out the small ones

- Ascending the ladder of powers and roots towards $X^{(2)}$ has the opposite effect.

$-\frac{1}{X}$	$\log_2 X$	X	X^2	X^3
-1	0	1	1	1
$\frac{1}{2}$ {	1 {	} 1	} 3	} 7
$-1/2$	1	2	4	8
$\frac{1}{6}$ {	0.59 {	} 1	} 5	} 19
$-1/3$	1.59	3	9	27
$\frac{1}{12}$ {	0.41 {	} 1	} 7	} 37
$-1/4$	2	4	16	64

- Power transformations are sensible only when all of the values of X are positive.
  - First of all, some of the transformations, such as log and square root, are undefined for negative or zero values.

- Second, the power transformations are not monotone when there are both positive and negative values in the data:

X	X^2
-2	4
-1	1
0	0
1	1
2	4

- We can add a positive constant (called a ‘start’) to each data value to make all of the values positive: $X \rightarrow (X+s)^p$:

X	$(X+3)^2$
-2	1
-1	4
0	9
1	16
2	25

- Alternatively, we can use the Yeo-Johnson family of modified power transformations (Yeo and Johnson, 2000), which are defined as follows for power p:

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

▶ Power transformations are effective only when the ratio of the biggest data values to the smallest ones is sufficiently large; if this ratio is close to 1, then power transformations are nearly linear; in the following example, $1995/1991 = 1.002 \approx 1$:

X	$\log_{10} X$
1991	3.2991
1 {	} 0.0002
1992	3.2993
1 {	} 0.0002
1993	3.2995
1 {	} 0.0002
1994	3.2997
1 {	} 0.0002
1995	3.2999

- Using a negative start produces the desired effect:

X	$\log_{10}(X - 1990)$
1991	0
1{	}0.301
1992	0.301
1{	}0.176
1993	0.477
1{	}0.125
1994	0.602
1{	}0.097
1995	0.699

▶ Using reasonable starts, if necessary, an adequate power transformation can usually be found in the range $-2 \leq p \leq 3$.

3.6.1 Transforming Skewness

▶ Power transformations can make a skewed distribution more symmetric. But why should we bother?

- 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.

▶ How a power transformation can eliminate a positive skew:

X	$\log_{10} X$
1	0
9 {	} 1
10	1
90 {	} 1
100	2
900 {	} 1
1000	3

- 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.

▶ For infant mortality in the UN data, the log transformation works well, as shown in Figure 14.

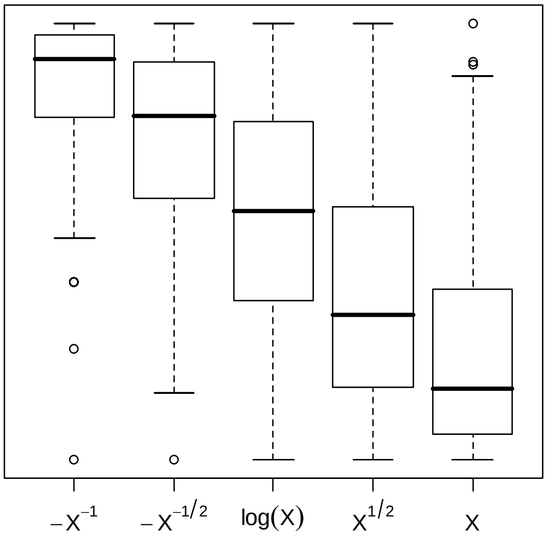

Figure 14. Boxplots for various transformations down the ladder of powers and roots for infant mortality in the UN datqa.

- If we have a choice between transformations that perform roughly equally well, we may prefer one transformation to another because of interpretability:
  - The log transformation has a convenient multiplicative interpretation (e.g. adding 1 to $\log_2 X$ doubles X; adding 1 to $\log_{10} X$ multiples X by 10.
  - In certain contexts, other transformations may have specific substantive meanings:
    - The inverse of time required to travel a fixed distance (e.g., hours for 1 km) is speed (km per hour).
    - The inverse of response latency (e.g., in a psychophysical experiment, in milliseconds) is response frequency (responses per 1000 seconds).

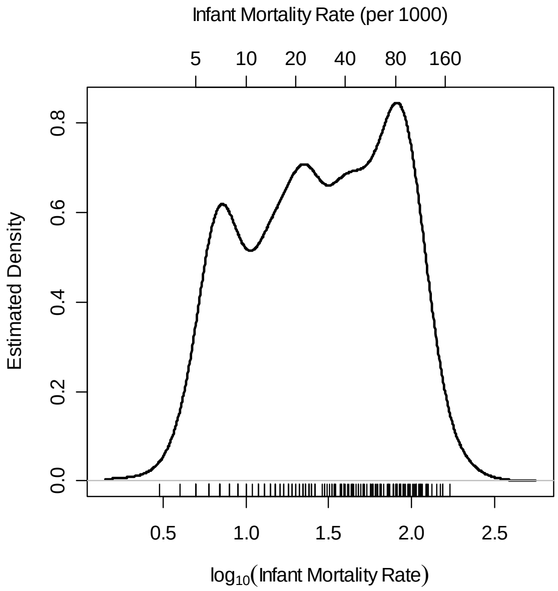

Figure 15. Adaptive-kernel density estimate for log-transformed infant mortality.

- · The square root of a measure of area (say, in m^2) is a linear measure of size (in meters).
- · The cube of a linear measure (say in cm) can be interpreted as a volume (cm^3).

▶ One can also label an axis with the original units, as in Figure 15.

3.6.2 Transforming Nonlinearity

▶ Power transformations can also be used to make many nonlinear relationships more nearly linear. Again, why bother?

- Linear relationships — expressible in the form $\widehat{Y} = a + bX$ — are particularly simple.
- When there are several explanatory variables, the alternative of nonparametric regression may not be feasible or may be difficult to visualize.
- There is a simple and elegant statistical theory for linear models.
- There are certain technical advantages to having linear relationships among the *explanatory* variables in a regression analysis.

▶ The following simple example suggests how a power transformation can serve to straighten a nonlinear relationship; here, $Y = \frac{1}{5}X^2$ (with no residual):

X	Y
1	0.2
2	0.8
3	1.8
4	3.2
5	5.0

- These ‘data’ are graphed in part (a) of Figure 16.
- We could replace Y by $Y' = \sqrt{Y}$, in which case $Y' = \sqrt{\frac{1}{5}}X$ [see (b)].
- We could replace X by $X' = X^2$, in which case $Y = \frac{1}{5}X'$ [see (c)].

▶ A power transformation works here because the relationship between Y and X is both monotone and simple. In Figure 17:

- the curve in (a) is simple and monotone;

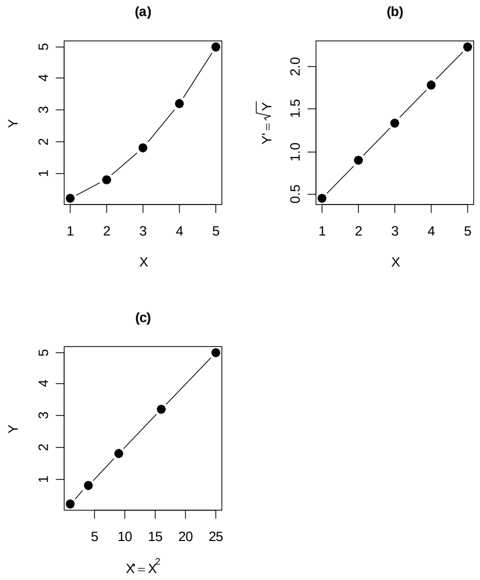

Figure 16. Transformating a nonlinear relationship (a) to linearity, (b) or (c).

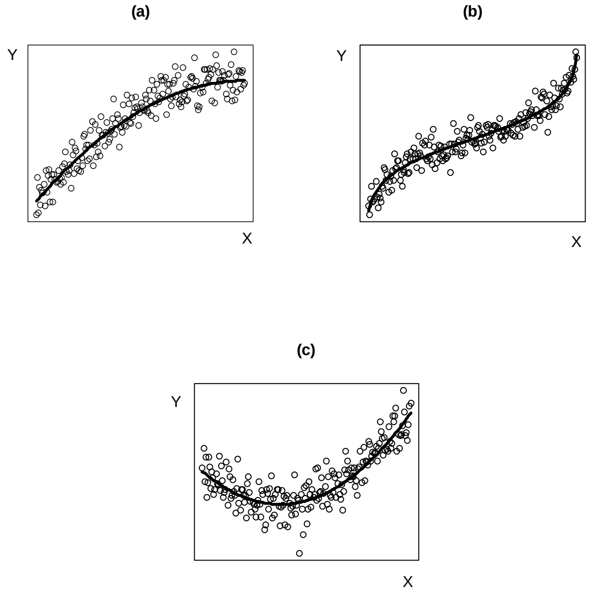

Figure 17. (a) A simple monotone relationship. (b) A monotone relationship that is not simple. (c) A simple nonmonotone relationship.

  - in (b) monotone, but not simple;
  - in (c) simple but not monotone.
    - In (c), we could fit a quadratic model, $\widehat{Y} = a + b_1 X + b_2 X^2$.

- ► Figure 18 introduces Mosteller and Tukey's 'bulging rule' for selecting a transformation.
  - For example, if the 'bulge' points *down* and to the *right*, we need to transform Y *down* the ladder of powers or X *up* (or both).
  - Figure 19 shows the relationship between prestige and average income for 102 Canadian occupations around 1970.
    - The relationship between prestige and income is clearly monotone and nonlinear.
    - Since the bulge points up and to the left, we can try transforming prestige up the ladder of powers or income down.
    - The cube-root transformation of income works reasonably well.

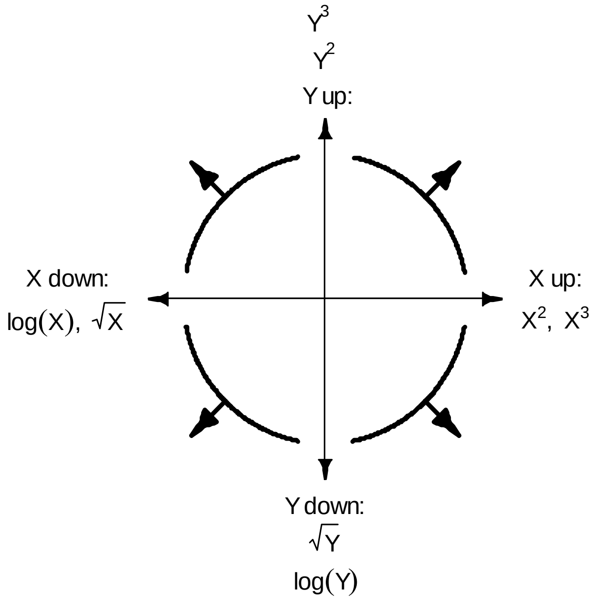

Figure 18. Mosteller and Tukey's bulging rule for selecting linearizing transformations.

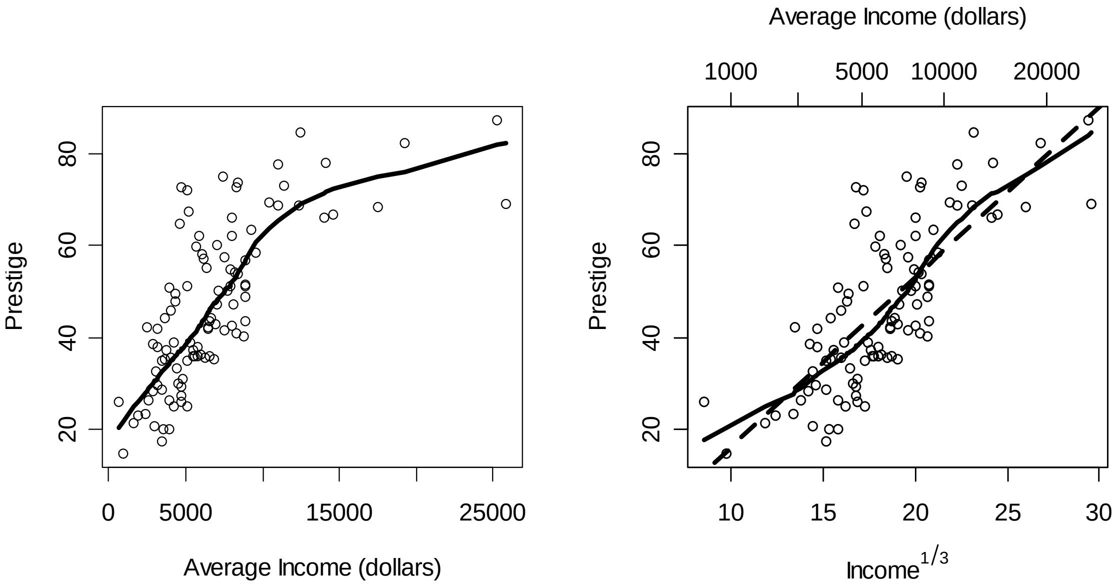

Figure 19. Transformating the relationship between prestige and income to (near) linearity: (left) original scatterplot; (right) with income transformed.

- A more extreme example appears in Figure 20, which shows the relationship between the infant-mortality rate and GDP per capita in the UN data.
  - The skewness of infant mortality and income makes the scatterplot difficult to interpret; the nonparametric regression reveals a nonlinear but monotone relationship.
  - The bulging rule suggests that infant mortality or income should be transformed down the ladder of powers and roots.

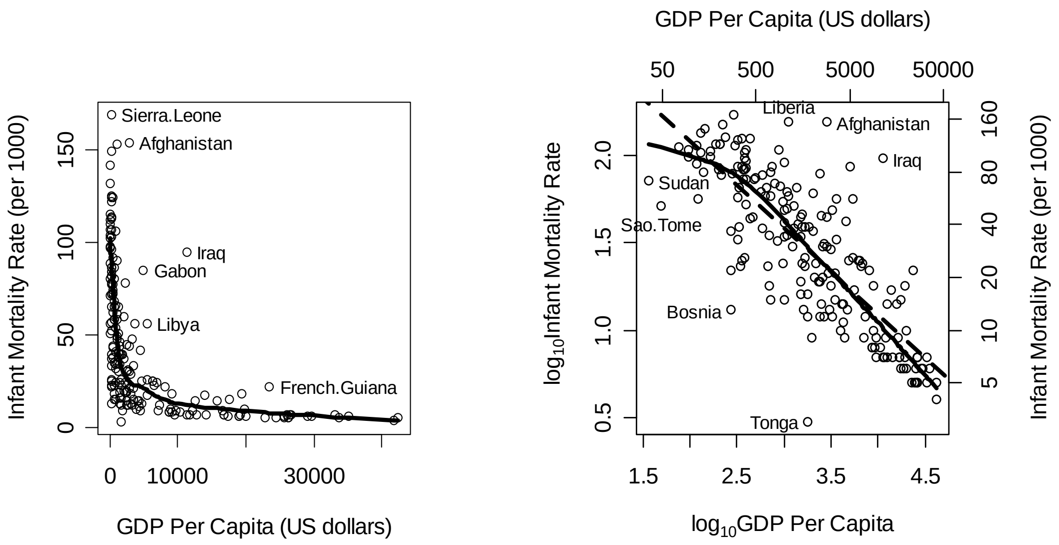

Figure 20. Transforming the relationship between infant mortality and GDP per capita.

