POLS 7014

1. Introduction

1. Goals

- To introduce the notion of regression analysis as a description of how the average value of a response variable changes with the value(s) of one or more explanatory variables.
- To show that this essential idea can be pursued 'nonparametrically' without making strong prior assumptions about the structure of the data.
- To introduce or review basic concepts: skewness, sampling variance, bias, outliers, etc.

2. Introduction

- *Regression analysis* traces the distribution of a *response* (or *dependent*) variable (denoted by Y) as a function of one or more *explanatory* (or *independent* or *predictor*) *variables* ($X_1, ..., X_k$):

$$p(y|x_1, ..., x_k) = f(x_1, ..., x_k)$$

  - $p(y|x_1, ..., x_k)$ represents the probability (or, for continuous Y, the probability density) of observing the specific value y of the response variable, *conditional* upon a set of specific values ($x_1, ..., x_k$) of the explanatory variables.

- Imagine, for example, that Y is individuals' income and that the X's are a variety of characteristics upon which income might depend, such as education, gender, age, and so on. In what follows, I restrict consideration to quantitative X's, such as years of education and age.

- Most discussions of regression analysis begin by assuming (see Figure 1, drawn for a single explanatory variable X)
  - that the conditional distribution of the response variable, $p(Y|x_1, ..., x_k)$, is a normal distribution
  - that the variance of Y conditional on the X's, denoted σ^2, is everywhere the same regardless of the specific values of $x_1, ..., x_k$
  - and that the expected value (the mean) of Y is a linear function of the X's:

$$\mu \equiv E(Y|x_1, ..., x_k) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

  - These assumptions, along with independent random sampling, lead to linear least-squares estimation.
- In contrast, I will pursue the notion of regression with as few preconceived assumptions as possible.

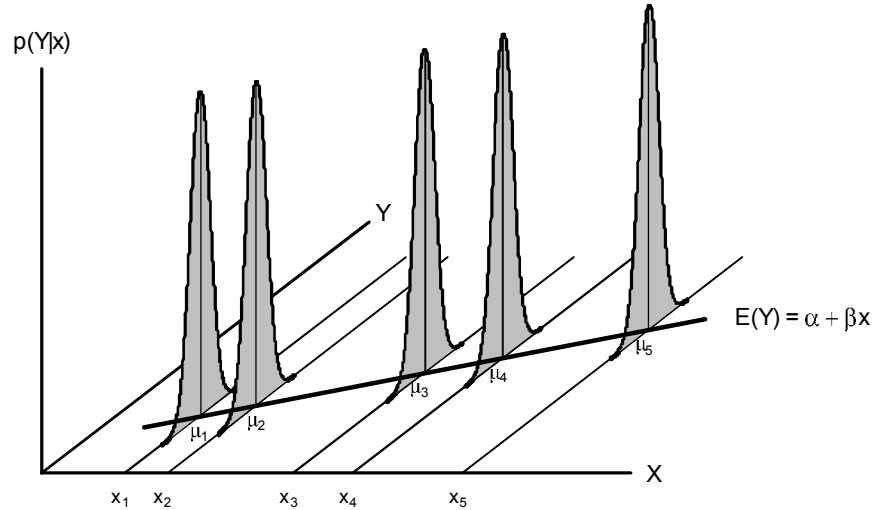

Figure 1. The usual assumptions: linearity, constant variance, and normality, for a single X.

- Figure 2 (for a single X) illustrates why we should not be too hasty to make the assumptions of normality, equal variance, and linearity:
  - **Skewness.** If the conditional distribution of Y is skewed then the mean will not be a good summary of its center.
  - **Multiple modes.** If the conditional distribution of Y is multimodal then it is intrinsically unreasonable to summarize its center with a single number.
  - **Heavy tails.** If the conditional distribution of Y is substantially non-normal — for example, heavy-tailed — then the sample mean will not be an efficient estimator of the center of the Y-distribution even when this distribution is symmetric.
  - **Unequal spread.** If the conditional variance of Y changes with the values of the X's then the efficiency of the usual least-squares estimates may be compromised; moreover, the nature of the dependence of the variance on the X's may itself be of interest.

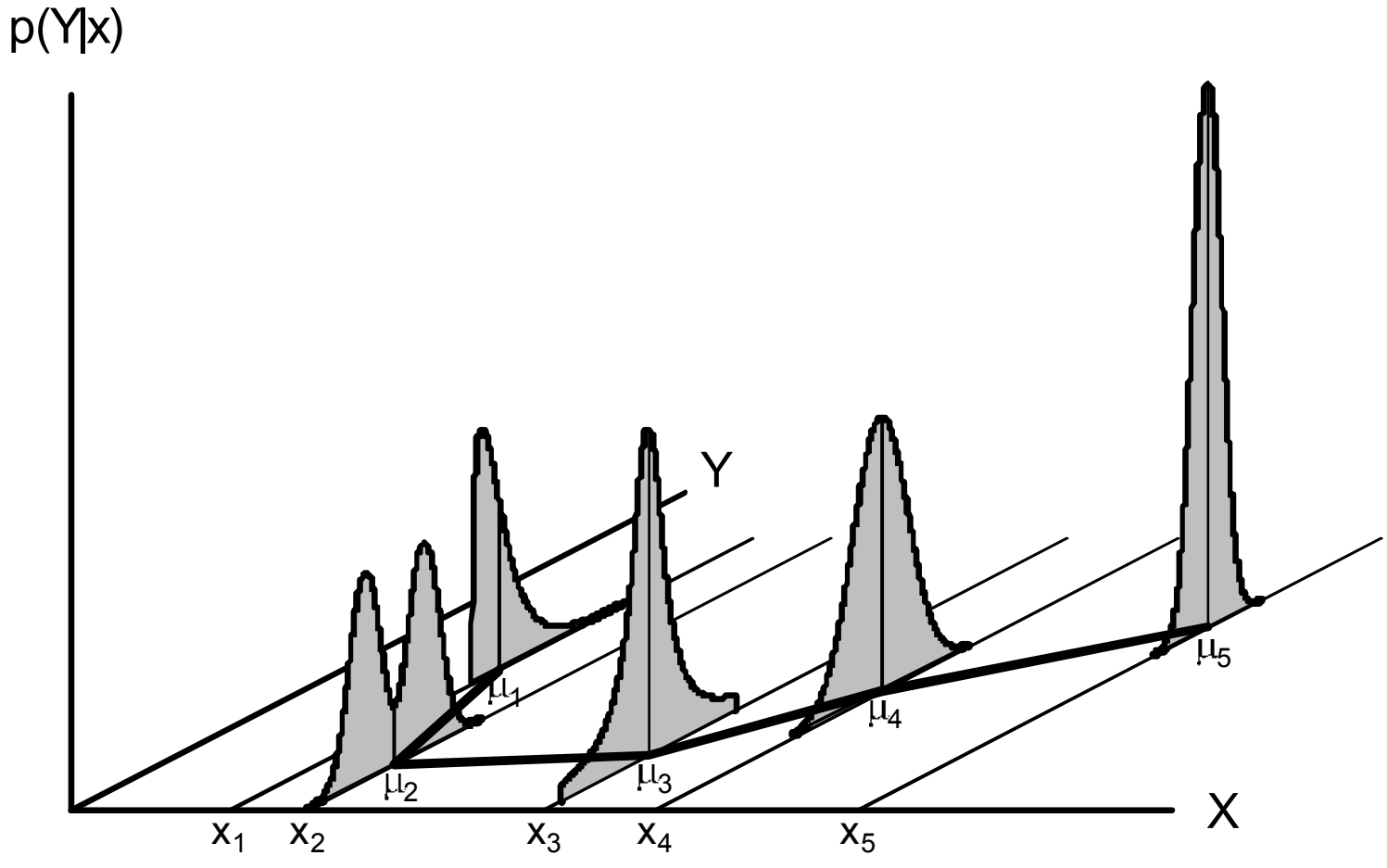

Figure 2. How the usual regression assumptions can fail.

