POL 213 – Spring 2024

Quantitative Analysis in Political Science II

Introduction

Lauren Peritz

U.C. Davis

lperitz@ucdavis.edu

April 2, 2024

Table of Contents

Acknowledgments

- Thanks to Chris Hare, Scott MacKenzie, and Jeff Lewis for sharing their slides, notes, and exercises from similar courses.

Quantitative Methods in Political Science

- **POL211:** Basics of data analysis in social science. You covered: measures of central tendencies and dispersion, graphical summaries of variables, probability distributions, confidence intervals, t-tests, and hypothesis testing.

Quantitative Methods in Political Science

- **POL211:** Basics of data analysis in social science. You covered: measures of central tendencies and dispersion, graphical summaries of variables, probability distributions, confidence intervals, t-tests, and hypothesis testing.
- **POL212:** Computational social science. You covered: linear regression model (OLS), transforming data, estimation and interpretation, diagnostics of results and/or potential fixes for violations of assumptions, MC simulations, trees & bias-variance tradeoff.

Quantitative Methods in Political Science

- **POL211:** Basics of data analysis in social science. You covered: measures of central tendencies and dispersion, graphical summaries of variables, probability distributions, confidence intervals, t-tests, and hypothesis testing.
- **POL212:** Computational social science. You covered: linear regression model (OLS), transforming data, estimation and interpretation, diagnostics of results and/or potential fixes for violations of assumptions, MC simulations, trees & bias-variance tradeoff.
- **POL213:** Linear and nonlinear regression. We will cover: deeper dive into OLS derivation, mechanics, robustness, diagnostics as well as fundamentals for maximum likelihood estimation (MLE) of models for categorical dependent variables.

Instructors

Professor:	Lauren Peritz
Email:	lperitz@ucdavis.edu
Office Hours:	Tuesday 1:00 - 3:00 PM
Location:	Zoom (passcode *marmot*) or in Kerr 680.
Teaching Assistant:	Lily Huang
Email:	yslhuang@ucdavis.edu
Office Hours:	Wednesday 1:00 - 3:00 PM
Location:	Kerr 675
TA Session	TBD

Syllabus & Course Requirements

- Lectures & Reading
  - Actively participate in class and ask questions. Complete weekly reading assignments in advance of class. If you don't complete the assigned readings, you're basically guaranteed to learn nothing.

Syllabus & Course Requirements

- Lectures & Reading
  - Actively participate in class and ask questions. Complete weekly reading assignments in advance of class. If you don't complete the assigned readings, you're basically guaranteed to learn nothing.
- Problem Sets (60%).
  - Three problem sets to be submitted via **Canvas**. Late submissions penalized by one letter grade per day or fraction thereof.
  - All problem sets must be written in LaTeXor `Markdown`.
  - You may collaborate with colleagues but must acknowledge and submit final independent work.

Syllabus & Course Requirements

- Lectures & Reading
  - Actively participate in class and ask questions. Complete weekly reading assignments in advance of class. If you don't complete the assigned readings, you're basically guaranteed to learn nothing.
- Problem Sets (60%).
  - Three problem sets to be submitted via **Canvas**. Late submissions penalized by one letter grade per day or fraction thereof.
  - All problem sets must be written in LaTeXor `Markdown`.
  - You may collaborate with colleagues but must acknowledge and submit final independent work.
- Midterm Exam (15%).
  - In class, closed book and note. May 16.

Syllabus & Course Requirements

- Lectures & Reading
  - Actively participate in class and ask questions. Complete weekly reading assignments in advance of class. If you don't complete the assigned readings, you're basically guaranteed to learn nothing.
- Problem Sets (60%).
  - Three problem sets to be submitted via **Canvas**. Late submissions penalized by one letter grade per day or fraction thereof.
  - All problem sets must be written in LaTeXor `Markdown`.
  - You may collaborate with colleagues but must acknowledge and submit final independent work.
- Midterm Exam (15%).
  - In class, closed book and note. May 16.
- Final Exam (25%).
  - Take home, no consultation or collaboration. Due June 10.

Prerequisites

1. POL 211 and POL 212 or equivalent.
2. Familiarity with fundamentals of calculus and matrix algebra.
3. Basic proficiency in **R**.