- Transforming both variables by taking logs makes the relationship nearly linear; the least-squares fit is:

$$\log_{10}\widehat{\text{Infant mortality}} = 3.06 - 0.493 \times \log_{10}\text{GDP}$$

  - Because both variables are expressed on log scales to the same base, the slope of this relationship has a simple interpretation: A one-percent increase in per-capita income is associated on average with an approximate half-percent decline in the infant-mortality rate.
  - Economists call this type of number an ‘elasticity.’

3.6.3 Transforming Non-Constant Spread

- When a variable has very different degrees of variation in different groups, it becomes difficult to examine the data and to compare differences in level across the groups.
  - Recall Ornstein's Canadian interlocking-directorate data, examining the relationship between number of interlocks and nation of control.
- Differences in spread are often systematically related to differences in level.
  - Using the median and hinge-spread (inter-quartile range) as indices of level and spread, respectively, the following table shows that there is indeed an association, if an imperfect one, between spread and level for Ornstein's data:

Nation of Control	*Lower Hinge*	*Median*	*Upper Hinge*	*Hinge Spread*
Other	3	14.5	23	20
Canada	5	12.0	29	24
United Kingdom	3	8.0	13	10
United States	1	5.0	12	11

- Tukey suggests graphing the log hinge-spread against the log median, as shown in Figure 21.
  - Because some firms maintained zero interlocks, I used a start of 1.
  - 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:
    - Express the linear fit as
      $$\text{log-spread } \approx a + b \text{ log-level}$$
    - Then the corresponding spread-stabilizing transformation uses the power $p = 1 - b$.

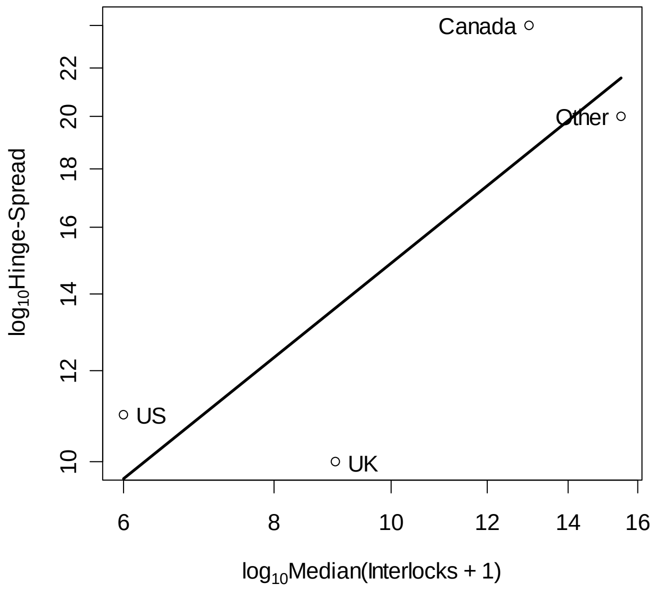

Figure 21. Spread-level plot for Ornstein's interlocking-directorate data.

- For Ornstein's data, the slope of the least-squares line is $b = 0.85$, suggesting the power transformation, $p = 1 - 0.85 = 0.15 \approx 0$ (i.e., log). See the Figure 22, using logs to the base 2 (and plotting on a log-scaled axis).

▶ The problems of unequal spread and skewness commonly occur together, because they often have a common origin:

- Here, the data represent frequency counts (*number* of interlocks); the impossibility of obtaining a negative count tends to produce positive skewness, together with a tendency for larger levels to be associated with larger spreads.

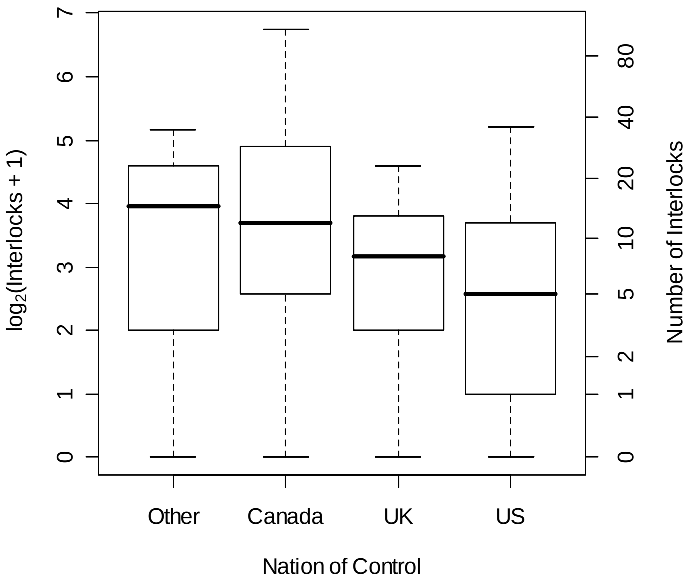

Figure 22. Ornstein’s interlocking-directorate data, log-transforming interlocks (with a start of 1).

3.7 Transforming Proportions

- ▶ Power transformations are often not helpful for proportions, since these quantities are bounded below by 0 and above by 1.
  - If the data values do not approach these two boundaries, then proportions can be handled much like other sorts of data.
  - Percents and many sorts of rates are simply rescaled proportions.
  - It is common to encounter ‘disguised’ proportions, such as the number of questions correct on an exam of fixed length.
- ▶ An example, drawn from the Canadian occupational prestige data, is shown in the stem-and-leaf display (a type of histogram) in Figure 23. The distribution is for the percentage of women among the incumbents of each of 102 occupations.

```
Unit: 1    Lines/stem: 2
     1|2 <--> 12

depth
 32  0|00000000000000111111222233334444
 44  0|555566777899
  8) 1|01111333
 50  1|5557779
 43  2|1344
 39  2|57
 37  3|01334
 32  3|99
 30  4|
 30  4|678
 27  5|224
 24  5|67
 22  6|3
 21  6|789
 18  7|024
 15  7|5667
 11  8|233
  8  8|
  8  9|012
  5  9|56667
```

Figure 23. Stem-and-leaf display of percent women in the Canadian occupational prestige data. Notice the “stacking up” near the boundaries of 0 and 100.

▶ Several transformations are commonly employed for proportions; the most important is the *logit* transformation:

$$P \rightarrow \text{logit}(P) = \log_e \frac{P}{1-P}$$

- The logit transformation is the log of the ‘odds,’ $P/(1-P)$.
- The ‘trick’ of the logit transformation is to remove the upper and lower boundaries of the scale, spreading out the tails of the distribution and making the resulting quantities symmetric about 0; for example:

P	$\frac{P}{1-P}$	logit
.05	$1/19$	-2.94
.1	$1/9$	-2.20
.3	$3/7$	-0.85
.5	1	0
.7	$7/3$	0.85
.9	$9/1$	2.20
.95	$19/1$	2.94

- The logit transformation is graphed in Figure 24. Note that the transformation is nearly linear in its center, between about $P = .2$ and $P = .8$.
- The logit transformations cannot be applied to proportions of exactly 0 or 1.
  - · If we have access to the original counts, we can define adjusted proportions
    $$P' = \frac{F + \frac{1}{2}}{N + 1}$$
    in place of P.
    - · Here, F is the frequency count in the focal category (e.g., number of women) and N is the total count (total number of occupational incumbents, women plus men).

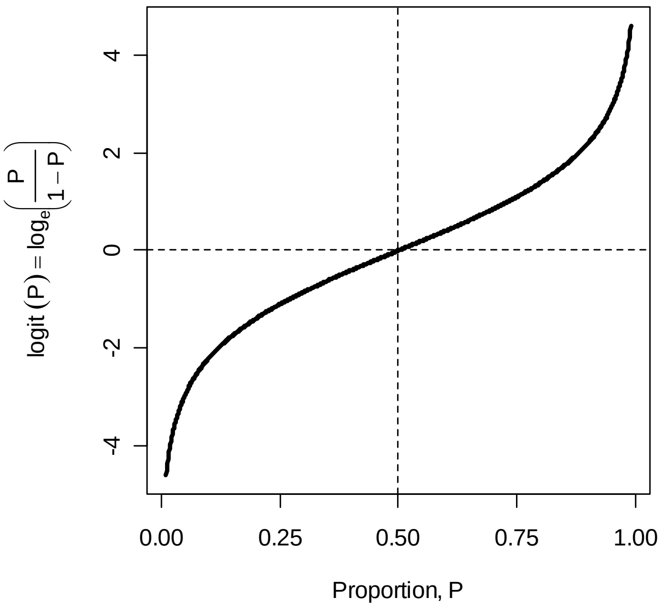

Figure 24. The logit transformation of a proportion.

- · If the original counts are not available, then we can remap the proportions to an interval that excludes 0 and 1.
  - · For example, $P' = .005 + .99 \times P$ remaps proportions to the interval [.005, .995].
- The distribution of logit(P'_{women}) for the Canadian occupational prestige data appears in Figure 25.
- We will encounter logits again when we talk about generalized linear models for categorical data.

```
Unit: 0.1    Lines/stem: 2
     1|2 <--> 1.2

depth
  5  -4|77777
  8  -3|444
 16  -3|55667888
 21  -2|01124
 31  -2|5567888999
 39  -1|01112344
 48  -1|556779999
 10) -0|0111333444
 44  -0|668889
 38   0|01233355889
 27   0|00122577889
 16   1|01111
 11   1|556
  8   2|23
  6   2|5
  5   3|00014
```

Figure 25. Logit-transformed percent women.

3.8 Summary: Data Craft

- ▶ Statistical graphs are central to effective data analysis, both in the early stages of an investigation and in statistical modeling.
- ▶ There are many useful univariate displays, including the traditional histogram.
  - Nonparametric density estimation may be employed to smooth a histogram.
  - Quantile comparison plots are useful for comparing data with a theoretical probability distribution.
  - Boxplots summarize some of the most important characteristics of a distribution, including center, spread, skewness, and the presence of outliers.

- The bivariate scatterplot is a natural graphical display of the relationship between two quantitative variables.
  - Interpretation of a scatterplot can often be assisted by graphing a nonparametric regression, which summarizes the relationship between the two variables.
  - Scatterplots of the relationship between discrete variables can be enhanced by randomly jittering the data.
- Parallel boxplots can be employed to display the relationship between a quantitative response variable and a discrete explanatory variable.
- Visualizing multivariate data is intrinsically difficult because we cannot directly examine higher-dimensional scatterplots.
  - Effective displays project the higher-dimensional point cloud onto two or three dimensions.
  - These displays include the scatterplot matrix and the dynamic three-dimensional scatterplot.

- Transformations can often facilitate the examination and modeling of data.
- The powers and roots are a particularly useful family of transformations: $X \to X^p$.
  - We employ the log transformation in place of X^0.
- Power transformations preserve the order of the data only when all values are positive, and are effective only when the ratio of largest to smallest data values is itself large.
  - When these conditions do not hold, we can impose them by adding a positive or negative start to all of the data values.
- Descending the ladder of powers (e.g., to $\log X$) tends to correct a positive skew; ascending the ladder of powers (e.g., to X^2) tends to correct a negative skew.
- Simple monotone nonlinearity can often be corrected by a power transformation of X, of Y, or of both variables.

- Mosteller and Tukey's 'bulging rule' assists in the selection of a transformation.

▶ When there is a positive association between the level of a variable in different groups and its spread, the spreads can be made more constant by descending the ladder of powers. A negative association between level and spread is less common, but can be corrected by ascending the ladder of powers.

▶ Power transformations are ineffective for proportions P that push the boundaries of 0 and 1, and for other variables (e.g., percents, rates, disguised proportions) that are bounded both below and above.

- The logit transformation, $P \rightarrow \log[P/(1-P)]$ often works well for proportions.

4. Unusual and Influential Data in Least-Squares Regression

▶ Linear statistical models make strong assumptions about the structure of data, which often do not hold in applications.

▶ The method of least-squares is very sensitive to the structure of the data, and can be markedly influenced by one or a few unusual observations.

▶ We could abandon linear models and least-squares estimation in favor of nonparametric regression and robust estimation.

▶ Alternatively, we can adapt and extend methods for examining and transforming data to diagnose problems with a linear model, and to suggest solutions.

4.1 Goals

- To distinguish among regression outliers, high-leverage observations, and influential observations.
- To show how outlyingness, leverage, and influence can be measured.
- To introduce added-variable ('partial-regression') plots as a means of displaying leverage and influence on particular coefficients.

4.2 Outliers, Leverage, and Influence

- Unusual data are problematic in linear models fit by least squares because they can unduly influence the results of the analysis, and because their presence may be a signal that the model fails to capture important characteristics of the data.
- Some central distinctions are illustrated in Figure 26 for the simple regression model $Y = \alpha + \beta X + \varepsilon$.
  - In simple regression, an *outlier* is an observation whose response-variable value is conditionally unusual given the value of the explanatory variable.
  - In contrast, a univariate outlier is a value of Y or X that is unconditionally unusual; such a value may or may not be a regression outlier.

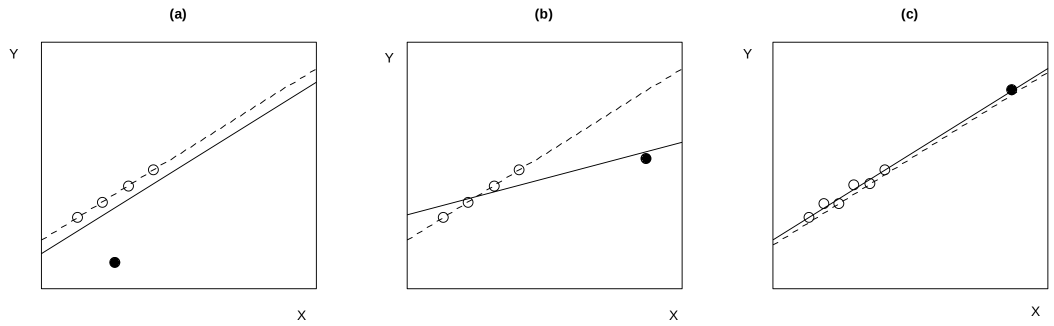

Figure 26. Unusual data in regression: (a) a low-leverage and hence uninfluential outlier; (b) a high-leverage and hence influential outlier; (c) a high-leverage in-line observation. In each case, the solid line is the least--squares line for all of the data; the broken line is the least-squares line with the unusual observation omitted.

- Regression outliers appear in (a) and (b).
  - In (a), the outlying observation has an X-value that is at the center of the X distribution; deleting the outlier has little impact on the least-squares fit.
  - In (b), the outlier has an unusual X-value; its deletion markedly affects both the slope and the intercept. Because of its unusual X-value, the outlying last observation in (b) exerts strong *leverage* on the regression coefficients, while the outlying middle observation in (a) is at a low-leverage point. The combination of high leverage with a regression outlier produces substantial *influence* on the regression coefficients.
  - In (c), the last observation has no influence on the regression coefficients even though it is a high-leverage point, because this observation is in line with the rest of the data.