- **Nonlinearity.** Although we are often in a position to expect that the values of Y will increase or decrease with some X, there is almost never good reason to assume *a priori* that the relationship between Y and X is linear; this problem is compounded when there are several X's.

▶ This is not to say, of course, that linear regression analysis or, more generally, linear statistical models, are of little practical use. Much of this course is devoted to the exposition of linear models. It is, however, prudent to begin with an appreciation of the limitations of linear models, since their effective use in data analysis frequently depends upon adapting to these limitations.

3. Naive Nonparametric Regression

- We have a large random sample of employed Canadians that includes hourly wages and years of education.
  - We could easily display the conditional distribution of wages for each of the values of education $(0, 1, 2, ..., 20)$ that occur in our data, as in Figure 3.
  - If we are interested in the population average or typical value of wages conditional on education, $\mu|x$, we could estimate (most of) these conditional averages very accurately using the sample means $\overline{Y}|x$ (see Figure 4).
    - Using the conditional means isn't a good idea here because the conditional distributions of wages given education are positively skewed.
  - Had we access to the entire population of employed Canadians, we could calculate $\mu|x$ directly.

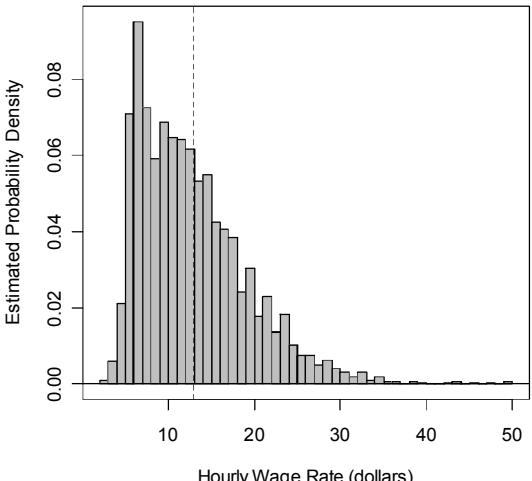

Figure 3. The conditional distribution of hourly wages for the 3384 employed Canadians in the SLID who had 12 years of education. The broken vertical line shows the conditional mean wages.

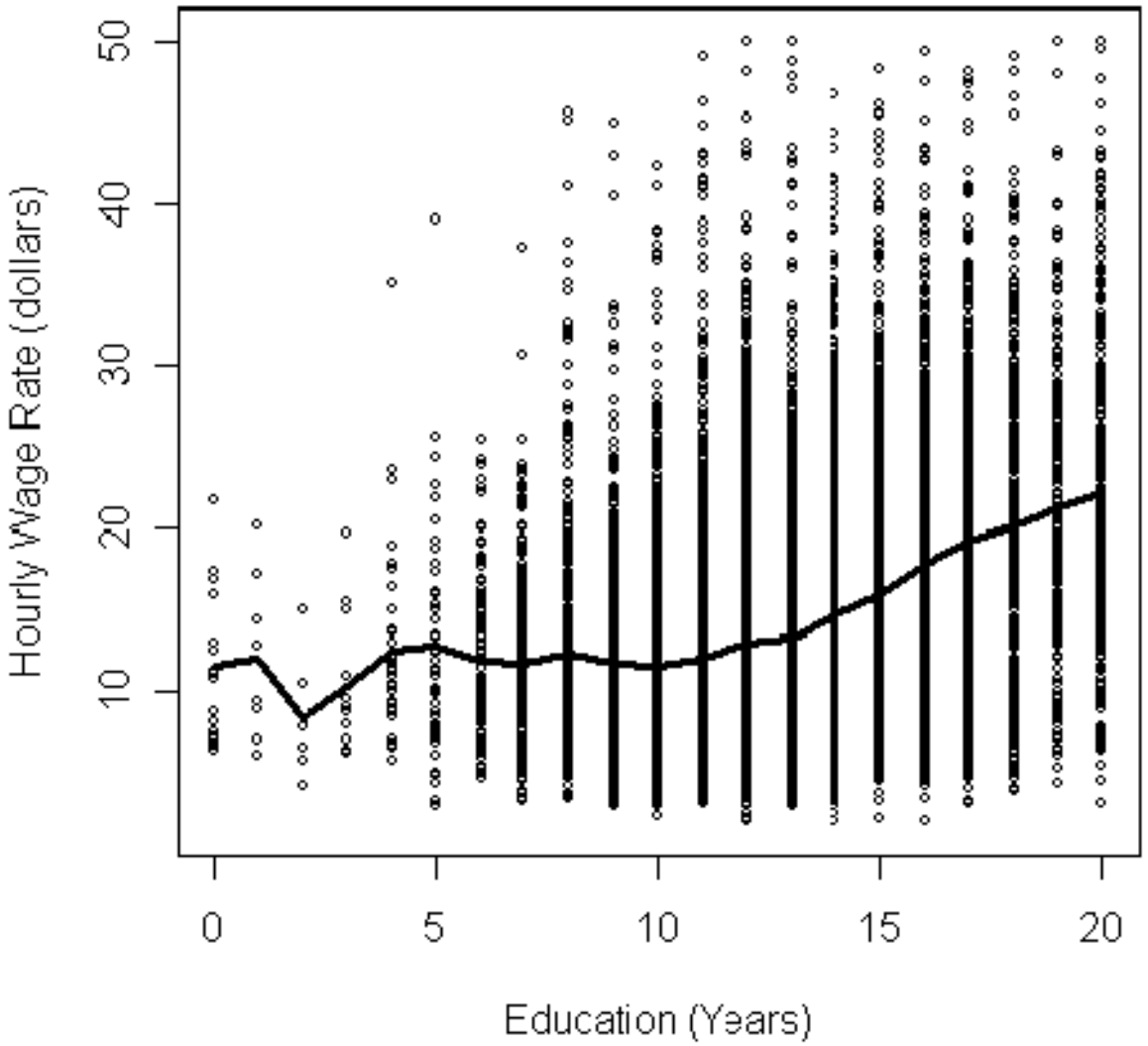

Figure 4. A scatterplot showing the relationship between hourly wages (in dollars) and education (in years) for a sample of 14,601 employed Canadians.

- Imagine now that X, along with Y, is a continuous variable.
  - For example, X is the reported weight in kilograms for each of a sample of individuals, and Y is their measured weight, again in kilograms.
  - We want to use reported weight to predict actual (i.e., measured) weight, and so we are interested in the mean value of Y as a function of X in the population of individuals from among whom the sample was randomly drawn:

$$\mu = E(Y|x) = f(x)$$

  - Even if the sample is large, replicated values of X will be rare because X is continuous, but for a large sample we can dissect the range of X into many narrow class intervals (or *bins*) of reported weight, each bin containing many observations; within each bin, we can display the conditional distribution of measured weight and estimate the conditional mean of Y with great precision.

- If we have fewer data at our disposal, we have to make do with fewer bins, each containing relatively few observations.
- This situation is illustrated in Figure 5, using data on reported and measured weight for each of 101 Canadian women engaged in regular exercise.
- Another example, using the prestige and income levels of 102 Canadian occupations in 1971, appears in Figure 6.
- The X-axes in these figures are carved into bins, each containing approximately 20 observations (the first and last bins contain the extra observations). The 'non-parametric regression line' displayed on each plot is calculated by connecting the points defined by the conditional response-variable means $\overline{Y}$ and the explanatory-variable means $\overline{X}$ in the five bins.