   UCD Resources:
   https://github.com/dsidavis/RFundamentals

Note: On (2) and (3), you don't need to be a pro, just understand some fundamentals and be willing to put in the time to learn.

Textbooks

Required:

- Gelman, Andrew, Jennifer Hill, and Aki Vehtari. 2021. Regression and Other Stories. New York: Cambridge University Press. https://avehtari.github.io/ROS-Examples/
- Fox, John F. 2016. Applied Regression Analysis & Generalized Linear Models. (Third Edition) Thousand Oaks: SAGE Publications.

Textbooks

Recommended:

- Gelman, Andrew and Jennifer Hill. 2007. Data Analysis Using Regression and Multilevel/ Hierarchical Models. New York: Cambridge University Press. *Excellent discussion of motivations for, derivations and implementation of multilevel models, widely applicable to social science.*
- James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. An Introduction to Statistical Learning with Applications in R. New York: Springer. *If you're looking for a light, user-friendly introduction to machine learning methods (and a discussion of how regression analysis fits into this world), this is the book for you.*
- Monogan, James E. III. 2015. Political Analysis Using R London: Springer. (Available electronically from the UC Davis library's website.) *A political-science focused discussion of statistical fundamentals with R applications.*
- Christopher R. Bilder and Thomas M. Loughin. 2015. Analysis of Categorical Data with R. Boca Raton: Chapman and Hall/CRC. *Great reference for models of categorical data beyond what can be covered in POL213.*

Resources

Here is our course Canvas page:

https://canvas.ucdavis.edu/courses/892125

Please check it frequently for syllabus updates, lecture slides, problem sets, announcements and other resources.

TA sessions with Lily. This is an essential component of the course and it will help you solidify your understanding of the material. Regular study sessions, problem sets, and discussion are important for grasping difficult material such as applied statistics.

Policies and Logistics

Adjustments: The syllabus is subject to revision. Changes will be announced ASAP and individual needs will be accommodated to the greatest extent possible, subject to university policy and instructor discretion.

Academic Integrity: I strictly adhere to the UC Davis Code of Academic Conduct. You are responsible for understanding the standards. Any suspected misconduct will be reported to Student Judicial Affairs.

Disabilities: UC Davis is committed to educational equity in the academic setting, and in serving a diverse student body. I encourage all students who are interested in learning more about the Student Disability Center to contact them directly at sdc@ucdavis.edu, or 530.752.3184. If you require accommodations, please submit your SDC Letter of Accommodation to me as soon as possible.

Policies and Logistics

Attendance: You should attend all class meetings. If you are feeling unwell, please don't come to class sick. In this case, I will be happy to set up a Zoom link so that you can participate remotely.

Methods Track: Students on methods subfield track should plan to (1) engage course content at deeper level of methodological rigor (be able to use linear algebra + calculus) and (2) answer all the "bonus" problems on the assignments.

Table of Contents

Regression Analysis:
Parametric (Linear) and Nonparametric Approaches

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 both with a parametric (e.g. linear) and nonparametric approach. The latter does not make strong prior assumptions about the structure of the data.
- To introduce or review basic concepts: skewness, sampling variance, bias, outliers, etc.

What is Regression Analysis?

Regression analysis traces the distribution of a response (dependent) variable (Y) as a function of one or more explanatory (predictor) variables ($X_1, X_2, ...X_k$).

$$p(y|x_1, x_2, ...x_k) = f(x_1, x_2, ...x_k)$$

$p(y|x_1, x_2, ...x_k)$ represents the probability (or for a continuous Y, the probability density) of observing the specific value y of the response variable conditional upon a set of specific values $(x_1, x_2, ...x_k)$ of the explanatory variables.

Example: Suppose Y is individual income and X's are characteristics such as education, age, geographical location, and so on. Let's restrict our example to quantitative X's: years of education and age.

What is Regression Analysis?

Most discussions of regression analysis begin by assuming:

- the conditional distribution of the response variable $p(Y|x_1, ...x_k)$, is a normal distribution
- the variance of Y condition on the X's, denoted σ^2 is everywhere the same regardless of the specific values of $x_1, ...x_k$
- that the expected value (mean) of Y is a linear function of the X's

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

- and that there is independent random sampling.

These assumptions lead to the *linear least squares estimation*. **Figure 1** illustrates this for a single X variable.