- The following heuristic formula helps to distinguish among the three concepts of influence, leverage and discrepancy (‘outlyingness’):

Influence on Coefficients = Leverage × Discrepancy

▶ A simple example with real data from Davis (1990) appears in Figure 27. The data record the measured and reported weight of 183 male and female subjects who engage in programs of regular physical exercise. Davis’s data can be treated in two ways:

1. We could regress reported weight (RW) on measured weight (MW), a dummy variable for sex (F, coded 1 for women and 0 for men), and an interaction regressor (formed as the product $MW \times F$):

$$\begin{array}{rcccccccc} \widehat{RW} & = & 1.36 & + & 0.990MW & + & 40.0F & - & 0.725(MW \times F) \\ & & (3.28) & & (0.043) & & (3.9) & & (0.056) \end{array}$$

$$R^2 = 0.89 \quad S_E = 4.66$$

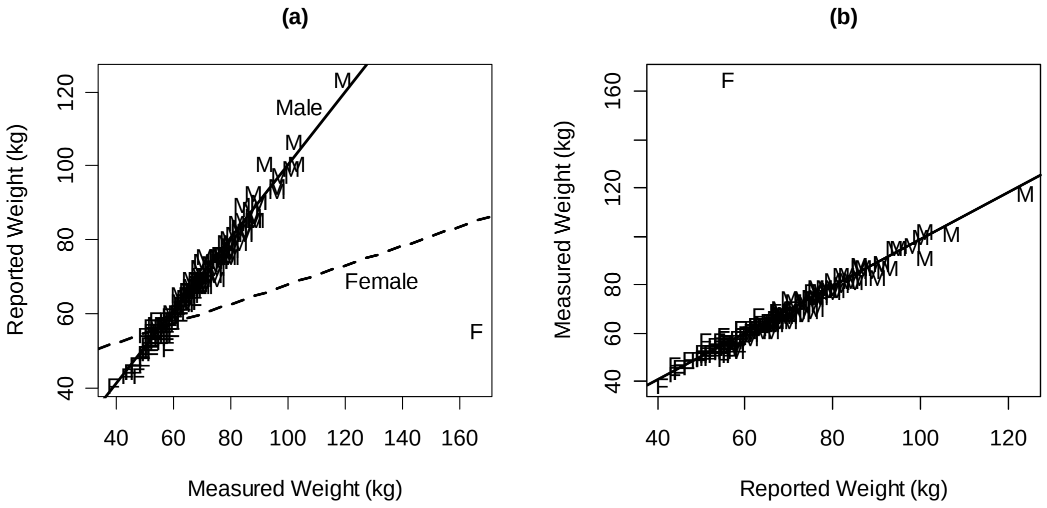

Figure 27. (a) Regressing reported weight on measured weight, sex, and their interaction; (b) regressing measured weight on reported weight, sex, and their interaction.

- Were these results taken seriously, we would conclude that men are unbiased reporters of their weights (because $A \approx 0$ and $B_1 \approx 1$), while women tend to over-report their weights if they are relatively light and under-report if they are relatively heavy.
- The figure makes it clear that the differential results for women and mean are due to one erroneous data point.
- Correcting the data produces the regression

$$\begin{array}{rlllll} \widehat{RW} & = 1.36 & + 0.990MW & + 1.98F & - 0.0567(MW \times F) \\ & \quad (1.58) & \quad (0.021) & \quad (2.45) & \quad (0.0385) \end{array}$$
$$R^2 = 0.97 \quad S_E = 2.24$$

2. We could regress measured weight on reported weight, sex, and their interaction:

$$\begin{array}{rlllll} \widehat{MW} & = 1.79 & + 0.969RW & + 2.07F & - 0.00953(MW \times F) \\ & \quad (5.92) & \quad (0.076) & \quad (9.30) & \quad (0.147) \end{array}$$
$$R^2 = 0.70 \quad S_E = 8.45$$

- The outlier does not have much impact on the regression coefficients because the value of RW for the outlying observation is near $\overline{RW}$ for women.
- There is, however, a marked effect on the multiple correlation and standard error: For the corrected data, $R^2 = 0.97$ and $S_E = 2.25$.

4.3 Assessing Leverage: Hat-Values

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

$$\widehat{Y}_j = h_{1j}Y_1 + h_{2j}Y_2 + \cdots + h_{jj}Y_j + \cdots + h_{nj}Y_n = \sum_{i=1}^{n} h_{ij}Y_i$$

  - Thus, the weight h_{ij} captures the contribution of observation Y_i to the fitted value $\widehat{Y}_j$: 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.
  - $1/n \le h_i \le 1$

  - The average hat-value is $\overline{h} = (k+1)/n$.
  - In simple-regression analysis, the hat-values measure distance from the mean of X:

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

  - In multiple regression, h_i measures distance from the centroid (point of means) of the X's, taking into account the correlational and variational structure of the X's, as illustrated for $k = 2$ in Figure 28. Multivariate outliers in the X-space are thus high-leverage observations. The response-variable values are not at all involved in determining leverage.

- For Davis's regression of reported weight on measured weight, the largest hat-value by far belongs to the 12*th* subject, whose measured weight was wrongly recorded as 166 kg.: $h_{12} = 0.714$. This quantity is many times the average hat-value, $\overline{h} = (3+1)/183 = 0.0219$.

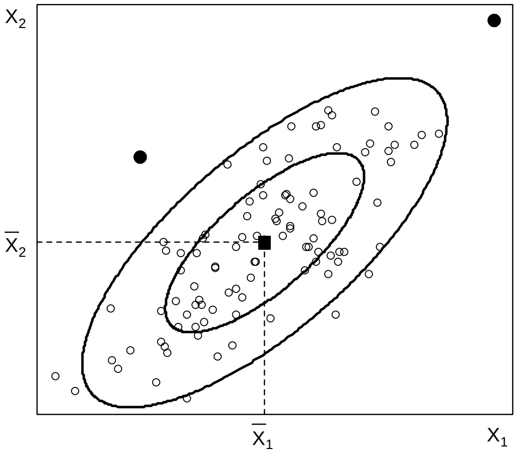

Figure 28. Contours of constant leverage in multiple regression with two explanatory variables, X_1 and X_2. The two observations marked with solid black dots have equal hat-values.

- Recall Duncan's data on the prestige, education, and income of 45 U.S. occupations in 1950. Here is the regression of prestige on income and education:

$$\begin{array}{rcccccccc}\widehat{\text{Prestige}} & = & -6.06 & + & 0.599 \times & \text{Income} & + & 0.546 \times & \text{Education} \\ & & (4.27) & & (0.120) & & & (0.098) & \end{array}$$

  - An index plot of hat-values for the observations in Duncan's regression is shown in Figure 29 (a), with a scatterplot for the explanatory variables in Figure 29 (b).

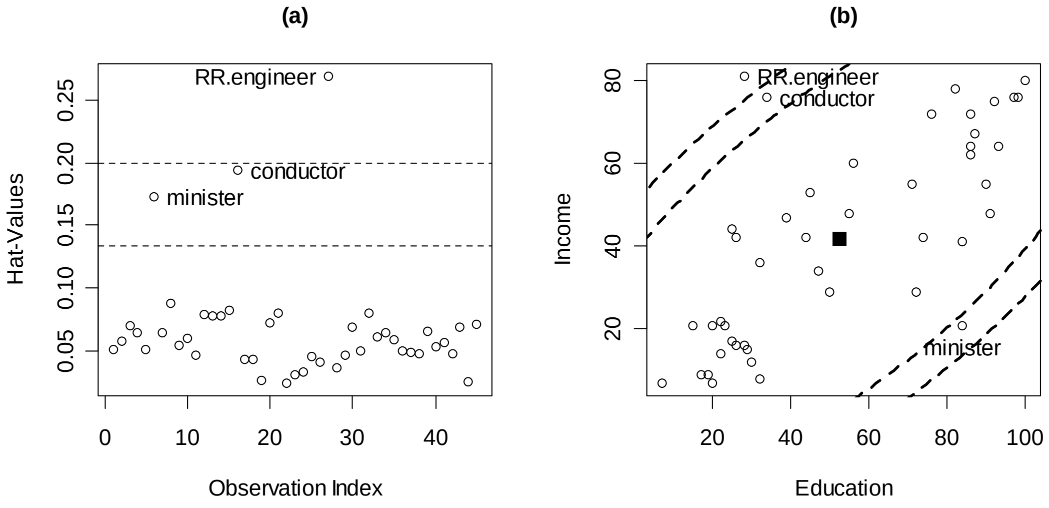

Figure 29. Duncan's occupational prestige regression: (a) hat-values; (b) scatterplot for education and income, showing contours of constant leverage at $2 \times \overline{h}$ and $3 \times \overline{h}$.

4.4 Detecting Outliers: Studentized Residuals

- Discrepant observations usually have large residuals, but even if the errors ε_i have equal variances (as assumed in the general linear model), the residuals E_i do not:

$$V(E_i) = \sigma_\varepsilon^2(1 - h_i)$$

  - High-leverage observations tend to have small residuals, because these observations can coerce the regression surface to be close to them.

- Although we can form a *standardized residual* by calculating

$$E_i' = \frac{E_i}{S_E\sqrt{1 - h_i}}$$

  this measure is slightly inconvenient because its numerator and denominator are not independent, preventing E_i' from following a t-distribution: When $|E_i|$ is large, $S_E = \sqrt{\sum E_i^2/(n - k - 1)}$, which contains E_i^2, tends to be large as well.

- Suppose that we refit the model deleting the ith observation, obtaining an estimate $S_{E(-i)}$ of σ_ε that is based on the remaining $n-1$ observations.
  - Then the *studentized residual*

$$E_i^* = \frac{E_i}{S_{E(-i)}\sqrt{1-h_i}}$$

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

  - An equivalent procedure for finding the studentized residuals employs a 'mean-shift' outlier model

$$Y = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k + \gamma D + \varepsilon$$

  where D is a dummy regressor set to one for observation i and zero for all other observations:

$$D = \begin{cases} 1 & \text{for obs. } i \\ 0 & \text{otherwise} \end{cases}$$

- Thus

$$\begin{aligned} E(Y_i) &= \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \gamma \\ E(Y_j) &= \alpha + \beta_1 X_{j1} + \cdots + \beta_k X_{jk} \text{ for } j \neq i \end{aligned}$$

  - It would be natural to specify this model if, before examining the data, we suspected that observation i differed from the others.
  - Then to test H_0: $\gamma = 0$, we can calculate $t_0 = \widehat{\gamma}/\text{SE}(\widehat{\gamma})$. This test statistic is distributed as t_{n-k-2} under H_0, and is the studentized residual E_i^*.

4.4.1 Testing for Outliers

- 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 $E_1^*, E_2^*, ..., E_n^*$. (It is not literally necessary to perform n auxiliary regressions.)
  - Usually, our interest then focuses on the largest absolute E_i^*, denoted $E_{\max}^*$.
  - 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 $E_{\max}^*$.
- One solution to this problem of simultaneous inference is to perform a *Bonferroni adjustment* to the p-value for the largest absolute E_i^*: Let $p' = \Pr(\, t_{n-k-2} > E_{\max}^*)$.
  - Then the Bonferroni p-value for testing the statistical significance of $E_{\max}^*$ is $p = 2np'$.

  - Note that a much larger $E_{\max}^*$ is required for a statistically significant result than would be the case for an ordinary individual t-test.
- Another approach is to construct a quantile-comparison plot for the studentized residuals, plotting against either the t or normal distribution.
- In Davis's regression of reported weight on measured weight, the largest studentized residual by far belongs to the incorrectly coded 12*th* observation, with $E_{12}^* = -24.3$.
  - Here, $n - k - 2 = 183 - 3 - 2 = 178$, and $\Pr(t_{178} > 24.3) \approx 10^{-58}$.
  - The Bonferroni p-value for the outlier test is $p \approx 2 \times 183 \times 10^{-58} = 4 \times 10^{-56}$, an unambiguous result.
- For Duncan's occupational prestige regression, the largest studentized residual belongs to *ministers*, with $E_{\text{minister}}^* = 3.135$.
  - The Bonferroni p-value is $2 \times 45 \times \Pr(t_{45-2-2} > 3.135) = .143$.

4.5 Measuring Influence

- Influence on the regression coefficients combines leverage and discrepancy.
- The most direct measure of influence simply expresses the impact on each coefficient of deleting each observation in turn:

  $$\text{dfbeta}_{ij} = B_j - B_{j(-i)} \text{ for } i = 1, ..., n \text{ and } j = 0, 1, ..., k$$

  where the B_j are the least-squares coefficients calculated for all of the data, and the $B_{j(-i)}$ are the least-squares coefficients calculated with the ith observation omitted. (So as not to complicate the notation here, I denote the least-squares intercept A as B_0.)
- One problem associated with using the dfbeta$_{ij}$ is their large number — $n(k+1)$.
  - It is useful to have a single summary index of the influence of each observation on the least-squares fit.

- Cook (1977) has proposed measuring the 'distance' between the B_j and the corresponding $B_{j(-i)}$ by calculating the F-statistic for the 'hypothesis' that $\beta_j = B_{j(-i)}$, for $j = 0, 1, ..., k$.
  - This statistic is recalculated for each observation $i = 1, ..., n$.
  - The resulting values should not literally be interpreted as F-tests, but rather as a distance measure that does not depend upon the scales of the X's.
  - Cook's statistic can be written (and simply calculated) as

    $$D_i = \frac{E_i'^2}{k+1} \times \frac{h_i}{1-h_i}$$

  - In effect, the first term in the formula for Cook's D is a measure of discrepancy, and the second 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 E_i^* against the h_i and to look for observations for which both are big. A slightly more sophisticated version of this plot that incorporates Cook's D is given below.
- For Davis's regression of reported weight on measured weight, Cook's D points to the obviously discrepant 12*th* observation:

  Cook's $D_{12} = 85.9$ (next largest, $D_{21} = 0.065$)
- For Duncan's regression, the largest Cook's D is for ministers, $D_6 = 0.566$.
  - Figure 30 displays a plot of studentized residuals versus hat-values, with the areas of the plotted circles proportional to values of Cook's D. The lines on the plot are at $E^* = \pm 2$ (on the vertical axis), and at $h = 2\overline{h}$ and $3\overline{h}$ (on the horizontal axis).
  - Four observations that exceed these cutoffs are identified on the plot.

  - Notice that *reporters* have a relatively large residual but are at a low-leverage point, while *railroad engineers* have high leverage but a small studentized residual.
- In developing the concept of influence in regression, I have focused on changes in regression coefficients. Other regression outputs, such as the coefficient sampling variances and covariances, are also subject to influence.

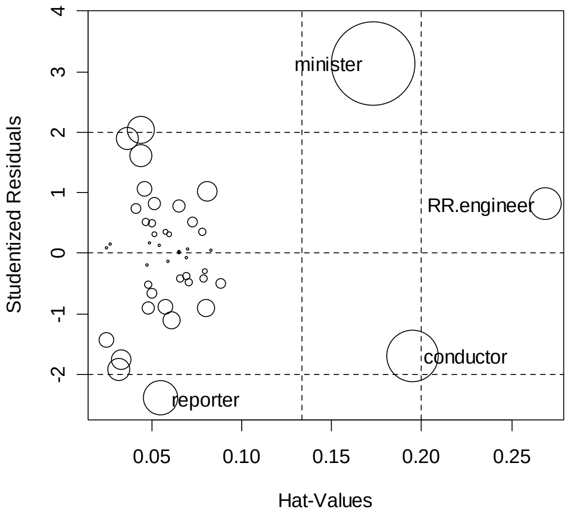

Figure 30. Influence plot for Duncan’s occupational prestige regression. The areas of the circles are proportional to Cook’s distance.

4.6 Numerical Cutoffs for Diagnostic Statistics

▶ I have refrained from suggesting specific numerical criteria for identifying noteworthy observations on the basis of measures of leverage and influence: I believe that it is generally more effective to examine the distributions of these quantities directly to locate unusual values.