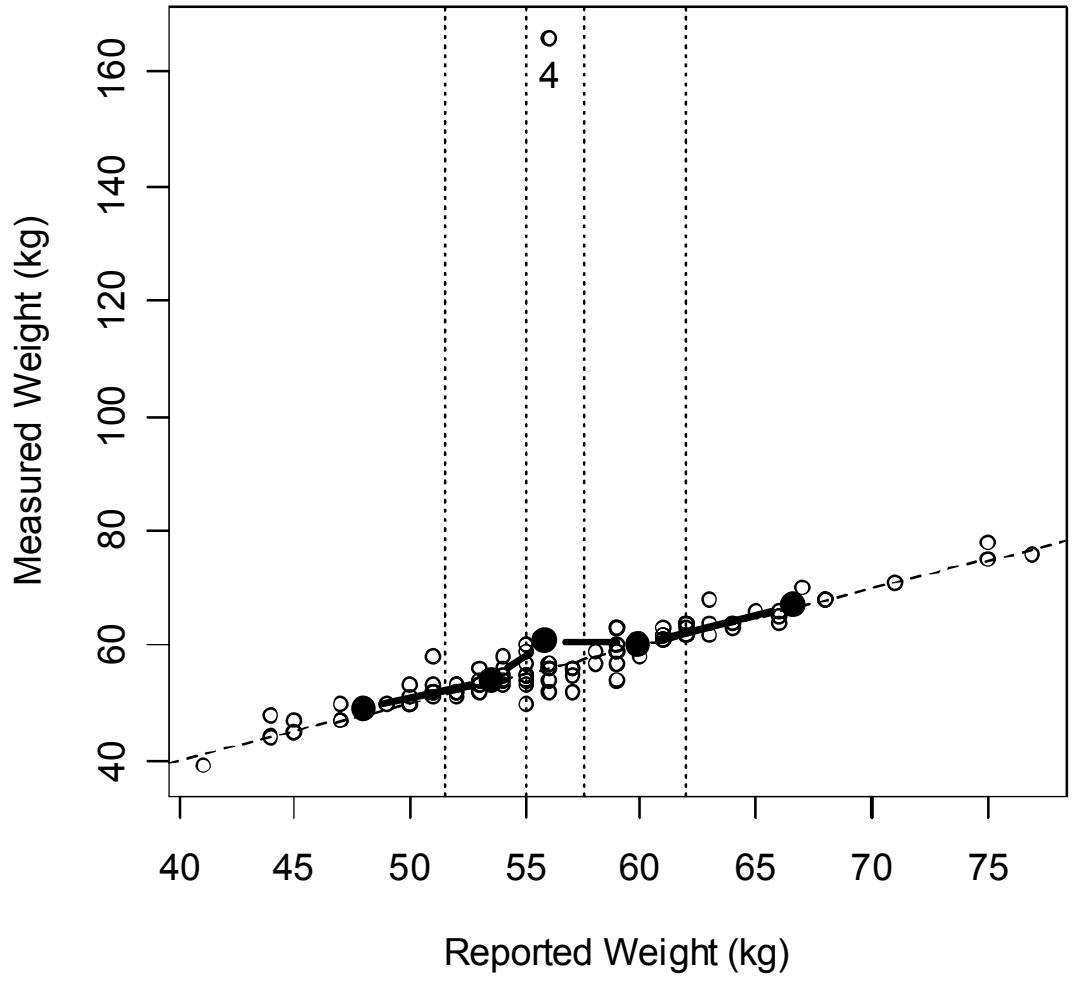

Figure 5. Naive nonparametric regression of measured on reported weight. The data are carved into fifths based on their X-values and the average Y in each fifth is calculated (the solid dots). Note the effect of the outlier (observation 4).

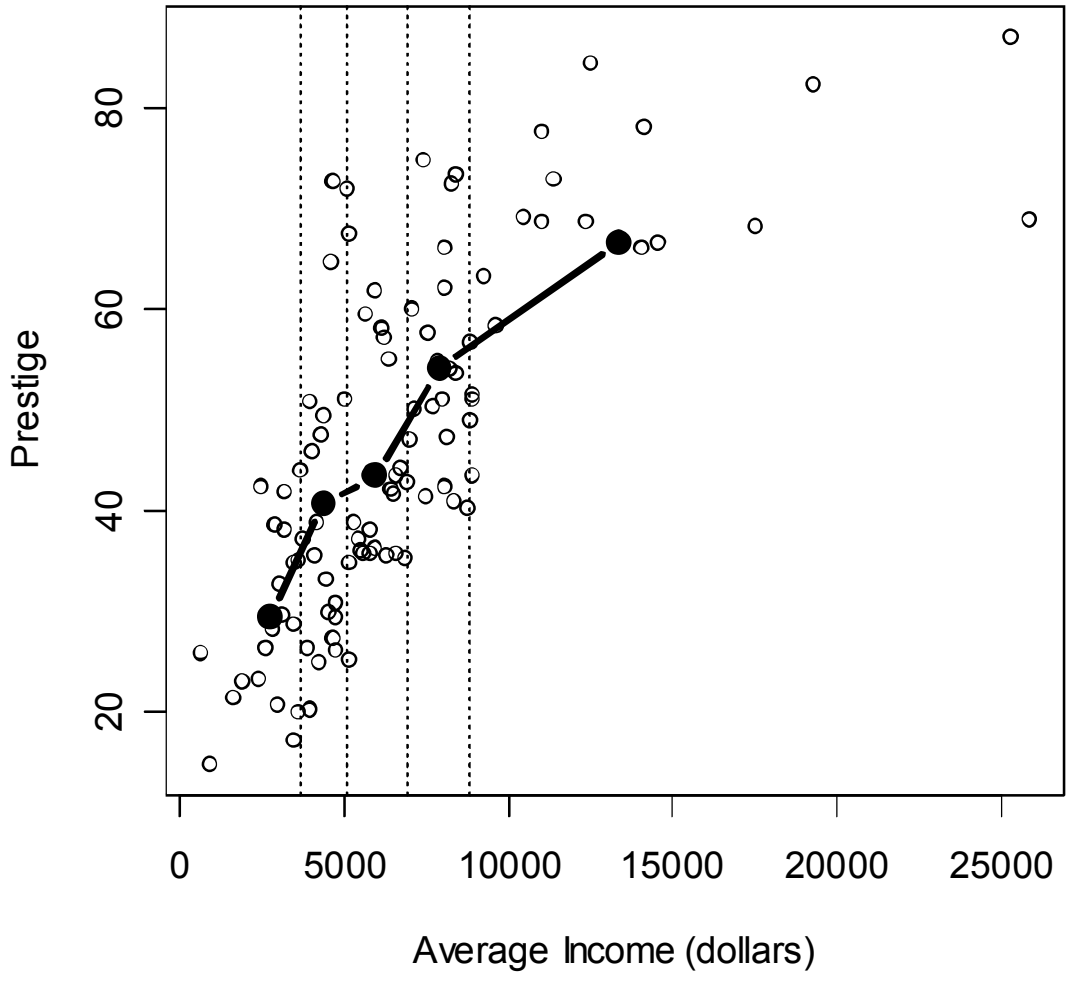

Figure 6. Naive nonparametric regression of occupational prestige on average income.

- There are two sources of error in this simple procedure of binning and averaging :
  - **Sampling error (variance)**. The conditional sample means $\overline{Y}$ will change if we select a new sample. Sampling error is minimized by using a small number of relatively wide bins, each with a substantial number of observations.
  - **Bias.** Let x_i denote the center of the ith bin (here, $i = 1, ..., 5$). If the population regression curve $f(x)$ is nonlinear within the interval, then the average population value of Y in the interval ($\overline{\mu}_i$) is usually different from the value of the regression curve at the center of the interval, $\mu_i = f(x_i)$, even if the x-values are evenly distributed within the interval. Bias is minimized by making the class-intervals as numerous and as narrow as possible (see Figure 7).

