Sociology 740   John Fox

Lecture Notes

1. Introduction

Copyright © 2014 by John Fox

---

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.

▶ To introduce the R statistical computing environment.

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.

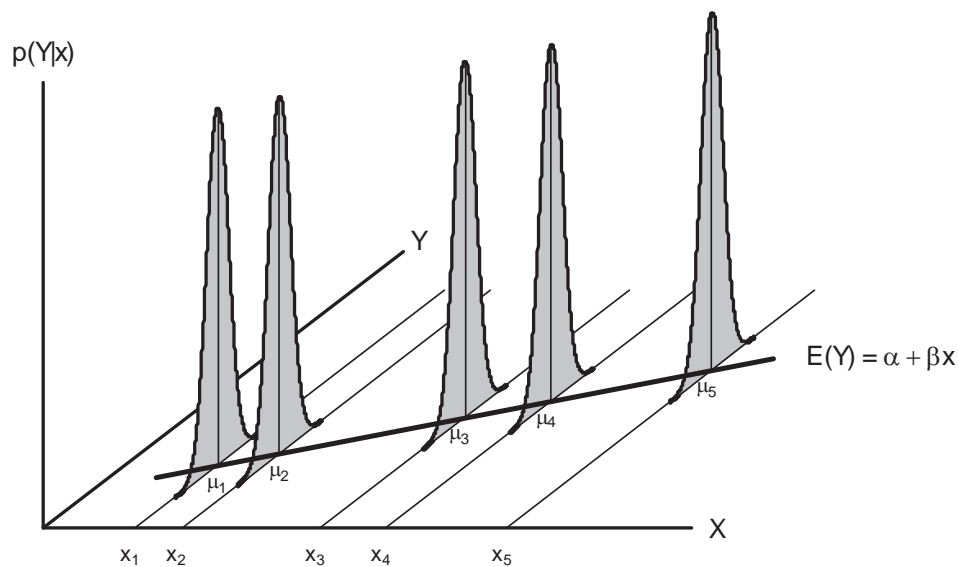▶ In contrast, I will pursue the notion of regression with as few preconceived assumptions as possible.

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.
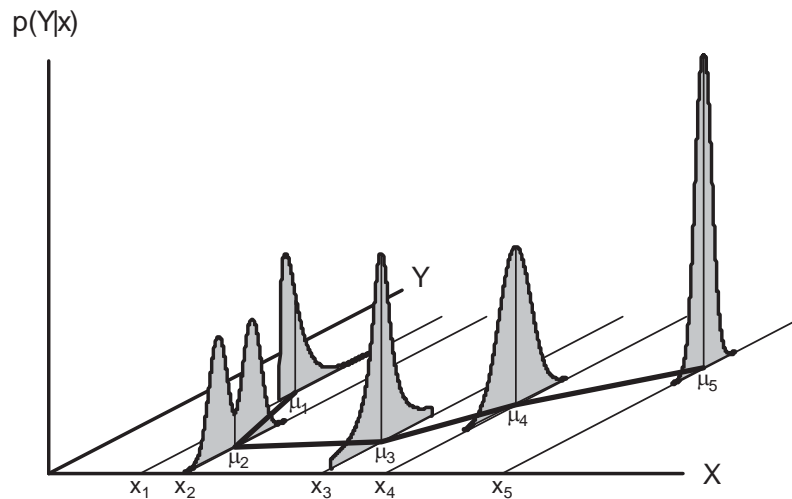
p(Y|x)

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.

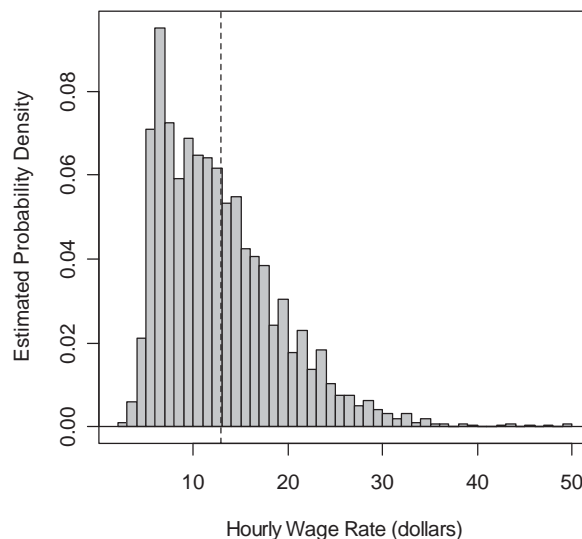- • Had we access to the entire population of employed Canadians, we could calculate $\mu|x$ directly.

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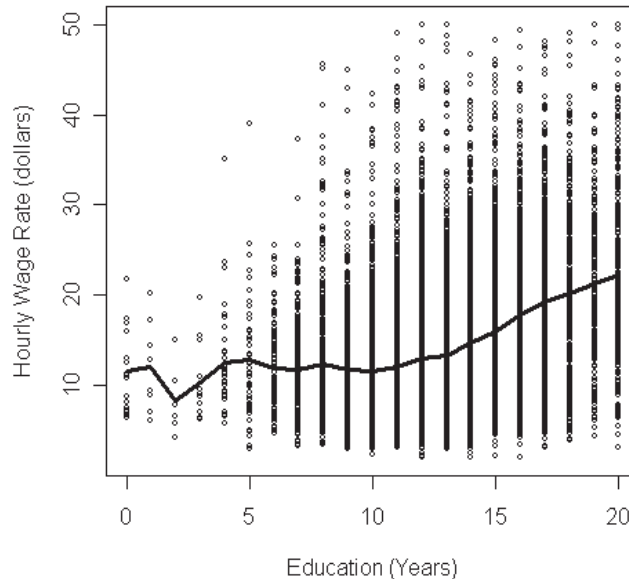
Figure 4. A scatterplot showing the relationship between hourly wages (in dollars) and education (in years) for a sample of 14,601 employed Canadians.