- For studentized residuals, outlier-testing provides a numerical cutoff, but even this is no substitute for graphical examination of the residuals.

▶ Nevertheless, numerical cutoffs can be of some use, as long as they are not given too much weight, and especially when they are employed to enhance graphical displays.

- A line can be drawn on a graph at the value of a numerical cutoff, and observations that exceed the cutoff can be identified individually.

▶ Cutoffs for a diagnostic statistic may be derived from statistical theory, or they may result from examination of the sample distribution of the statistic.

- Cutoffs may be absolute, or they may be adjusted for sample size.
  - For some diagnostic statistics, such as measures of influence, absolute cutoffs are unlikely to identify noteworthy observations in large samples.
  - In part, this characteristic reflects the ability of large samples 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:

4.6.1 Hat-Values

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

4.6.2 Studentized Residuals

- Beyond the issue of ‘statistical significance,’ it sometimes helps to call attention to residuals that are relatively large.
- Under ideal conditions, about five percent of studentized residuals are outside the range $|E_i^*| \leq 2$. It is therefore reasonable to draw attention to observations outside this range.

4.6.3 Measures of Influence

- Many cutoffs have been suggested for different measures of influence, including the following size-adjusted cutoff for *Cook’s D,* due to Chatterjee and Hadi:

$$D_i > \frac{4}{n-k-1}$$

- Absolute cutoffs for D, such as $D_i > 1$, risk missing relatively influential data.

4.7 Joint Influence: Added-Variable Plots

- As illustrated in Figure 31, subsets of observations can be *jointly influential* or can offset each other's influence.
  - 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, but the sequential approach is not always successful.
- Although it is possible to generalize deletion statistics to subsets of several points, the very large number of subsets usually renders this approach impractical.
- An attractive alternative is to employ graphical methods, and a particularly useful influence graph is the *added-variable plot* (also called a *partial-regression plot* or an *partial-regression leverage plot*).

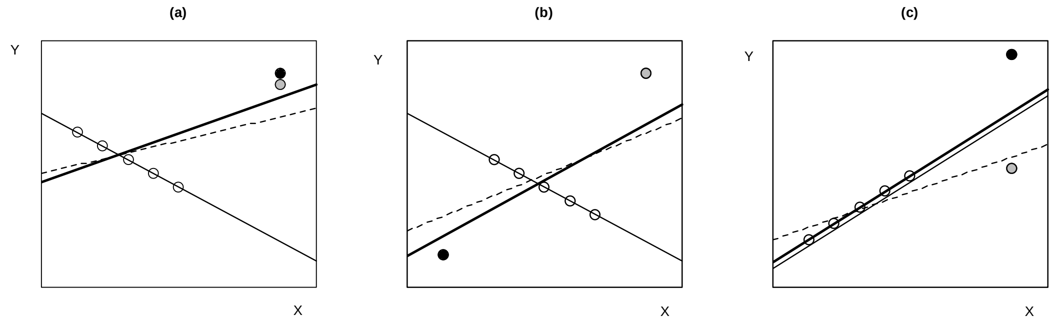

Figure 31. Jointly influential observations: (a) a pair of jointly influential points; (b) a widely separated jointly infuential pair; (c) two points that offset each other's influence. In each case the heavier solid line is the least-squares line for all of the data, the broken line deletes the black point, and the lighter solid line deletes both the gray and the black points.

- Let $Y_i^{(1)}$ represent the residuals from the least-squares regression of Y on all of the X's with the exception of X_1:
$$Y_i = A^{(1)} + B_2^{(1)} X_{i2} + \cdots + B_k^{(1)} X_{ik} + Y_i^{(1)}$$
- Likewise, $X_i^{(1)}$ are the residuals from the least-squares regression of X_1 on all the other X's:
$$X_{i1} = C^{(1)} + D_2^{(1)} X_{i2} + \cdots + D_k^{(1)} X_{ik} + X_i^{(1)}$$
- 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 effects of $X_2, ..., X_k$ are 'removed.'
- The residuals $Y^{(1)}$ and $X^{(1)}$ have the following interesting properties:

1. The slope from the least-squares regression of $Y^{(1)}$ on $X^{(1)}$ is simply the least-squares slope B_1 from the full multiple regression.

2. The residuals from the simple regression of $Y^{(1)}$ on $X^{(1)}$ are the same as those from the full regression:
$$Y_i^{(1)} = B_1 X_i^{(1)} + E_i$$
No constant is required, because both $Y^{(1)}$ and $X^{(1)}$ have means of 0.
3. The variation of $X^{(1)}$ is the conditional variation of X_1 holding the other X's constant and, as a consequence, the standard error of B_1 in the auxiliary simple regression
$$\text{SE}(B_1) = \frac{S_E}{\sqrt{\sum X_i^{(1)^2}}}$$
is (except for df) the multiple-regression standard error of B_1. Unless X_1 is uncorrelated with the other X's, its conditional variation is smaller than its marginal variation — much smaller, if X_1 is strongly collinear with the other X's.

- Plotting $Y^{(1)}$ against $X^{(1)}$ permits us to examine leverage and influence on B_1. Because of properties 1–3, this plot also provides a visual impression of the precision of estimation of B_1.
- Similar added-variable plots can be constructed for the other regression coefficients:

$$\text{Plot } Y^{(j)} \text{ versus } X^{(j)} \text{ for each } j = 0, ..., k$$

▶ Illustrative added-variable plots are shown in Figure 32, using data from Duncan's regression of occupational prestige on the income and educational levels of 45 U.S. occupations:

$$\begin{array}{lcccccc} \widehat{\text{Prestige}} & = & -6.06 & + & 0.599 \times \text{ Income} & + & 0.546 \times \text{ Education} \\ & & (4.27) & & (0.120) & & (0.098) \end{array}$$

$$R^2 = 0.83 \quad S_E = 13.4$$

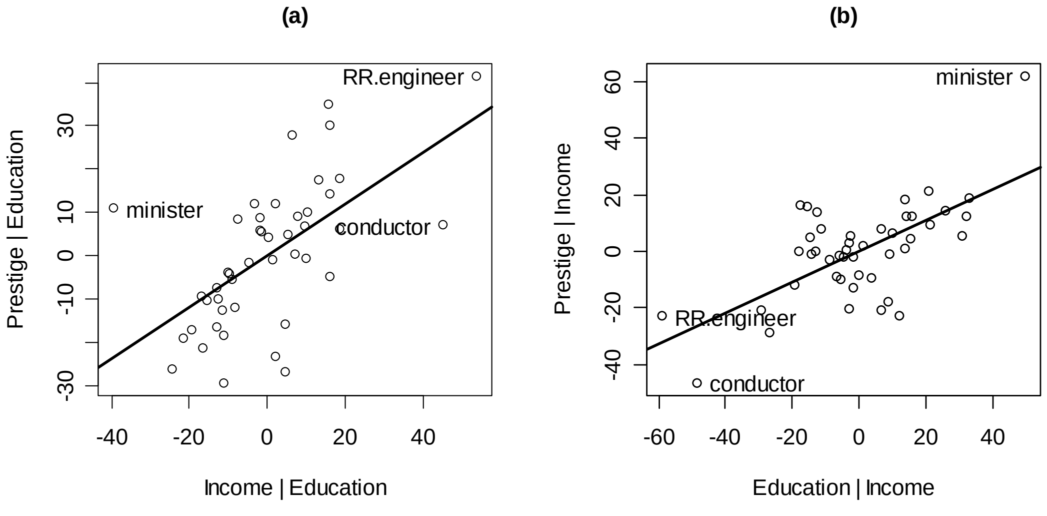

Figure 32. Added-variable plots for Duncan's occupational prestige regression, (a) for income, and (b) for education.

- The added-variable plot for income (a) reveals three unusual data points:
  - *ministers*, whose income is unusually low given the educational level of the occupation; and
  - *railroad conductors* and *railroad engineers*, whose incomes are unusually high given education.
  - Together, *ministers* and *railroad conductors* reduce the income slope; *railroad engineers*, while a high-leverage point, are more in line with the rest of the data.
  - Remember that the horizontal variable in this added-variable plot is the residual from the regression of income on education, and thus values far from 0 in this direction are for occupations with incomes that are unusually high or low given their levels of education.

- The added-variable plot for education (b) shows that the same three observations have relatively high leverage on the education coefficient:
  - *ministers* and *railroad conductors* tend to increase the education slope;
  - *railroad engineers* appear to be closer in line with the rest of the data.

▶ Deleting *ministers* and *conductors* produces the fitted regression

$$\begin{array}{rcccccc}\widehat{\text{Prestige}} & = & -6.41 & + & 0.867 \times \text{ Income} & + & 0.332 \times \text{ Education} \\ & & (3.65) & & (0.122) & & (0.099)\end{array}$$

$$R^2 = 0.88 \quad S_E = 11.4$$

which has a larger income slope and smaller education slope than the original regression.

- The estimated standard errors are likely optimistic, because relative outliers have been trimmed away.

- Deleting *railroad engineers*, along with *ministers* and *conductors*, further increases the income slope and decreases the education slope, but the change is not dramatic: $B_{\text{Income}} = 0.931, B_{\text{Education}} = 0.285$.

4.8 Should Unusual Data Be Discarded?

▶ Although problematic data should not be ignored, they also should not be deleted automatically and without reflection:

▶ It is important to investigate *why* an observation is unusual.

- Truly bad data (e.g., as in Davis's regression) can be corrected or thrown away.
- When a discrepant data-point is correct, we may be able to understand why the observation is unusual.
  - For Duncan's regression, for example, it makes sense that ministers enjoy prestige not accounted for by the income and educational levels of the occupation.
  - In a case like this, we may choose to deal separately with an outlying observation.

- Outliers or influential data may motivate model respecification.
  - For example, the pattern of outlying data may suggest the introduction of additional explanatory variables.
    - If, in Duncan's regression, we can identify a variable that produces the unusually high prestige of ministers (net of their income and education), and if we can measure that variable for other observations, then the variable could be added to the regression.
  - 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.
  - We must, however, be careful to avoid 'over-fitting' the data — permitting a small portion of the data to determine the form of the model.

- Except in clear-cut cases, we are justifiably reluctant to delete observations or to respecify the model to accommodate unusual data.
  - Some researchers reasonably adopt alternative estimation strategies, such as robust regression, which continuously downweights outlying data rather than simply including or discarding them.
  - Because these methods assign zero or very small weight to highly discrepant data, however, the result is generally not very different from careful application of least squares, and, indeed, robust-regression weights can be used to identify outliers.

4.9 Summary:Unusual and Influential Data

- Unusual data are problematic in linear models fit by least squares because they can substantially influence the results of the analysis, and because they may indicate that the model fails to capture important features of the data.
- Observations with unusual combinations of explanatory-variables values have high *leverage* in a least-squares regression. The hat-values h_i provide a measure of leverage. A rough cutoff for noteworthy hat-values is $h_i > 2\overline{h} = 2(k+1)/n$.

- A regression *outlier* is an observation with an unusual response-variable value given its combination of explanatory-variable values. The studentized residuals E_i^* can be used to identify outliers, through graphical examination or a Bonferroni test for the largest absolute E_i^*. If the model is correct (and there are no true outliers), then each studentized residual follows a t-distribution with $n-k-2$ degrees of freedom.
- Observations that combine high leverage with a large studentized residual exert substantial *influence* on the regression coefficients. Cook's D-statistic provides a summary index of influence on the coefficients. A rough cutoff is $D_i > 4/(n-k-1)$.
- Subsets of observations can be jointly influential. Added-variable plots are useful for detecting joint influence on the regression coefficients. The added-variable plot for the regressor X_j is formed using the residuals from the least-squares regressions of X_j and Y on all of the other X's.

- Outlying and influential data should not be ignored, but they also should not simply be deleted without investigation. ‘Bad’ data can often be corrected. ‘Good’ observations that are unusual may provide insight into the structure of the data, and may motivate respecification of the statistical model used to summarize the data.

5. Diagnosing Nonlinearity and Other Ills

5.1 Goals

- To introduce simple methods for detecting non-normality, non-constant error variance, and nonlinearity.
- To show how these problems can often be corrected by transformation and other approaches.
- To demonstrate the application of the method of maximum likelihood to regression diagnostics.

5.2 Example: The SLID Data

- To illustrate the methods described here, I will primarily use data from the 1994 wave of Statistics Canada's Survey of Labour and Income Dynamics (SLID).
- The SLID data set that I use includes 3997 employed individuals who were between 16 and 65 years of age and who resided in Ontario.
- Regressing the composite hourly wage rate on a dummy variable for sex (code 1 for males), education (in years), and age (also in years) produces the following results:

$$
\begin{aligned}
\widehat{\text{Wages}} &= \underset{(0.599)}{-8.124} + \underset{(0.2070)}{3.474} \times \text{Male} \qquad + \underset{(0.0087)}{0.2613} \times \text{Age} \\
&\qquad\qquad + \underset{(0.0343)}{0.9296} \times \text{Education} \\
R^2 &= .3074
\end{aligned}
$$

5.3 Non-Normally Distributed Errors

- The assumption of normally distributed errors is almost always arbitrary, but the central-limit theorem assures that inference based on the least-squares estimator is approximately valid. Why should we be concerned about non-normal errors?
  - Although the *validity* of least-squares estimation is robust the *efficiency* of least squares is not: The least-squares estimator is maximally efficient among unbiased estimators when the errors are normal. For heavy-tailed errors, the efficiency of least-squares estimation decreases markedly.
  - Highly skewed error distributions, aside from their propensity to generate outliers in the direction of the skew, compromise the interpretation of the least-squares fit as a conditional typical value of Y.

- A multimodal error distribution suggests the omission of one or more discrete explanatory variables that divide the data naturally into groups.

▶ Quantile-comparison plots are useful for examining the distribution of the residuals, which are estimates of the errors.

- We compare the sample distribution of the studentized residuals, E_i^*, with the quantiles of the unit-normal distribution, $N(0, 1)$, or with those of the t-distribution for $n - k - 2$ degrees of freedom.
- Even if the model is correct, the studentized residuals are not an *independent* random sample from t_{n-k-2}. Correlations among the residuals depend upon the configuration of the X-values, but they are generally negligible unless the sample size is small.
- At the cost of some computation, it is possible to adjust for the dependencies among the residuals in interpreting a quantile-comparison plot.

▶ The quantile-comparison plot is effective in displaying the tail behavior of the residuals: Outliers, skewness, heavy tails, or light tails all show up clearly.

▶ Other univariate graphical displays, such as histograms and density estimates, effectively supplement the quantile-comparison plot.

▶ Figure 33 shows a t quantile-comparison plot and a density estimate for the studentized residuals from the SLID regression.

- The distribution of the studentized residuals is positively skewed and there may be more than one mode.
- The positive skew in the residual distribution can be corrected by transforming the *response variable* down the ladder of powers, in this case using logs, producing the residual distribution shown in Figure 34.
  - · The resulting residual distribution has a slight negative skew, but I preferred the log transformation to the 1/3 power for interpretability.