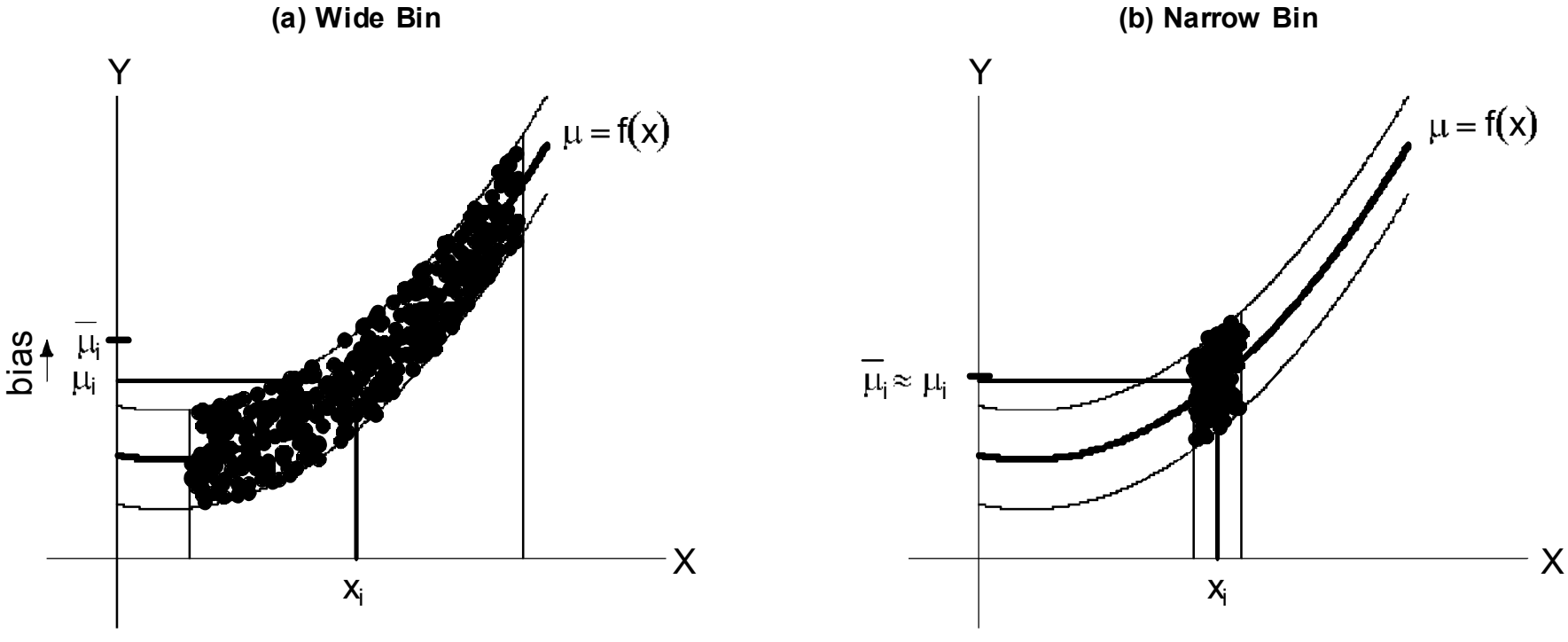

Figure 7. A narrow bin (b) generally produces less bias in estimating the regression curve than a wide bin (a).

- As is typically the case in statistical estimation, reducing bias and reducing sampling variance work at cross purposes.
  - Only if we select a very large sample can we have our cake and eat it too.
  - Naive nonparametric regression is, under very broad conditions, a *consistent* estimator of the population regression curve. As the sample size gets larger (i.e., as $n \to \infty$), we can insure that the intervals grow successively narrower, yet each contains more data.
- When there is more than one explanatory variable naive nonparametric regression is less practical:
  - Suppose, for example, that we have three discrete explanatory variables, each with ten values. There are, then, $10^3 = 1,000$ combinations of values of the three variables, and within each such combination there is a conditional distribution of Y [i.e., $p(Y|x_1, x_2, x_3)$].

- Even if the X's are independently distributed — implying equal expected numbers of observations for each of the $1,000$ combinations — we would require a very large sample indeed to calculate the conditional means of Y with sufficient precision.
- The situation is even worse when the X's are continuous, since dissecting the range of each X into as few as ten class intervals might introduce substantial bias into the estimation.
- The problem of dividing the data into too many parts grows exponentially more serious as the number of X's increases. Statisticians therefore often refer to the intrinsic sparseness of multivariate data as the 'curse of dimensionality.'

4. Local Regression

- There are much better methods of nonparametric regression than binning and averaging. We often will use a method called local regression as a data-analytic tool to smooth scatterplots.
  - Local regression produces a smoothed fitted value $\widehat{Y}$ corresponding to any X-value in the range of the data — usually, at the data-values x_i.
  - To find smoothed values, the procedure fits n linear (or polynomial) regressions to the data, one for each observation i, emphasizing the points with X-values that are near x_i. This procedure is illustrated in Figure 8.
- Here are the details (but don't worry about them):

1. *Choose the span:* Select a fraction of the data $0 < s \leq 1$ (called the *span* of the smoother) to include in each fit, corresponding to $m \equiv [s \times n]$ data values. Often $s = \frac{1}{2}$ or $s = \frac{2}{3}$ works well. Larger values of s produce smoother results.