▶ 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. (The example isn't plausible since reported weight would not literally be continuous, but imagine that it is.)

- 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.
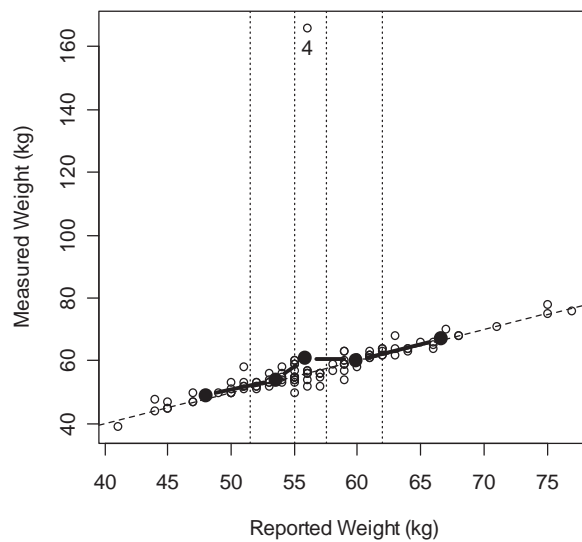

Figure 5. Naive nonparametric regression of measured on reported weight. The data are carved into fifths based on their X-values and the average Y in each fifth is calculated (the solid dots). Note the effect of the outlier (observation 4).
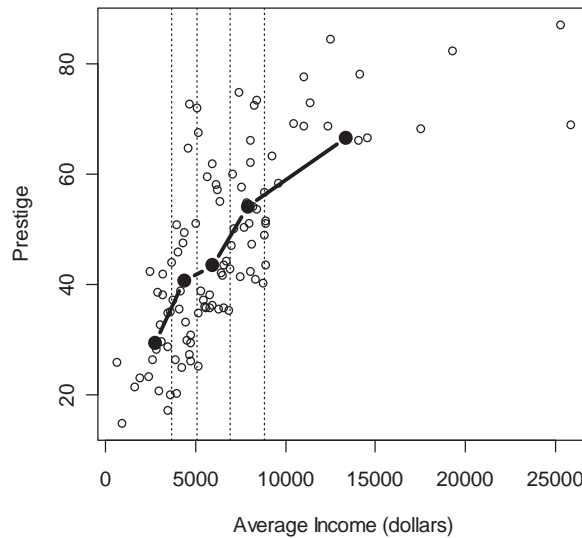
Figure 6. Naive nonparametric regression of occupational prestige on average income.

▶ 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).
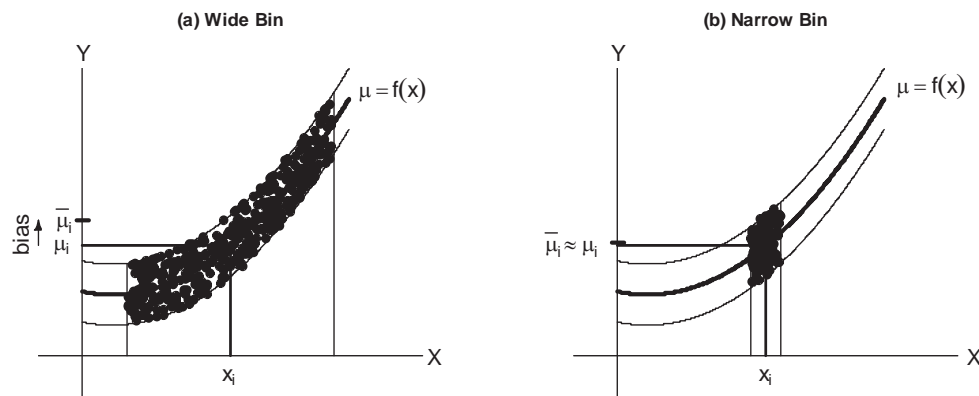
(a) Wide Bin

(b) Narrow Bin

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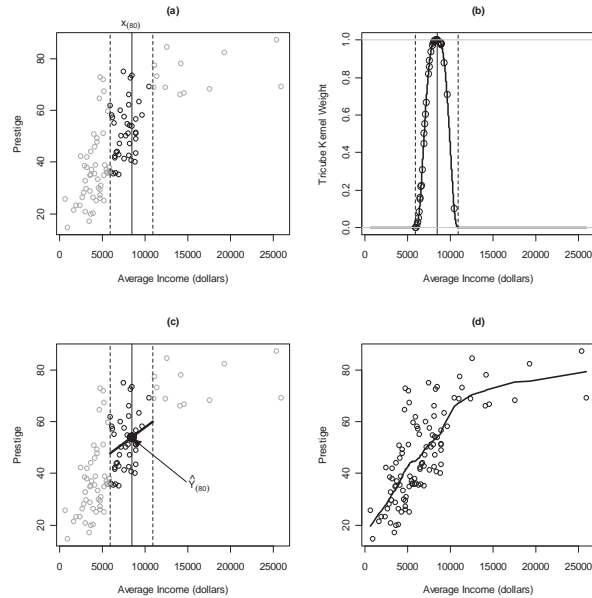
Figure 8.   Local linear regression of occupational prestige on income, showing the computation of the fit at $x_{(80)}$.

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.