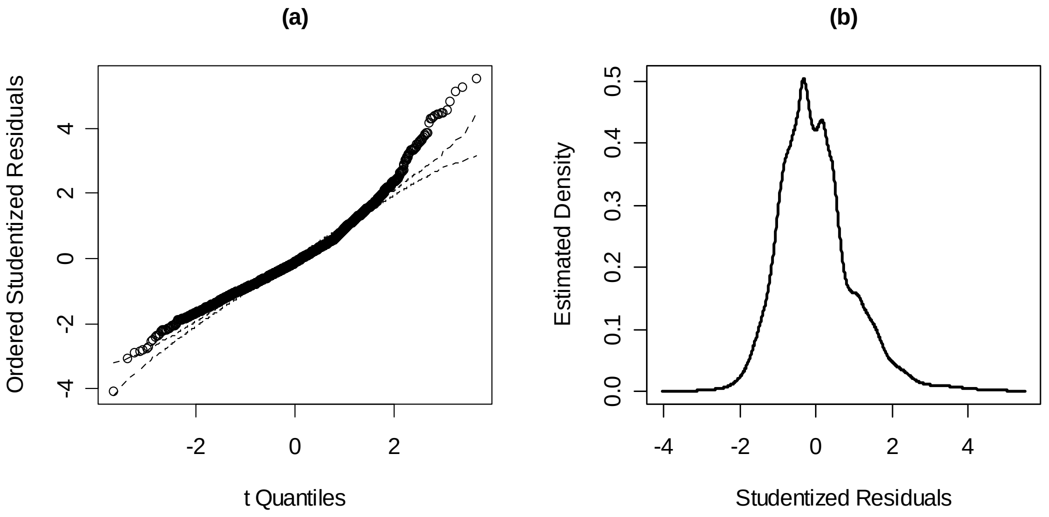

Figure 33. (a) Quantile-comparison plot with point-wise 95-percent simulated confidence envelope and (b) adaptive kernel-density estimate for the studentized residuals from the SLID regression.

- Note that the residual distribution is heavy-tailed and possibly bimodal.

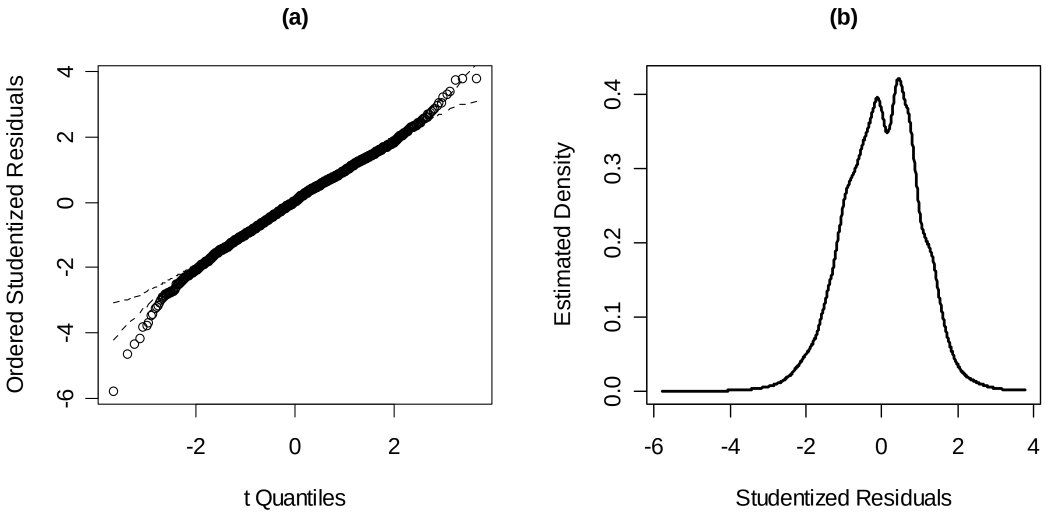

Figure 34. (a) Quantile-comparison plot, and (b) adaptive kernel-density estimate for the studentized residuals from the SLID regression with wages log-transformed.

5.4 Non-Constant Error Variance

▶ Although the least-squares estimator is unbiased and consistent even when the error variance is not constant, its efficiency is impaired, and the usual formulas for coefficient standard errors are inaccurate.

- Non-constant error variance is sometimes termed ‘heteroscedasticity.’

▶ Because the regression surface is k-dimensional, and imbedded in a space of $k+1$ dimensions, it is generally impractical to assess the assumption of constant error variance by direct graphical examination of the data.

▶ 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.

- Plotting residuals against Y (as opposed to $\widehat{Y}$) is generally unsatisfactory, because the plot will be 'tilted'
  - · There is a built-in linear correlation between Y and E, since $Y = \widehat{Y} + E$.
  - · The least-squares fit insures that the correlation between $\widehat{Y}$ and E is zero, producing a plot that is much easier to examine for evidence of non-constant spread.
- 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 $\widehat{Y}$.
- 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.

▶ Figure 35 shows a plot of studentized residuals against fitted values and a spread-level plot for the SLID regression.

- The increasing spread with increasing $\widehat{Y}$ suggests moving Y down the ladder of powers to stabilize the variance.
- The slope of the line in the spread-level plot is $b = 0.9994$, suggesting the transformation $p = 1 - 0.9994 = 0.0006 \approx 0$ (i.e., the log transformation).
- After log-transforming Y, the diagnostic plots look much better (Figure 36).

▶ There are alternatives to transformation for dealing with non-constant error variance.

- Weighted-least-squares (WLS) regression, for example, can be used, down-weighting observations that have high variance.
- It is also possible to correct the estimated standard errors of the ordinary least squares (OLS) estimates for non-constant spread.

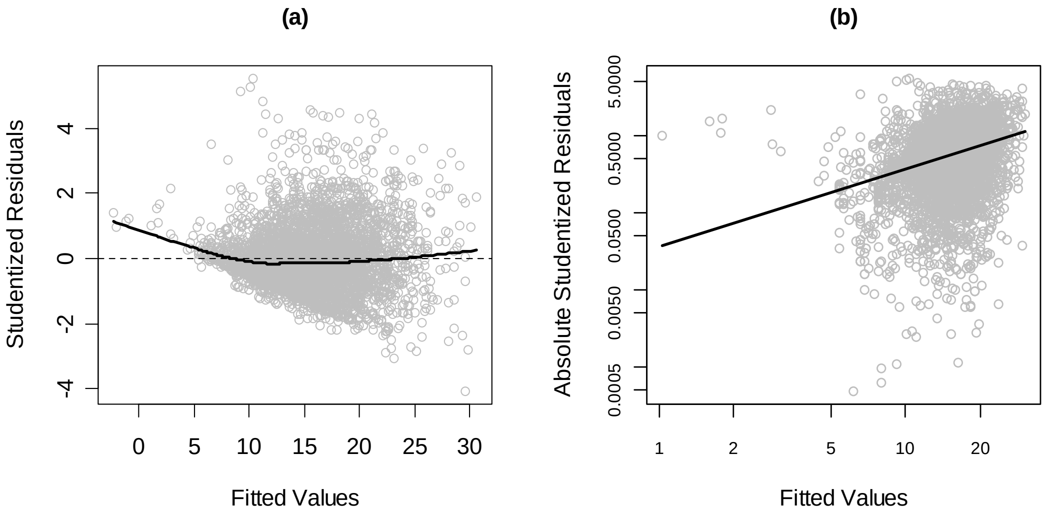

Figure 35. (a) Studentized residuals vs. fitted values, and (b) spread-level plot for the SLID regression. A few observations with $\widehat{Y} \leq 0$ were removed from (b), and the line is fit by robust regression.

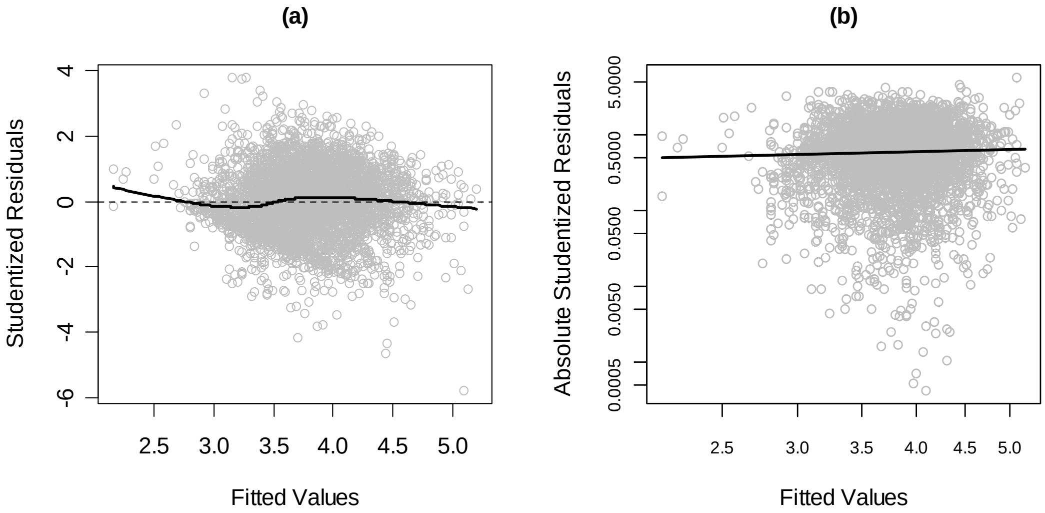

Figure 36. (a) Studentized residuals versus fitted values, and (b) spread--level plot for the SLID regression after log-transforming wages.

- ▶ Non-constant error variance is a serious problem only when it is relatively extreme — say when the magnitude (i.e., the standard deviation) of the errors varies by more than a factor of about three — that is, when the largest error variance is more than about 10 times the smallest (although there are cases where this simple rule fails to offer sufficient protection).

5.5 Nonlinearity

- ▶ The assumption that the average error, $E(\varepsilon)$, is everywhere zero implies that the specified regression surface accurately reflects the dependency of Y on the X's.
  - The term 'nonlinearity' is therefore not used in the narrow sense here, although it includes the possibility that a partial relationship assumed to be linear is in fact nonlinear.
  - If, for example, two explanatory variables specified to have additive effects instead interact, then the average error is not zero for all combinations of X-values.
  - If nonlinearity, in the broad sense, is slight, then the fitted model can be a useful approximation even though the regression surface $E(Y|X_1, ...X_k)$ is not captured precisely.
  - In other instances, however, the model can be seriously misleading.

- The regression surface is generally high dimensional, even after accounting for regressors (such as dummy variables, interactions, and polynomial terms) that are functions of a smaller number of fundamental explanatory variables.
  - As in the case of non-constant error variance, it is necessary to focus on particular patterns of departure from linearity.
  - The graphical diagnostics discussed in this section are two-dimensional projections of the $(k+1)$-dimensional point-cloud of observations $\{Y_i, X_{i1}, ..., X_{ik}\}$.

5.5.1 Component+Residual Plots

- Although it is useful in multiple regression to plot Y against each X, these plots 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 an individual X, ignoring the other X's.
- Plotting residuals or studentized residuals against each X is frequently helpful for detecting departures from linearity.
  - As Figure 37 illustrates, however, residual plots cannot distinguish between monotone and non-monotone nonlinearity.
    - The distinction is important because monotone nonlinearity frequently can be 'corrected' by simple transformations.
    - Case (a) might be modeled by $Y = \alpha + \beta\sqrt{X} + \varepsilon$.

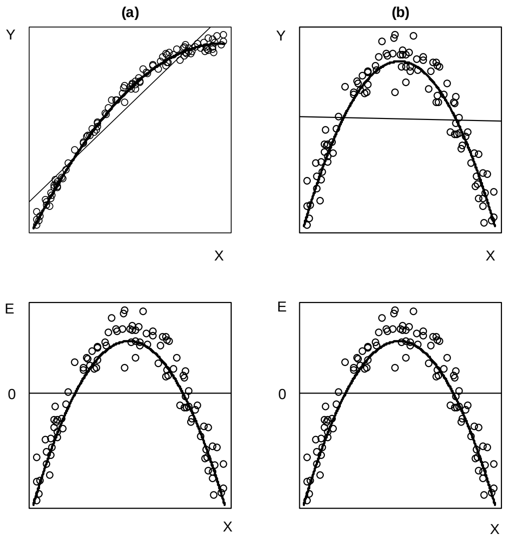

Figure 37. The residual plots of E versus X (bottom) are identical, even though the regression of Y on X in (a) is monotone while that in (b) is non-monotone.

· Case (b) cannot be linearized by a power transformation of X, and might instead be dealt with by the quadratic regression, $Y = \alpha + \beta_1 X + \beta_2 X^2 + \varepsilon$.

- ▶ Added-variable plots, introduced previously for detecting influential data, can reveal nonlinearity and suggest whether a relationship is monotone.
  - These plots are not always useful for locating a transformation, however: The added-variable plot adjusts X_j for the other X's, but it is the unadjusted X_j that is transformed in respecifying the model.
- ▶ *Component+residual plots*, also called *partial-residual plots* (as opposed to partial-regression = added-variable plots) are often an effective alternative.
  - Component+residual plots are not as suitable as added-variable plots for revealing leverage and influence.
  - The partial residual for the jth explanatory variable is

$$E_i^{(j)} = E_i + B_j X_{ij}$$

- In words, add back the linear component of the partial relationship between Y and X_j to the least-squares residuals, which may include an unmodeled nonlinear component.
- Then plot $E^{(j)}$ versus X_j.
- By construction, the multiple-regression coefficient B_j is the slope of the simple linear regression of $E^{(j)}$ on X_j, but nonlinearity may be apparent in the plot as well.

▶ The component+residual plots in Figure 38 are for age and education in the SLID regression, using log-wages as the response.

- Both plots look nonlinear:
  - · It is not entirely clear whether the partial relationship of log wages to age is monotone, simply tending to level off at the higher ages, or whether it is non-monotone, turning back down at the far right.

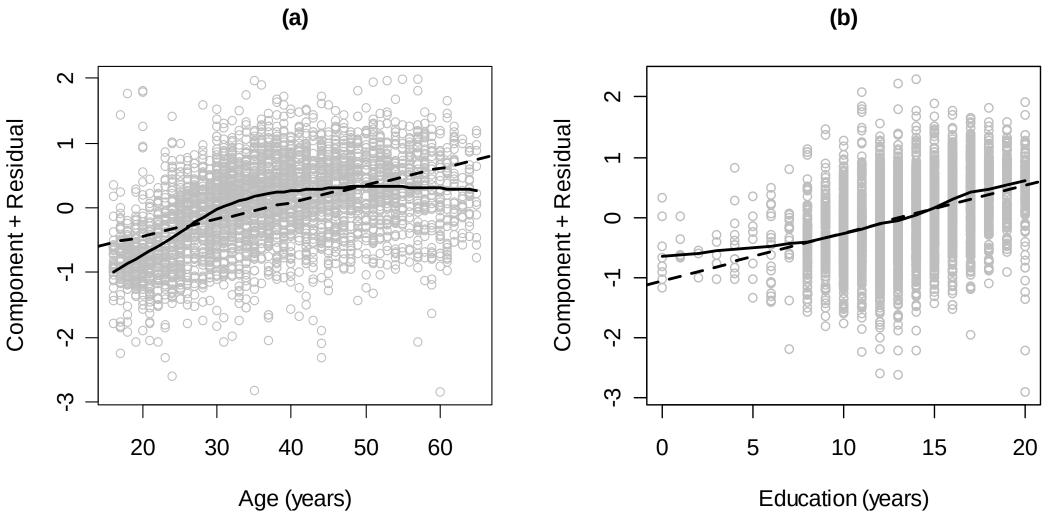

Figure 38. Component-plus-residual plots for age and education in the SLID regression of log wages on these variables and sex. A lowess smooth (span = 0.4) and least-squares line is shown on each graph.