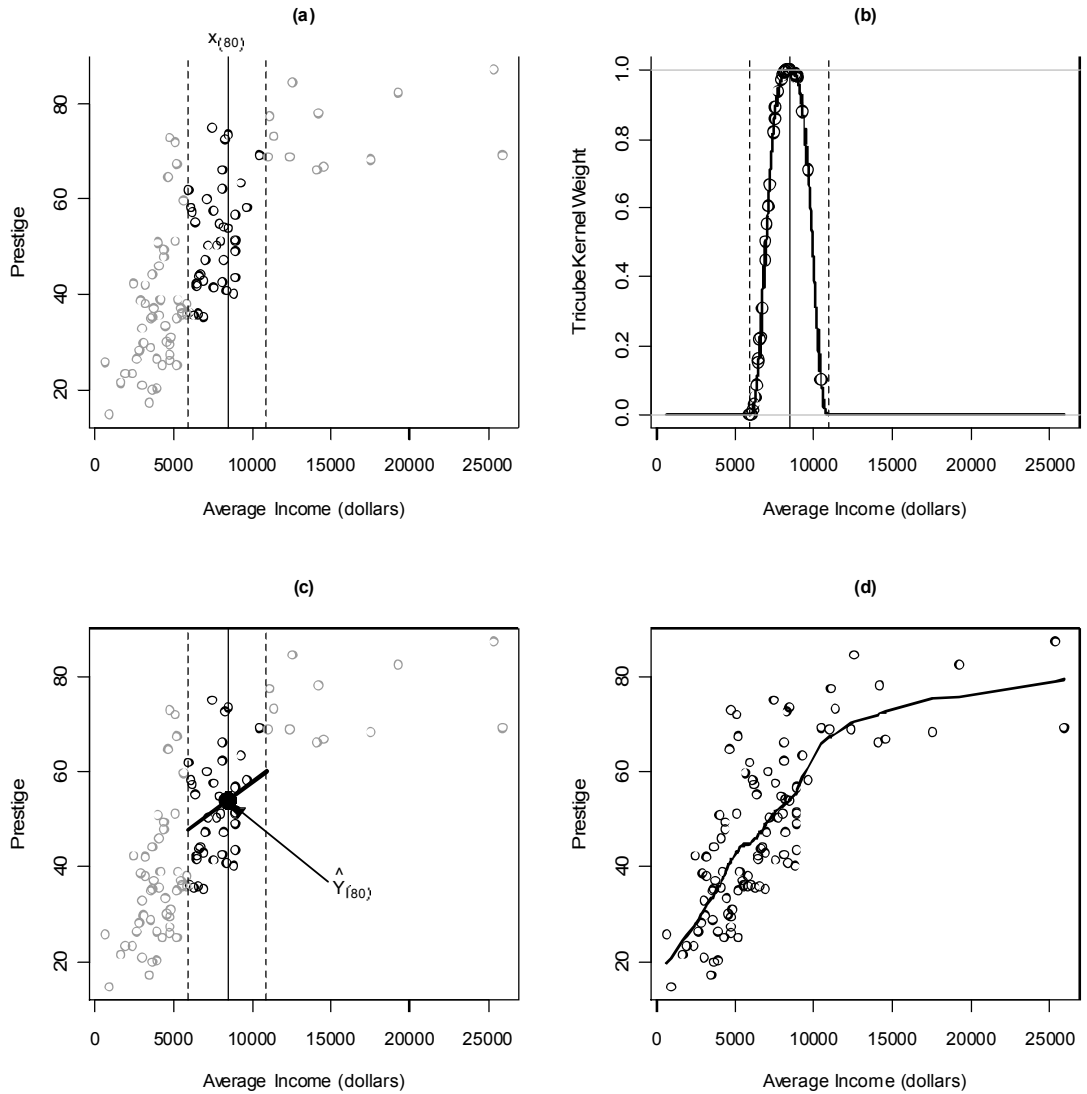

Figure 8. Local linear regression of occupational prestige on income, showing the computation of the fit at $x_{(80)}$.

2. *Locally weighted regressions:* For each $i = 1, 2, ..., n$, select the m values of X closest to x_i, denoted $x_{i1}, x_{i2}, ..., x_{im}$. The window half-width for observation i is then the distance to the farthest x_{ij}; that is, $h_i \equiv \max_{j=1}^{m} |x_{ij} - x_i|$. In panel (a) of Figure 8 the span is selected to include the $m = 40$ nearest neighbours of the focal value $x_{(80)}$ (which denotes the 80th ordered X-value).
   a. *Calculate weights:* For each of the m observations in the window, compute the weight

$$w_{ij} \equiv w_t\left(\frac{x_{ij} - x_i}{h_i}\right)$$

   where $w_t(\cdot)$ is the *tricube* weight function (see panel b):

$$w_t(z_{ij}) = \begin{cases} (1 - |z_{ij}|^3)^3 & \text{for } |z_{ij}| < 1 \\ 0 & \text{for } |z_{ij}| \geq 1 \end{cases}$$

   The tricube function assigns greatest weight to observations at the centre of the window and weights of 0 outside of the window.

b. *Local WLS fit:* Having computed the weights, fit the local regression equation

$$Y_{ij} = A_i + B_{i1}x_{ij} + E_{ij}$$

to minimize $\sum_{j=1}^{m} w_{ij}E_{ij}^2$ (i.e., by *weighted least squares*).

c. *Fitted value:* Compute the fitted value

$$\widehat{Y}_i = A_i + B_{i1}x_i$$

One regression equation is fit, and one fitted value is calculated, for each $i = 1, ..., n$ [see panel (c)]. Connecting these fitted values produces the nonparametric regression smooth [panel (d)].

5. Summary

- Regression analysis examines the relationship between a quantitative response variable Y and one or more quantitative explanatory variables, $X_1, ..., X_k$. Regression analysis traces the conditional distribution of Y — or some aspect of this distribution, such as its mean — as a function of the X's.
- In very large samples, and when the explanatory variables are discrete, it is possible to estimate a regression by directly examining the conditional distribution of Y given the X's. When the explanatory variables are continuous, we can proceed similarly by dissecting the X's into a large number of narrow bins.
- Local regression allows us to trace how the average Y changes with X even in small samples.

POLS 7014

2. Examining and Transforming Data

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).

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

- 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.

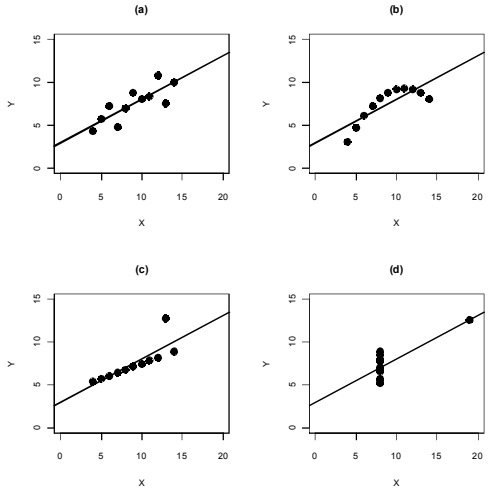

Figure 1. Anscombe's "quartet": Each data set has the same linear least--squares regression of Y on X.

- 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. Univariate Displays

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.

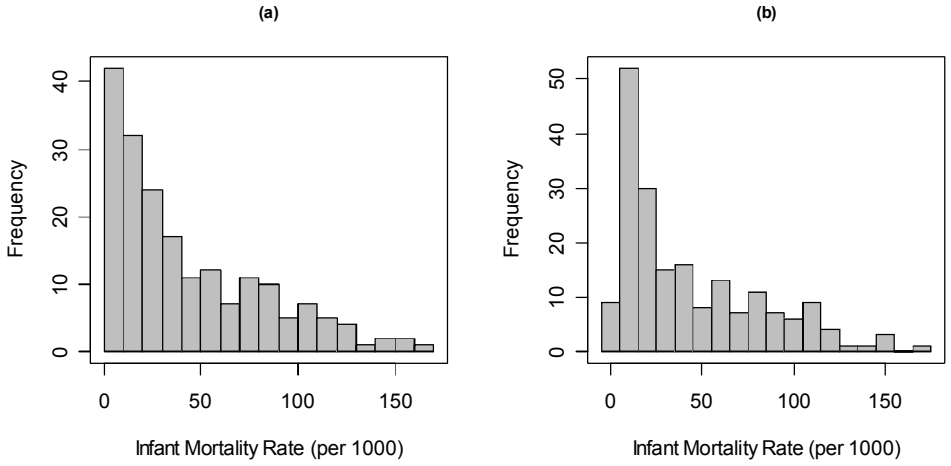

Figure 2. Two histograms with the same bin width but different origins for infant mortality in the United Nations data.

- 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 income, 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.

3.2 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.
  - Details are in the text
- An example is shown in Figure 3.

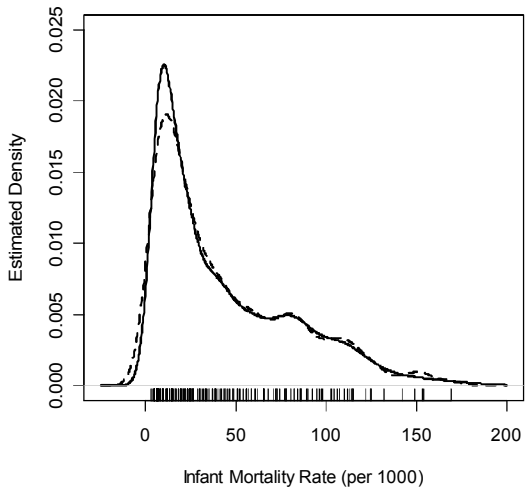

Figure 3. Kernel (broken line) and adaptive-kernel (solid line) density estimators for the distribution infant mortality. A "one-dimensional scatterplot" (or "rug plot") of the observations is shown at the bottom.

3.3 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:

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

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

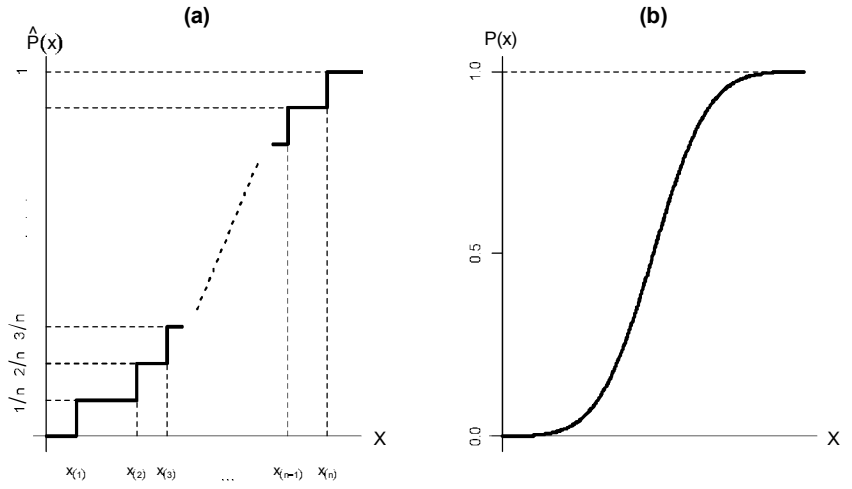

Figure 4. (a) Typical ECDF; (b) typical CDF.

► 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 $X_{(i)} \simeq z_i$.
- 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:

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

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

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

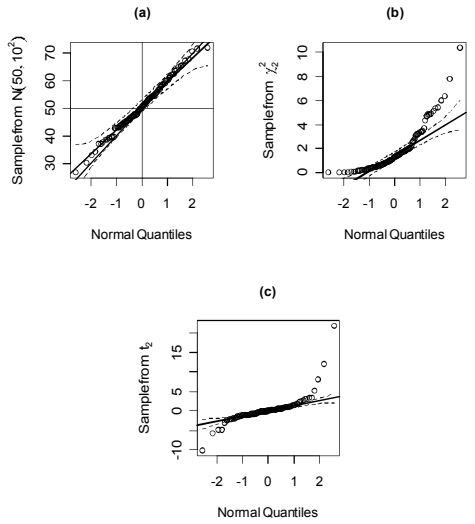

Figure 5. Normal quantile comparison plots for samples of size $n = 100$ drawn from three distributions.

- 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.

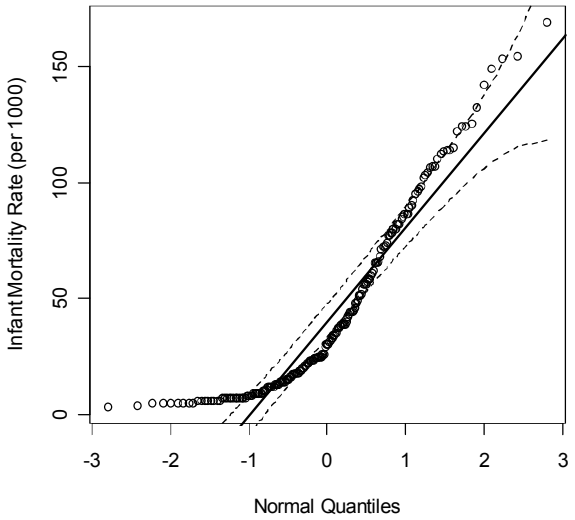

Figure 6. Normal quantile-comparison plot for infant mortality.

3.4 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.

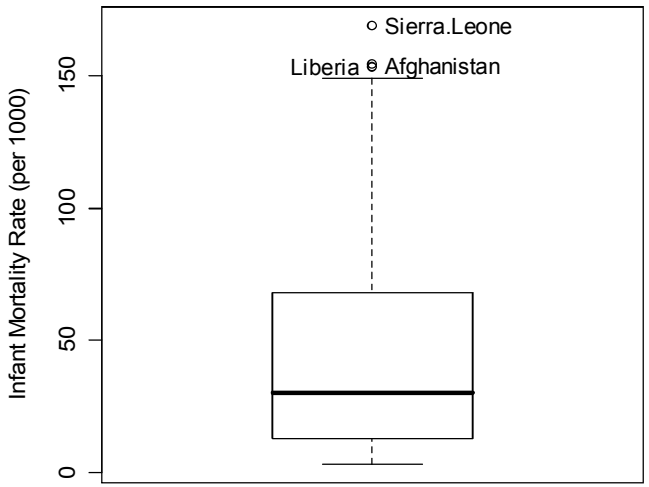

Figure 7. Boxplot of infant mortality.

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:
   $$\mathsf{F}_L = H_L - 1.5 \times H\text{-spread}$$
   $$\mathsf{F}_U = H_U + 1.5 \times H\text{-spread}$$
   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.

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 the social sciences 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.

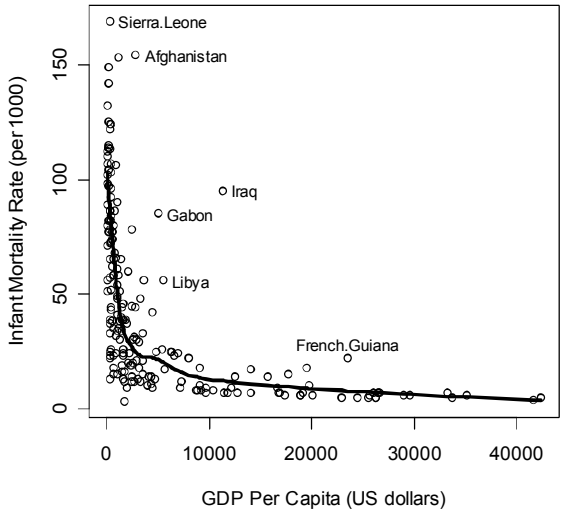

Figure 8. Scatterplot of infant morality by GDP per capita, for the UN data. The solid line is for a lowess smooth with a span of .5.

- 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.

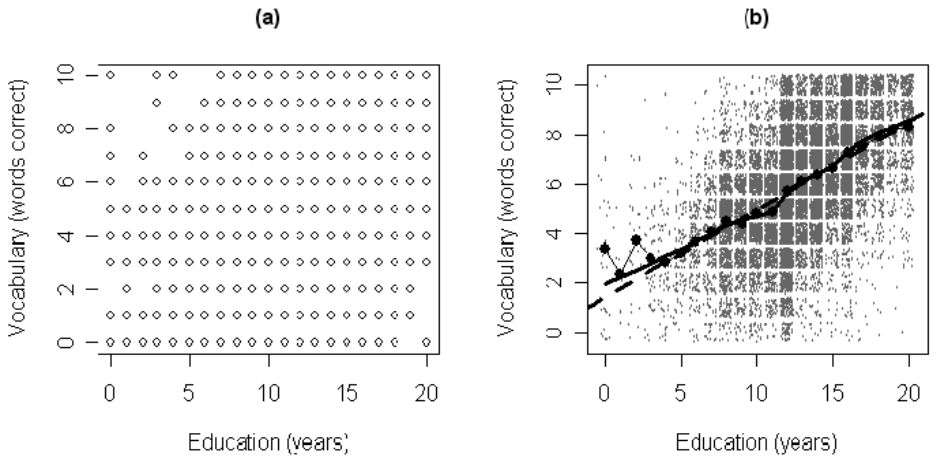

Figure 9. Vocabulary score by education: (a) original scatterplot; (b) jittered, with the least-squares lines, lowess line (for span = .2), and conditional means.

- 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.

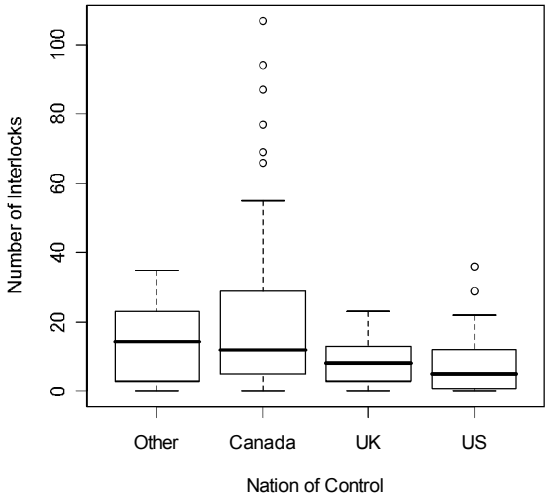

Figure 10. Parallel boxplots of number of interlocks by nation of control, for Ornstein's interlocking-directorate data.