- The partial relationship of log wages to education is clearly monotone, and the departure from linearity is not great—except at the lowest levels of education, where data are sparse; we should be able to linearize this partial relationship by moving education *up* the ladder of powers, because the bulge points to the right.
- Trial and error experimentation suggests that the quadratic specification for age works better, producing the following fit to the data:

$$\begin{aligned}
\widehat{\log_2 \text{Wages}} &= \underset{(0.0834)}{0.5725} + \underset{(0.0180)}{0.3195} \times \text{Male} + \underset{(0.0046)}{0.1198} \times \text{Age} \\
&\quad - \underset{(0.000059)}{0.001230} \times \text{Age}^2 + \underset{(0.000113)}{0.002605} \times \text{Education}^2 \\
R^2 &= .3892
\end{aligned}$$

- We can take two approaches to constructing component+residual plots for this respecified model:

1. We can plot partial residuals for each of age and education against the corresponding explanatory variable. In the case of age, the partial residuals are computed as
$$E_i^{(\text{Age})} = 0.1198 \times \text{Age}_i - 0.001230 \times \text{Age}_i^2 + E_i$$
and for education,
$$E_i^{(\text{Education})} = 0.002605 \times \text{Education}_i^2 + E_i$$
See the upper panels of Figure 39; the solid lines are the *partial fits* (i.e., the components) for the two explanatory variables,
$$\widehat{Y}_i^{(\text{Age})} = 0.1198 \times \text{Age}_i - 0.001230 \times \text{Age}_i^2$$
$$\widehat{Y}_i^{(\text{Education})} = 0.002605 \times \text{Education}_i^2$$
2. We can plot the partial residuals against the partial fits. See the two lower panels of Figure 39.

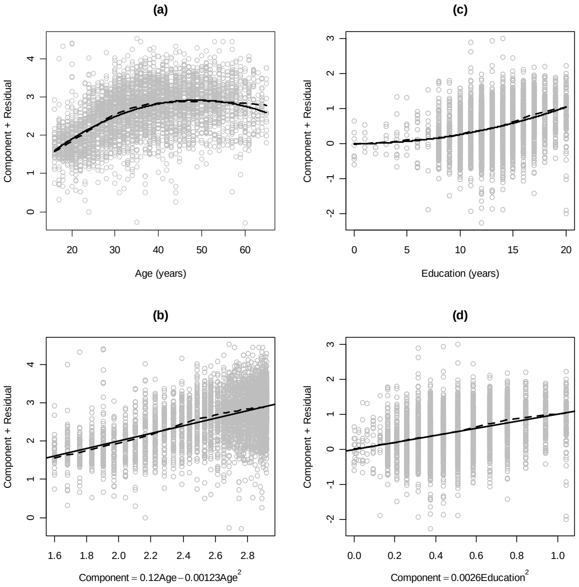

Figure 39. Component-plus-residual plots for age [panels (a) and (b)] and education [panels (c) and (d)] in the respecified model fit to the SLID data.

▶ Interpretation of the respecified SLID regression model is complicated by the transformation of the response ($\log_2$wages), the transformation of education, and the use of a quadratic for age.

- The coefficient of the dummy variable for sex, 0.3195, implies that at fixed levels of age and education, men on average earn $2^{0.3195} = 1.25$ times (i.e., 25 percent more) than women.
- ‘Effect displays’ for the partial relationship of wages to age and education are shown in Figure 40. Each effect plot is obtained by setting the other variable (e.g., education in the case of age) to its mean, while the focal variable (e.g., age) takes on its range of values in the data, computing the fitted value of the response under the model for each value of the focal variable.

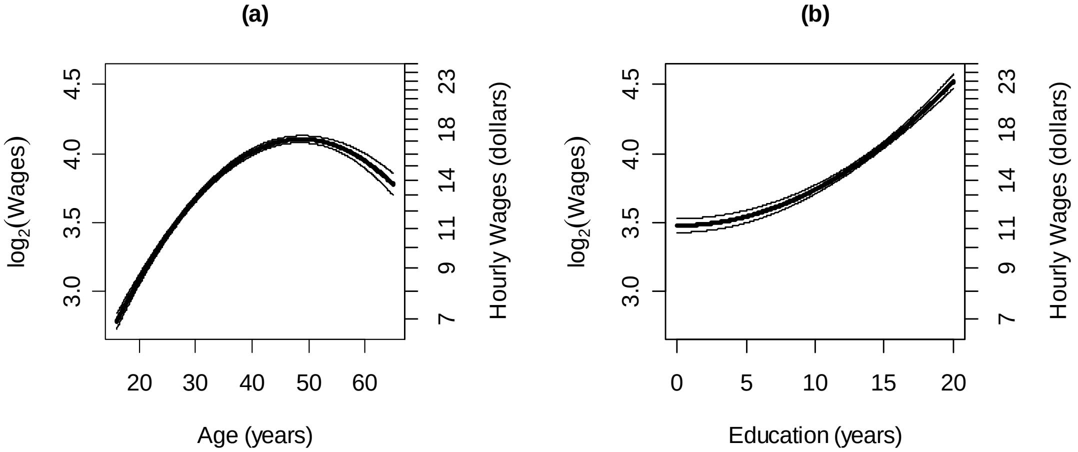

Figure 40. Effect displays for age and education in the regression of log wages on a quadratic in age, the square of education, and sex. The lighter lines give 95-percent point-wise confidence envelopes around the fits.

5.5.2 When Do Component+Residual Plots Work?

- Imagine that the following model accurately describes the data:

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

  - That is, the partial relationship between Y and X_1 is (potentially) nonlinear, characterized by the function $f(X_1)$, while the other explanatory variables, $X_2, ..., X_k$ enter the model linearly.

- Instead of fitting this model to the data, we fit the 'working model'

$$Y_i = \alpha' + \beta'_1 X_{i1} + \beta'_2 X_{i2} + \cdots + \beta'_k X_{ik} + \varepsilon'_i$$

and construct a component+residual plot for the working model.

- The partial residuals estimate

$$\varepsilon_i^{(1)} = \beta'_1 X_{i1} + \varepsilon'_i$$

  - What we would really like to estimate, however, is $f(X_{i1}) + \varepsilon_i$, which, apart from random error, will tell us the partial relationship between Y and X_1.

- Cook (1993) shows that $\varepsilon_i^{(1)} = f(X_{i1}) + \varepsilon_i$, as desired, under either of two circumstances:

- The function $f(X_1)$ is linear.
- The *other* explanatory variables $X_2, ..., X_k$ are each linearly related to X_1. That is,

$$E(X_{ij}) = \alpha_{j1} + \beta_{j1} X_{i1} \text{ for } j = 2, ..., k$$

▶ If there are *nonlinear* relationships between other X's and X_1, then the component+residual plot for X_1 may appear nonlinear even if the true partial regression is linear.

▶ The second result suggests a practical procedure for improving the chances that component+residual plots will provide accurate evidence of nonlinearity:

- If possible, transform the explanatory variables to linearize the relationships among them.

▶ Evidence suggests that weak nonlinearity is not especially problematic, but strong nonlinear relationships among the explanatory variables can invalidate the component+residual plot as a useful diagnostic display.

- There are more sophisticated versions of component+residual plots that are more robust.

5.6 Discrete Data

▶ Discrete explanatory and response variables often lead to plots that are difficult to interpret, a problem that can be rectified by ‘jittering’ the plotted points.

- A discrete *response* variable also violates the assumption that the errors in a linear model are normally distributed.
- Discrete *explanatory* variables, in contrast, are perfectly consistent with the general linear model, which makes no distributional assumptions about the X’s, other than independence between the X’s and the errors.
- 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.

5.6.1 Testing for Nonlinearity (‘Lack of Fit’)

▶ Recall the data on vocabulary and education collected in the U.S. General Social Survey. Years of education in this dataset range between 0 and 20 (see Figure 41). We model the relationship between vocabulary score and education in two ways:

1. Fit a linear regression of vocabulary on education:
$$Y_i = \alpha + \beta X_i + \varepsilon_i \qquad \text{(Model 1)}$$
2. Model education with a set of 20 dummy regressors (treating 0 years as the baseline category):
$$Y_i = \alpha' + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \cdots + \gamma_{20} D_{i,20} + \varepsilon_i' \qquad \text{(Model 2)}$$

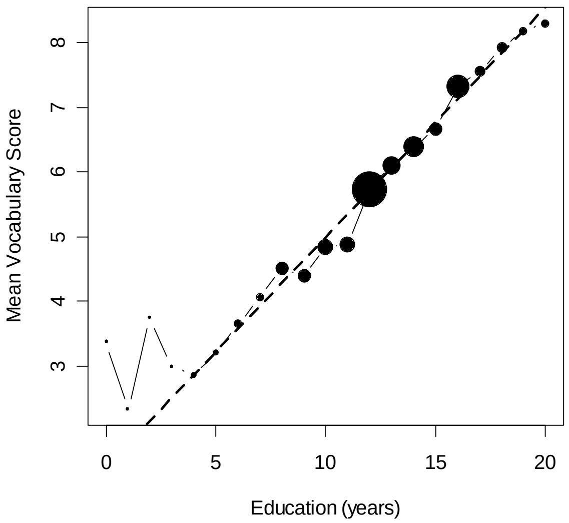

Figure 41. Mean vocabulary score by years of education. The size of the points is proportional to the number of observations. The broken line is the least-squares line.

- Contrasting the two models produces a test for nonlinearity, because the first model, specifying a linear relationship between vocabulary and education, is a special case of the second, which can capture *any* pattern of relationship between $E(Y)$ and X.
  - The resulting incremental F-test for nonlinearity appears in the following ANOVA table:

Source	*SS*	*df*	*F*	*p*
Education (*Model 2*)	26,099	20	374.44	$\ll .0001$
Linear (*Model 1*)	25,340	1	7,270.99	$\ll .0001$
Nonlinear (*"lack of fit"*)	759	19	11.46	$\ll .0001$
Error (*"pure error"*)	75,337	21,617		
Total	101,436	21,637		

- Note that while it is highly statistically significant, the nonlinear component accounts for very little of the variation in vocabulary scores.

▶ The incremental F-test for nonlinearity can easily be extended to a discrete explanatory variable — say X_1 — in a multiple-regression model.

- Here, we need to contrast the 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 $D_1, ..., D_{m-1}$ are dummy regressors constructed to represent the m categories of X_1.

5.6.2 Testing for Non-Constant Error Variance

▶ A discrete X (or combination of X's) partitions the data into m groups (as in analysis of variance).

- Let Y_{ij} denote the ith of n_j response-variable scores in group j.
- If the error variance is constant across groups, then the within-group sample variances

$$S_j^2 = \frac{\sum_{i=1}^{n_j} (Y_{ij} - \overline{Y}_j)^2}{n_j - 1}$$

should be similar.

  · Tests that examine the S_j^2 directly do not maintain their validity well when the distribution of the errors is non-normal.

▶ The following simple F-test (called Levene's test) is both robust and powerful:

- Calculate the values
$$Z_{ij} \equiv |Y_{ij} - \widetilde{Y}_j|$$
where $\widetilde{Y}_j$ is the median response-variable value in group j.
- Then perform a one-way analysis-of-variance of the Z_{ij} over the m groups.
- If the error variance is not constant across the groups, then the group means $\overline{Z}_j$ will tend to differ, producing a large value of the F-test statistic.

- For the vocabulary data, where education partitions the $21,638$ observations into $m = 21$ groups, $F_0 = 4.26$, with 20 and $21,617$ degrees of freedom, for which $p \ll .0001$. There is, therefore, strong evidence of non-constant spread in vocabulary across the categories of education, though, as revealed in Figure 42, the within-group standard deviations are not very different (discounting the small numbers of individuals with very low levels of education).

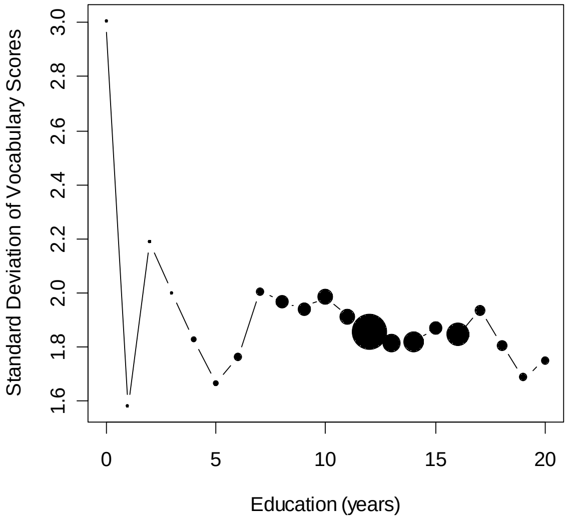

Figure 42. Standard deviation of vocabulary scores by education. The relative size of the points is proportional to the number of observations.

5.7 Maximum-Likelihood Methods

- A statistically sophisticated approach to selecting a transformation of Y or an X is to imbed the linear model in a more general nonlinear model that contains a parameter for the transformation.
  - If several variables are potentially to be transformed then there may be several such parameters.
- Suppose that the transformation is indexed by a single parameter λ, and that we can write down the likelihood for the model as a function of the transformation parameter and the usual regression parameters: $L(\lambda, \alpha, \beta_1, ..., \beta_k, \sigma_\varepsilon^2)$.
  - Maximizing the likelihood yields the maximum-likelihood estimate of λ along with the MLEs of the other parameters.
  - Now suppose that $\lambda = \lambda_0$ represents *no* transformation (e.g., $\lambda_0 = 1$ for the power transformation Y^λ).

- A likelihood-ratio test, Wald test, or score test of H_0: $\lambda = \lambda_0$ assesses the evidence that a transformation is required.
- A disadvantage of the likelihood-ratio and Wald tests is that they require finding the MLE, which usually requires iteration.
  - In contrast, the slope of the log-likelihood at λ_0 — on which the score test depends — generally can be assessed or approximated without iteration.
  - Often, the score test can be formulated as the t-statistic for a new regressor, called a *constructed variable*, to be added to the linear model.
  - Moreover, an added-variable plot for the constructed variable then can reveal whether one or a small group of observations is unduly influential in determining the transformation.

5.7.1 Box-Cox Transformation of Y

- Box and Cox suggest power transformation of Y with the object of normalizing the error distribution.
- The general Box-Cox model is

$$Y_i^{(\lambda)} = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

where the errors ε_i are independently $N(0, \sigma_\varepsilon^2)$, and

$$Y_i^{(\lambda)} = \begin{cases} \dfrac{Y_i^\lambda - 1}{\lambda} & \text{for } \lambda \neq 0 \\ \log_e Y_i & \text{for } \lambda = 0 \end{cases}$$

  - Note that all of the Y_i must be positive.

- ▶ A simple procedure for finding the MLE is to evaluate the maximized $\log_e L(\alpha, \beta_1, ..., \beta_k, \sigma_\varepsilon^2 | \lambda)$, called the *profile log-likelihood*, for a range of values of λ, say between -2 and $+2$.
  - If this range turns out not to contain the maximum of the log-likelihood, then the range can be expanded.
  - To test H_0: $\lambda = 1$, calculate the likelihood-ratio statistic
    $$G_0^2 = -2[\log_e L(\lambda = 1) - \log_e L(\lambda = \widehat{\lambda})]$$
    which is asymptotically distributed as χ^2 with one degree of freedom under H_0.
  - Equivalently, a 95-percent confidence interval for λ includes those values for which
    $$\log_e L(\lambda) > \log_e L(\lambda = \widehat{\lambda}) - 1.92$$
    - · The figure 1.92 comes from $1/2 \times \chi_{1,.05}^2 = 1/2 \times 1.96^2$.
- ▶ Figure 43 shows a plot of the profile log-likelihood against λ for the original SLID regression.