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.

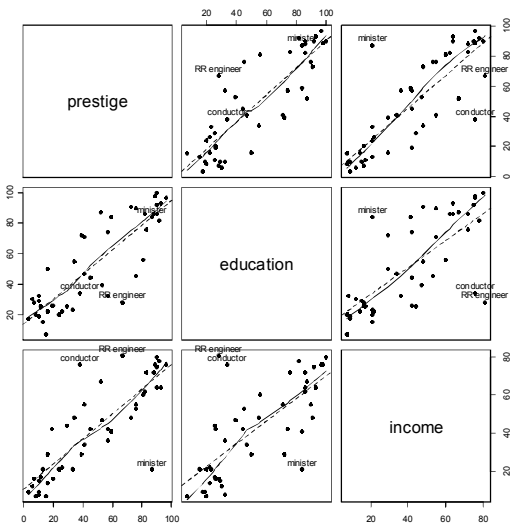

Figure 11. Scatterplot matrix for prestige, income, and education in Duncan's occupational prestige data.

  - 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 Davis's data on weight and reported weight.)

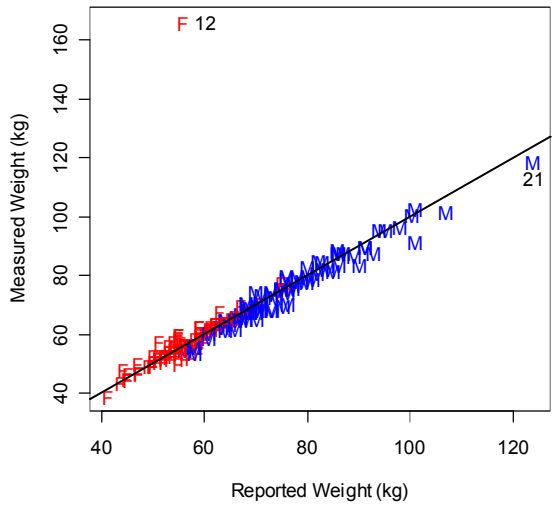

Figure 12. Measured by reported weight for 183 men (M) and women (F) engaged in regular exercise.

▶ 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.

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*):

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

▶ 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
2	1/2	$-1/2$
3	1/3	$-1/3$
4	1/4	$-1/4$

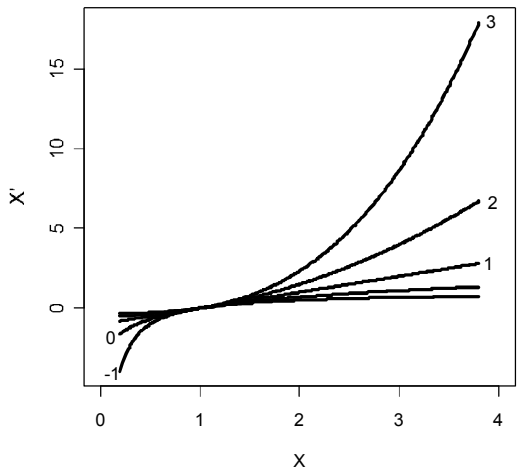

Figure 13. The Box-Cox familily of modified power transformations, $X^{(p)} = (X^p - 1)/p$, for values of $p = -1, 0, 1, 2, 3$. When $p = 0$, $X^{(p)} = \log_e X$.

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

$$\begin{aligned}
\log_{10} 100 = 2 &\iff 10^2 = 100 \\
\log_{10} 0.01 = -2 &\iff 10^{-2} = \tfrac{1}{10^2} = 0.01 \\
\log_{10} 10 = 1 &\iff 10^1 = 10 \\
\log_2 8 = 3 &\iff 2^3 = 8 \\
\log_2 \left(\tfrac{1}{8}\right) = -3 &\iff 2^{-3} = \tfrac{1}{2^3} = \tfrac{1}{8} \\
\log_b 1 = 0 &\iff b^0 = 1
\end{aligned}$$

▶ 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: $X \to (X+s)^p$:

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

- 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 $-2 \le p \le 3$.

7. 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.

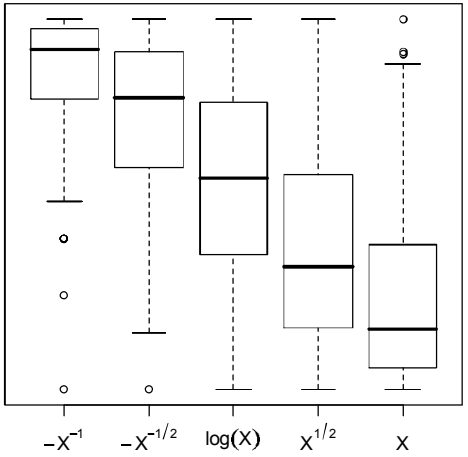

Figure 14. Boxplots for various transformations down the ladder of powers and roots for infant mortality in the UN datqa.

- 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).

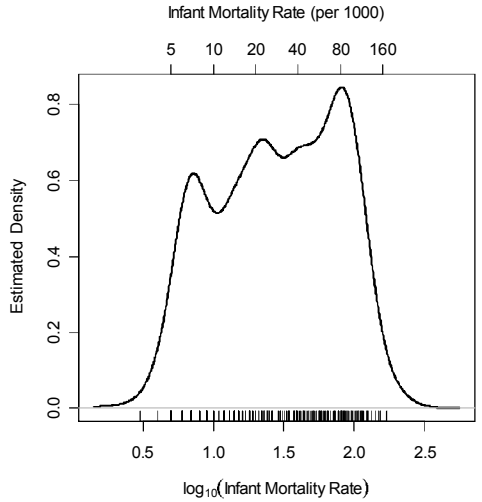

Figure 15. Adaptive-kernel density estimate for log-transformed infant mortality.

- 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.

8. 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;

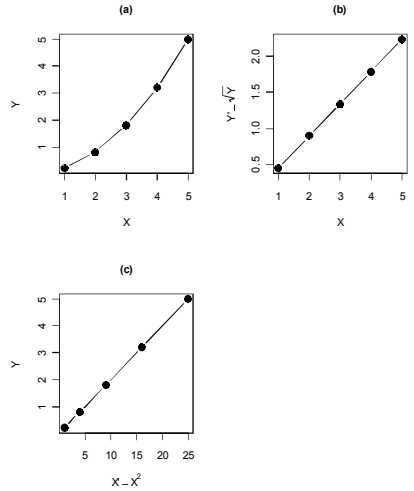

Figure 16. Transformating a nonlinear relationship (a) to linearity, (b) or (c).

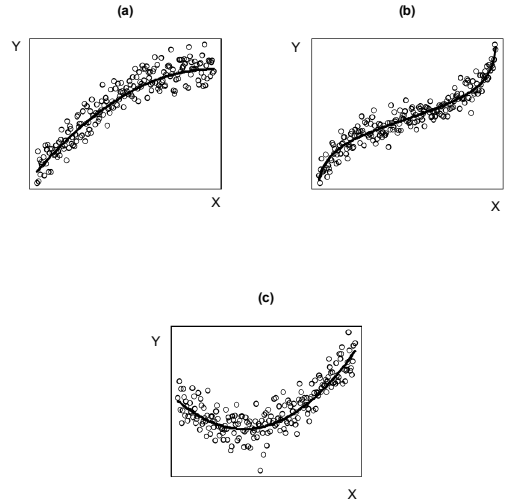

Figure 17. (a) A simple monotone relationship. (b) A monotone relationship that is not simple. (c) A simple nonmonotone relationship.