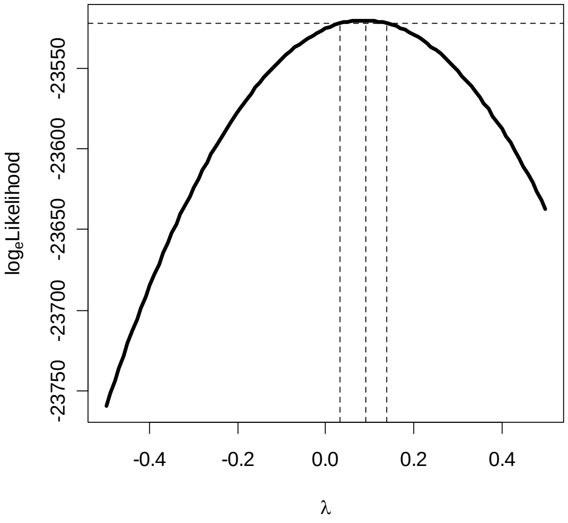

Figure 43. Box-Cox transformations for the SLID regression of wages on sex, age, and education. The profile log-likelihood is plotted against the transformation parameter λ.

- The maximum-likelihood estimate of λ is $\widehat{\lambda} = 0.09$, and a 95% confidence interval, marked out by the intersection of the line near the top of the graph with the profile log-likelihood, runs from 0.04 to 0.13.

▶ Atkinson has proposed an approximate score test for the Box-Cox model, based on the constructed variable

$$G_i = Y_i \left[\log_e \left(\frac{Y_i}{\widetilde{Y}} \right) - 1 \right]$$

where $\widetilde{Y}$ is the *geometric mean* of Y:

$$\widetilde{Y} \equiv (Y_1 \times Y_2 \times \cdots \times Y_n)^{\frac{1}{n}}$$

- The augmented regression, including the constructed variable, is then

$$Y_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \phi G_i + \varepsilon_i$$

- The t-test of H_0: $\phi = 0$, that is, $t_0 = \widehat{\phi}/\text{SE}(\widehat{\phi})$, assesses the need for a transformation.

- An estimate of λ (though not the MLE) is given by $\widetilde{\lambda} = 1 - \widehat{\phi}$.
- The added-variable plot for the constructed variable G shows influence and leverage on $\widehat{\phi}$, and hence on the choice of λ.
- Atkinson's constructed-variable plot for the interlocking-directorate regression is shown in Figure 44.
  - · The coefficient of the constructed variable in the regression is $\widehat{\phi} = 1.454$, with SE$(\widehat{\phi}) = 0.026$, providing overwhelmingly strong evidence of the need to transform Y.
  - · The suggested transformation, $\widetilde{\lambda} = 1 - 1.454 = -0.454$, is far from the MLE.

▶ The Box-Cox method can also be applied to the marginal distribution of a variable, or multivariately, to the joint distribution of several variables.

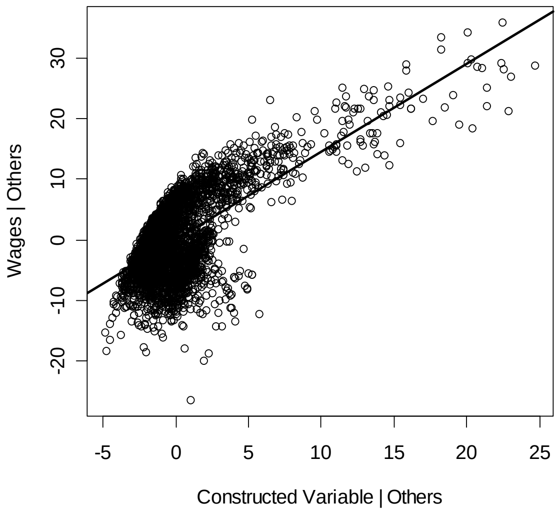

Figure 44. Constructed-variable plot for the Box-Cox transformation of wages in the SLID regression. The least-squares line is shown on the plot.

5.7.2 Box-Tidwell Transformation of the X's

- Now, consider the model
$$Y_i = \alpha + \beta_1 X_{i1}^{\gamma_1} + \cdots + \beta_k X_{ik}^{\gamma_k} + \varepsilon_i$$
where the errors are independently distributed as $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, and all of the X_{ij} are positive.

- The parameters of this model — $\alpha, \beta_1, ..., \beta_k, \gamma_1, ..., \gamma_k$, and σ_ε^2 — could be estimated by general nonlinear least squares, but Box and Tidwell suggest instead a computationally more efficient procedure that also yields a constructed-variable diagnostic:

1. Regress Y on $X_1, ..., X_k$, obtaining $A, B_1, ..., B_k$.
2. Regress Y on $X_1, ..., X_k$ *and* the constructed variables $X_1 \log_e X_1, ..., X_k \log_e X_k$, obtaining $A', B_1', ..., B_k'$ and $D_1, ..., D_k$.

3. The constructed variable $X_j \log_e X_j$ can be used to assess the need for a transformation of X_j by testing the null hypothesis H_0: $\delta_j = 0$, where δ_j is the population coefficient of $X_j \log_e X_j$ in step 2. Added-variable plots for the constructed variables are useful for assessing leverage and influence on the decision to transform the X's.

4. A preliminary estimate of the transformation parameter γ_j (not the MLE) is given by

$$\widetilde{\gamma}_j = 1 + \frac{D_j}{B_j}$$

- This procedure can be iterated through steps 1, 2, and 4 until the estimates of the transformation parameters stabilize, yielding the MLEs $\widehat{\gamma}_j$.

- Consider the SLID regression of log wages on sex, education, and age.
  - The dummy regressor for sex is not a candidate for transformation, of course, but I will consider power transformations of age and education.
    - Recall that we were initially undecided about whether to model the age effect as a quadratic or as a transformation down the ladder of powers and roots.
  - To make power transformations of age more effective, I use a negative start of 15 (recall that age ranges from 16 to 65).
  - The coefficients of $(\text{Age} - 15) \times \log_e(\text{Age} - 15)$ and Education$\times$ $\log_e$Education in the step-2 augmented model are, respectively, $D_{\text{Age}} = -0.04699$ with SE$(D_{\text{Age}}) = 0.00231$, and $D_{\text{Education}} = 0.05612$ with SE$(D_{\text{Education}}) = 0.01254$.
  - Both score tests are statistically significant, but there is much stronger evidence of the need to transform age.

- The first-step estimates of the transformation parameters are

$$\widetilde{\gamma}_{\text{Age}} = 1 + \frac{D_{\text{Age}}}{B_{\text{Age}}} = 1 + \frac{-0.04699}{0.02619} = -0.79$$

$$\widetilde{\gamma}_{\text{Education}} = 1 + \frac{D_{\text{Education}}}{B_{\text{Education}}} = 1 + \frac{0.05612}{0.08061} = 1.69$$

- The fully iterated MLEs of the transformation parameters are $\widehat{\gamma}_{\text{Age}} = 0.051$ and $\widehat{\gamma}_{\text{Education}} = 1.89$ — very close to the log transformation of started-age and the square of education.
- Constructed-variable plots for the transformation of age and education are shown in Figure 45.

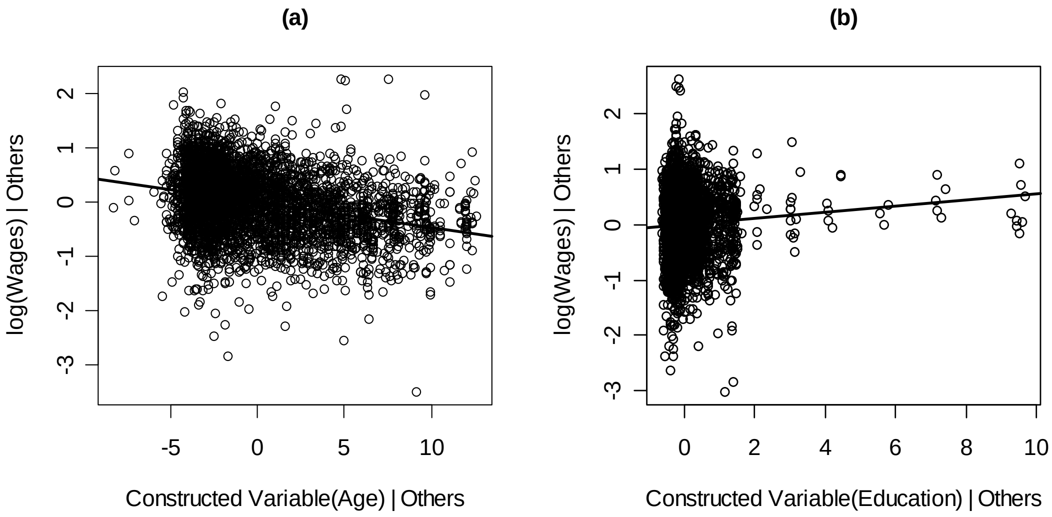

Figure 45. Constructed-variable plots for the Box-Tidwell transformation of (a) age and (b) education in the SLID regression of log wages on sex, age, and education.

5.7.3 Non-Constant Error Variance Revisited

- ▶ Breusch and Pagan develop a score test for heteroscedasticity based on the specification:

  $$\sigma_i^2 \equiv V(\varepsilon_i) = g(\gamma_0 + \gamma_1 Z_{i1} + \cdots + \gamma_p Z_{ip})$$

  where $Z_1, ..., Z_p$ are known variables, and where the function $g(\cdot)$ is quite general.
  - • The same test was independently derived by Cook and Weisberg.
- ▶ The score statistic for the hypothesis that the σ_i^2 are all the same, which is equivalent to H_0: $\gamma_1 = \cdots = \gamma_p = 0$, can be formulated as an auxiliary-regression problem.
  - • Let $U_i \equiv E_i^2/\widehat{\sigma}_\varepsilon^2$, where $\widehat{\sigma}_\varepsilon^2 = \sum E_i^2/n$ is the MLE of the error variance. Regress U on the Z's:

    $$U_i = \eta_0 + \eta_1 Z_{i1} + \cdots + \eta_p Z_{ip} + \omega_i$$

- • Breusch and Pagan show that the score statistic

  $$S_0^2 = \frac{\sum(\widehat{U}_i - \overline{U})^2}{2}$$

  is asymptotically distributed as χ^2 with p degrees of freedom under the null hypothesis of constant error variance.
- • Here, the $\widehat{U}_i$ are fitted values from the regression of U on the Z's, and thus S_0^2 is half the regression sum of squares from the auxiliary regression.

- ▶ To apply this result, it is necessary to select Z's, the choice of which depends upon the suspected pattern of non-constant error variance.
  - • Employing $X_1, ..., X_k$ in the auxiliary regression, for example, permits detection of a tendency of the error variance to increase (or decrease) with the values of one or more of the explanatory variables in the main regression.

- Cook and Weisberg suggest regressing U on the fitted values from the main regression (i.e., $U_i = \eta_0 + \eta_1 \widehat{Y}_i + \omega_i$), producing a one-degree-of-freedom score test to detect the common tendency of the error variance to increase with the level of the response variable.
  - · Anscombe suggests correcting detected heteroscedasticity by transforming Y to $Y^{(\widetilde{\lambda})}$ with $\widetilde{\lambda} = 1 - 1/2\widehat{\eta}_1\overline{Y}$.

▶ Applied to the initial SLID regression of wages on sex, age, and education, an auxiliary regression of U on $\widehat{Y}$ yields $\widehat{U} = -0.3449 + 0.08652\widehat{Y}$, and $S_0^2 = 567.66/2 = 283.83$ on 1 degree of freedom, for which $p \approx 0$.

- The suggested variance-stabilizing transformation using Anscombe's rule is

$$\widetilde{\lambda} = 1 - \frac{1}{2}(0.08652)(15.545) = 0.33$$

- An auxiliary regression of U on the explanatory variables in the main regression yields $S_0^2 = 579.08/2 = 289.54$ on $k = 3$ degrees of freedom.
  - · The score statistic for the more general test is not much larger than that for the regression of U on $\widehat{Y}$, implying that the pattern of non-constant error variance is indeed for the spread of the errors to increase with the level of Y.

5.8 Summary: Nonlinearity and Other Ills

- ▶ Heavy-tailed errors threaten the efficiency of least-squares estimation; skewed and multimodal errors compromise the interpretation of the least-squares fit.
  - Non-normality can often be detected by examining the distribution of the least-squares residuals, and frequently can be corrected by transforming the data.
- ▶ It is common for the variance of the errors to increase with the level of the response variable.
  - This pattern of non-constant error variance can often be detected in a plot of residuals against fitted values.

  - Strategies for dealing with non-constant error variance include transformation of the response variable to stabilize the variance; the substitution of weighted-least-squares estimation for ordinary least squares; and the correction of coefficient standard errors for heteroscedasticity.
  - A rough rule of thumb is that non-constant error variance seriously degrades the least-squares estimator only when the ratio of the largest to smallest variance is about 10 or more.
- ▶ Simple forms of nonlinearity can often be detected in component+residual plots.
  - Once detected, nonlinearity can frequently be accommodated by variable transformations or by altering the form of the model (to include a quadratic term in an explanatory variable, for example).
  - Component+residual plots adequately reflect nonlinearity when the explanatory variables are themselves not strongly nonlinearly related.

- ▶ Discrete explanatory variables divide the data into groups.
  - A simple incremental F-test for nonlinearity compares the sum of squares accounted for by the linear regression of Y on X with the sum of squares accounted for by differences in the group means.
  - Likewise, tests of non-constant variance can be based upon comparisons of spread in the different groups.
- ▶ A statistically sophisticated general approach to selecting a transformation of Y or an X is to imbed the linear-regression model in a more general model that contains a parameter for the transformation.
  - The Box-Cox procedure selects a power transformation of Y to normalize the errors.
  - The Box-Tidwell procedure selects power transformations of the X's to linearize the regression of Y on the X's.
  - In both cases, 'constructed-variable' plots help us to decide whether individual observations are unduly influential in determining the transformation parameters.

- ▶ Simple score tests are available to determine the need for a transformation and to test for non-constant error variance.

6. Diagnostics for Generalized Linear Models

6.1 Goals

- To review the structure of generalized linear models.
- To show how diagnostics for unusual data and nonlinearity can be extended to generalized linear models.

6.2 Review: The Structure of Generalized Linear Models

- A generalized linear model consists of three components:

1. A *random component*, specifying the conditional distribution of the response variable, Y_i, given the explanatory variables.
   - Traditionally, the random component is a member of an "exponential family" — the normal (Gaussian), binomial, Poisson, gamma, or inverse-Gaussian families of distributions — but generalized linear models have been extended beyond the exponential families.
   - The Gaussian and binomial distributions are familiar.
   - Poisson distributions are often used in modeling count data. Poisson random variables take on non-negative integer values, $0, 1, 2, \ldots$. Some examples are shown in Figure 46.