- 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).
- Recall the relationship between prestige and income for 102 Canadian occupations, shown again in Figure 19.
  - 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.

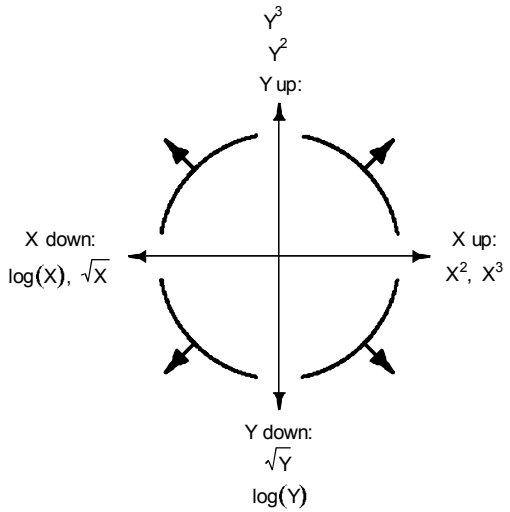

Figure 18. Mosteller and Tukey's bulging rule for selecting linearizing transformations.

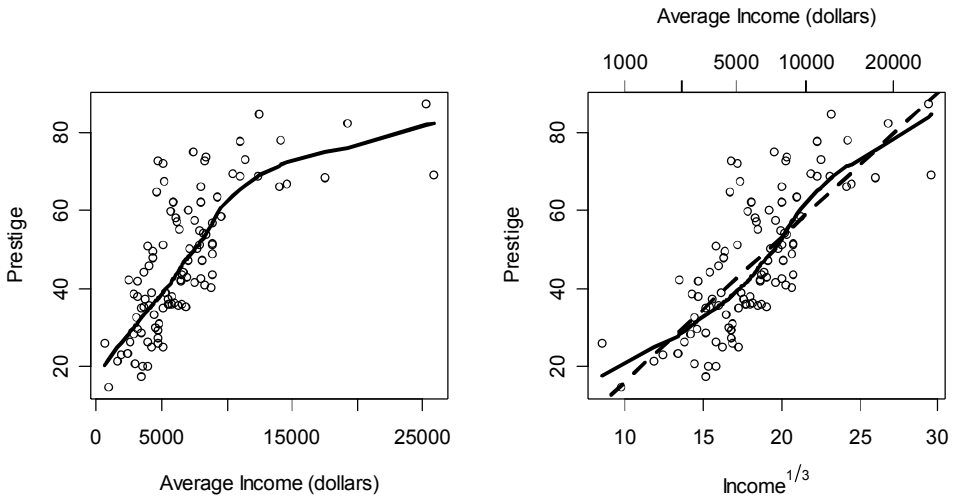

Figure 19. Transformating the relationship between prestige and income to (near) linearity: (left) original scatterplot; (right) with income transformed.

- 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.

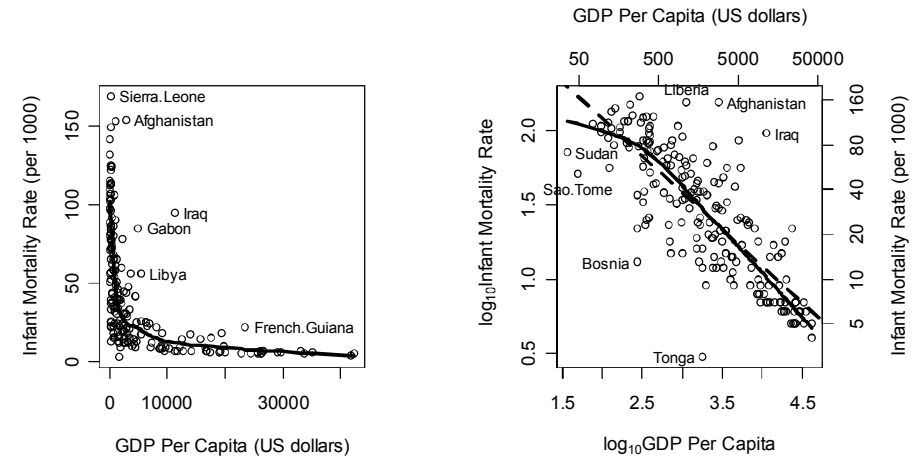

Figure 20. Transforming the relationship between infant mortality and GDP per capita.

- 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.'

9. 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 $p = 1 - b$.

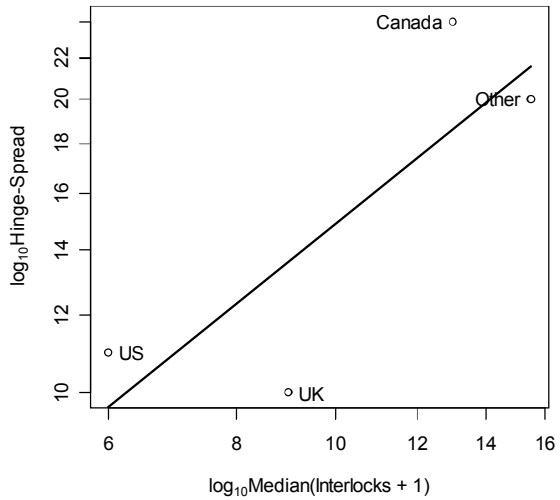

Figure 21. Spread-level plot for Ornstein's interlocking-directorate data.

  - 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.

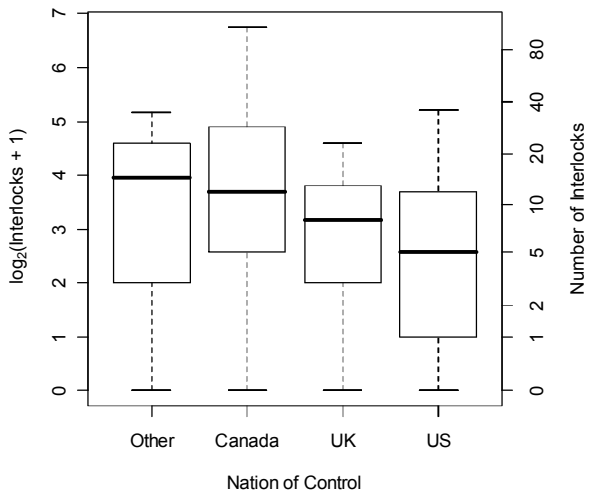

Figure 22. Ornstein's interlocking-directorate data, log-transforming interlocks (with a start of 1).

10. 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|00000000000000011111122223333444 4
 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).

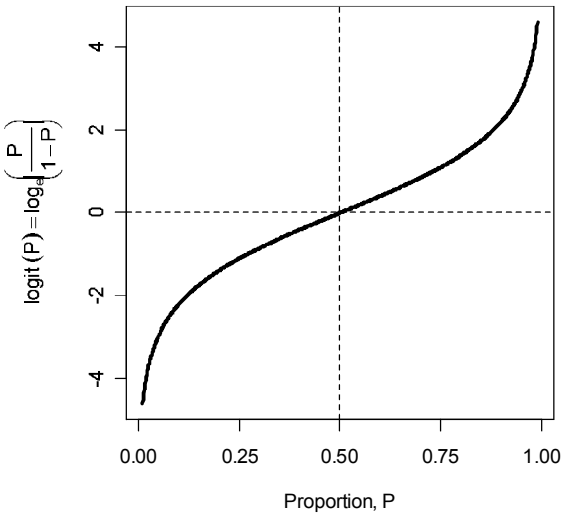

Figure 24. The logit transformation of a proportion.

  - 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.

11. Summary

- 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.