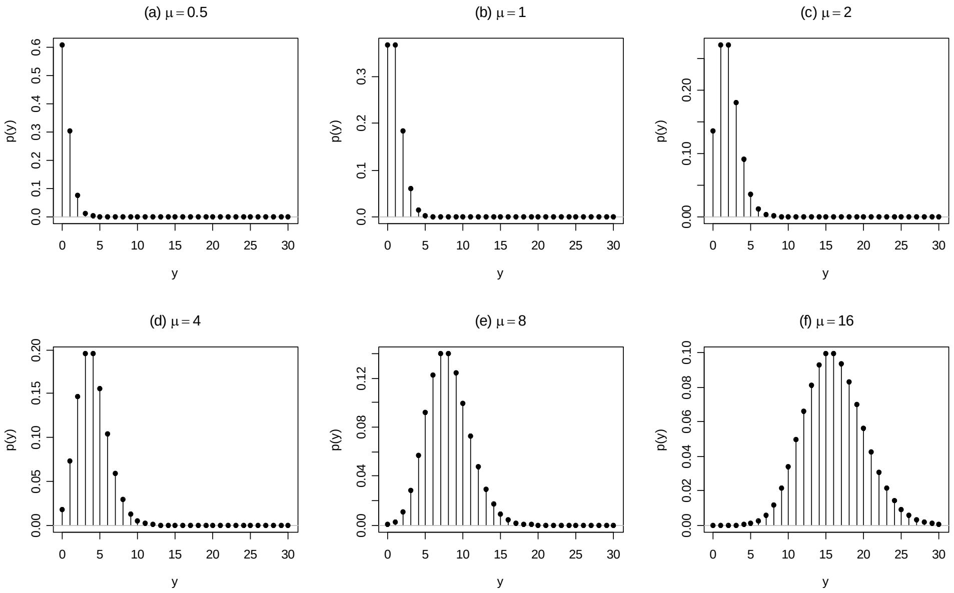

Figure 46. Poisson distributions for various values of the “rate” parameter (mean) μ.

- The gamma and inverse-Gaussian distributions are for positive continuous data; some examples are given in Figure 47.

2. A linear function of the regressors, called the *linear predictor*,

$$\eta_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$$

on which the expected value μ_i of Y_i depends.

- The X’s may include quantitative predictors, but they may also include transformations of predictors, polynomial terms, contrasts generated from factors, interaction regressors, etc.

3. An invertible *link function* $g(\mu_i) = \eta_i$, which transforms the expectation of the response to the linear predictor.

- The inverse of the link function is sometimes called the *mean function*: $g^{-1}(\eta_i) = \mu_i$.

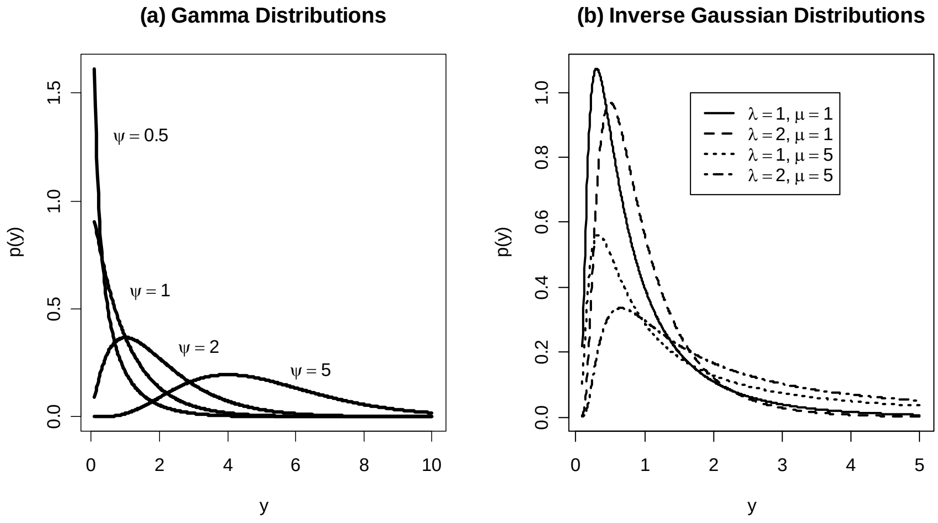

Figure 47. (a) Several gamma distributions for “scale” $\omega = 1$ and various values of the “shape” parameter ψ. (b) Inverse-Gaussian distributions for several combinations of values of the mean μ and “inverse-dispersion” λ.

- Standard link functions and their inverses are shown in the following table:

Link	$\eta_i = g(\mu_i)$	$\mu_i = g^{-1}(\eta_i)$
identity	μ_i	η_i
log	$\log_e \mu_i$	e^{η_i}
inverse	μ_i^{-1}	η_i^{-1}
inverse-square	μ_i^{-2}	$\eta_i^{-1/2}$
square-root	$\sqrt{\mu_i}$	η_i^2
logit	$\log_e \frac{\mu_i}{1-\mu_i}$	$\frac{1}{1+e^{-\eta_i}}$
probit	$\Phi^{-1}(\mu_i)$	$\Phi(\eta_i)$
log-log	$-\log_e[-\log_e(\mu_i)]$	$\exp[-\exp(-\eta_i)]$
complementary log-log	$\log_e[-\log_e(1-\mu_i)]$	$1-\exp[-\exp(\eta_i)]$

- The logit, probit, and complementary-log-log links are for *binomial data*, where Y_i represents the observed proportion and μ_i the expected proportion of “successes” in n_i binomial trials — that is, μ_i is the probability of a success.

- · For the probit link, Φ is the standard-normal cumulative distribution function, and Φ^{-1} is the standard-normal quantile function.
- · An important special case is *binary data*, where all of the binomial trials are 1, and therefore all of the observed proportions Y_i are either 0 or 1.

▶ For distributions in the exponential families, the conditional variance of Y is a function of the mean μ together with a dispersion parameter ϕ (as shown in the table below).

- For the binomial and Poisson distributions, the dispersion parameter is fixed to 1.
- For the Gaussian distribution, the dispersion parameter is the usual error variance, which we previously symbolized by σ_ε^2 (and which doesn't depend on μ).

Family	*Canonical Link*	*Range of* Y_i	$V(Y_i\|\eta_i)$
Gaussian	identity	$(-\infty, +\infty)$	ϕ
binomial	logit	$\frac{0, 1, ..., n_i}{n_i}$	$\frac{\mu_i(1-\mu_i)}{n_i}$
Poisson	log	$0, 1, 2, ...$	μ_i
gamma	inverse	$(0, \infty)$	$\phi\mu_i^2$
inverse-Gaussian	inverse-square	$(0, \infty)$	$\phi\mu_i^3$

- ▶ The *canonical link* for each familiy is not only the one most commonly used, but also arises naturally from the general formula for distributions in the exponential families.
  - Other links may be more appropriate for the specific problem at hand
  - One of the strengths of the GLM paradigm — in contrast, for example, to transformation of the response variable in a linear model — is the separation of the link function from the conditional distribution of the response.
- ▶ GLMs are typically fit to data by the method of maximum likelihood.
  - Denote the maximum-likelihood estimates of the regression parameters as $\widehat{\alpha}, \widehat{\beta}_1, ..., \widehat{\beta}_k$.
    - · These imply an estimate of the mean of the response, $\widehat{\mu}_i = g^{-1}(\widehat{\alpha} + \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_k x_{ik})$.

- The log-likelihood for the model, maximized over the regression coefficients, is

$$\log_e L_0 = \sum_{i=1}^{n} \log_e p(\widehat{\mu}_i, \phi; y_i)$$

  where $p(\cdot)$ is the probability or probability-density function corresponding to the family employed.

- A “saturated” model, which dedicates one parameter to each observation, and hence fits the data perfectly, has log-likelihood

$$\log_e L_1 = \sum_{i=1}^{n} \log_e p(y_i, \phi; y_i)$$

- Twice the difference between these log-likelihoods defines the *residual deviance* under the model, a generalization of the residual sum of squares for linear models:

$$D(\mathbf{y}; \widehat{\boldsymbol{\mu}}) = 2(\log_e L_1 - \log_e L_0)$$

- Dividing the deviance by the estimated dispersion produces the *scaled deviance*: $D(\mathbf{y}; \widehat{\boldsymbol{\mu}})/\widehat{\phi}$.
- Likelihood-ratio tests can be formulated by taking differences in the residual deviance for nested models.
- For models with an estimated dispersion parameter, one can alternatively use incremental F-tests.
- Wald tests for individual coefficients are formulated using the estimated asymptotic standard errors of the coefficients.

6.3 Outlier, Leverage, and Influence Diagnostics for GLMs

- ▶ Most regression diagnostics extend straightforwardly to generalized linear models.
- ▶ These extensions typically take advantage of the computation of maximum-likelihood estimates for generalized linear models by iterated weighted least squares (the procedure typically used to fit GLMs).

6.3.1 Hat-Values

- Hat-values for a generalized linear model can be taken directly from the final iteration of the IWLS procedure
- They have the usual interpretation — except that the hat-values in a GLM depend on Y as well as on the configuration of the X's.

6.3.2 Residuals

- Several kinds of residuals can be defined for generalized linear models:
  - *Response residuals* are simply the differences between the observed response and its estimated expected value: $Y_i - \widehat{\mu}_i$.
  - *Working residuals* are the residuals from the final WLS fit.
    - These may be used to define partial residuals for component-plus-residual plots (see below).
  - *Pearson residuals* are case-wise components of the Pearson goodness-of-fit statistic for the model:
  $$\frac{\widehat{\phi}^{1/2}(Y_i - \widehat{\mu}_i)}{\sqrt{\widehat{V}(Y_i|\eta_i)}}$$
  where ϕ is the dispersion parameter for the model and $V(Y_i|\eta_i)$ is the variance of the response given the linear predictor.

- *Standardized Pearson residuals* correct for the conditional response variation and for the leverage of the observations:
$$R_{Pi} = \frac{Y_i - \widehat{\mu}_i}{\sqrt{\widehat{V}(Y_i|\eta_i)(1-h_i)}}$$
.
- *Deviance residuals*, D_i, are the square-roots of the case-wise components of the residual deviance, attaching the sign of $Y_i - \widehat{\mu}_i$.

▶ *Standardized deviance residuals* are
$$R_{Di} = \frac{D_i}{\sqrt{\widehat{\phi}(1-h_i)}}$$

▶ Several different approximations to *studentized residuals* have been suggested.

- To calculate exact studentized residuals would require literally refitting the model deleting each observation in turn, and noting the decline in the deviance.

- Here is an approximation due to Williams:
$$E_i^* = \sqrt{(1-h_i)R_{Di}^2 + h_i R_{Pi}^2}$$
where, once again, the sign is taken from $Y_i - \widehat{\mu}_i$.
- A Bonferroni outlier test using the standard normal distribution may be based on the largest absolute studentized residual.

6.3.3 Influence Measures

- An approximation to Cook's distance influence measure is

$$D_i = \frac{R_{Pi}^2}{\widehat{\phi}(k+1)} \times \frac{h_i}{1-h_i}$$

- Approximate values of dfbeta$_{ij}$ (influence on each coefficient) may be obtained directly from the final iteration of the IWLS procedure.
- There are two largely similar extensions of added-variable plots to generalized linear models, one due to Wang and another to Cook and Weisberg.

6.4 Nonlinearity Diagnostics

- Component-plus-residual plots also extend straightforwardly to generalized linear models.
  - Nonparametric smoothing of the resulting scatterplots can be important to interpretation, especially in models for binary responses, where the discreteness of the response makes the plots difficult to examine.
  - Similar effects can occur for binomial and Poisson data.
- Component-plus-residual plots use the linearized model from the last step of the IWLS fit.
  - For example, the partial residual for X_j adds the working residual to $B_j X_{ij}$.
  - The component-plus-residual plot graphs the partial residual against X_j.

- An illustrative component+residual plot, for assets in an over-dispersed Poisson regression fit to Ornstein's interlocking-directorate data appears in Figure 48.
  - This plot is difficult to examine because of the large positive skew in assets, but it appears as if the assets slope is a good deal steeper at the left than at the right.
  - I therefore investigated transforming assets down the ladder of powers and roots, eventually arriving at the log transformation, the component+residual plot for which appears quite straight (Figure 49).

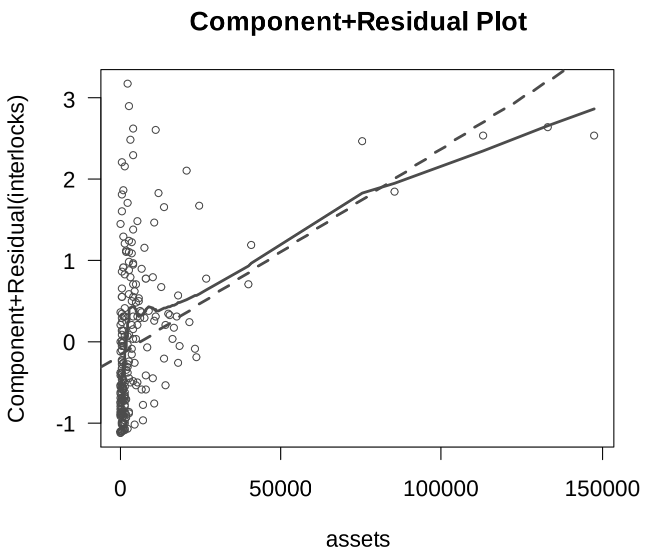

Figure 48. Component+residual plot for assets in the over-dispersed Poisson regression for Ornstein's data.

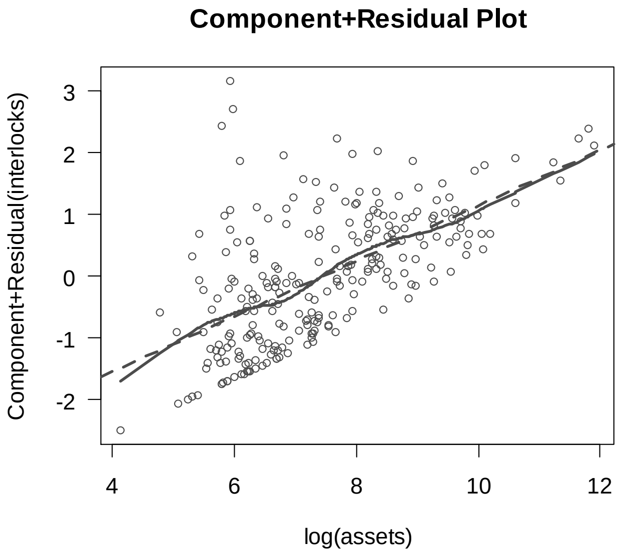

Figure 49. Component+residual plot for log(assets) in the respecified over-dispersed Poisson regression model for Ornstein's data.

6.5 Summary: Diagnostics for GLMs

▶ Generalized linear models (GLMs) consist of three components:

(a) A random component specifying the conditional distribution of the response variable Y given the explanatory variables, traditionally a member of an exponential family — the normal (Gaussian), binomial, Poisson, gamma, or inverse-Gaussian families of distributions.

· For distributions in exponential families, the conditional variance of Y is a function of μ, the mean of Y, and of a dispersion parameter ϕ; in the binomial and Poisson families, ϕ is fixed to 1.

(b) A linear predictor, $\eta_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$.

(c) A link function $g(\mu_i) = \eta_i$, which transforms the expectation of the response to the linear predictor; the inverse of the link is the mean function, $g^{-1}(\eta_i) = \mu_i$.

- Traditional GLMs are fit to data by maximum likelihood.
- Most standard linear-model diagnostics may be generalized to GLMs. These include hat-values, studentized residuals, Cook's distances, added-variable plots, and component-plus-residual plots (among others).