In contrast, we can pursue the notion of regression with as few assumptions as possible.

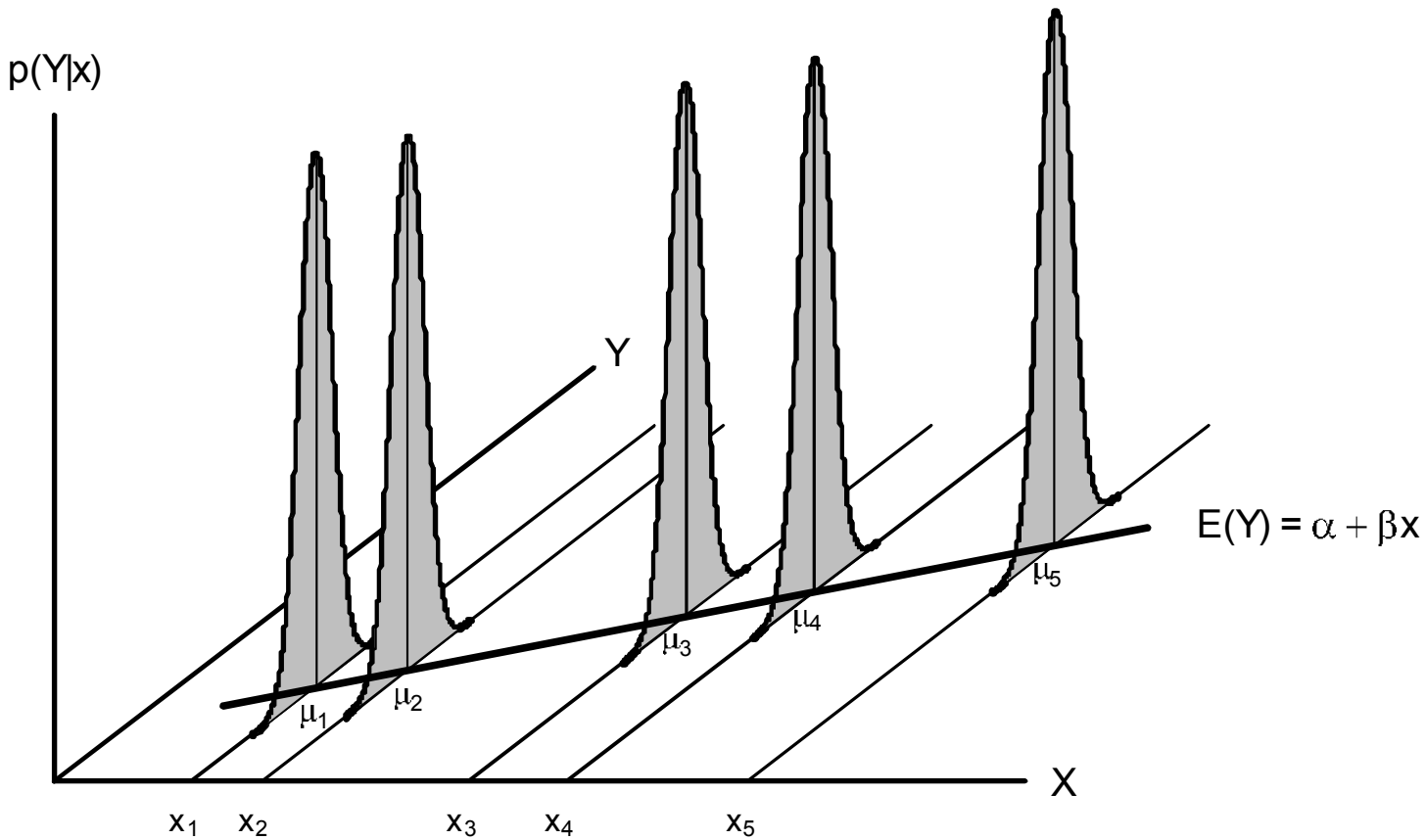


Figure 1. The usual assumptions: linearity, constant variance, and normality, for a single X.

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 multi-modal then it is intrinsically unreasonable to summarize its center with a single number.
- **Non-normal:** If the conditional distribution is not normal (e.g. heavy tails) 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 varaince of Y changes with the values of the X's then the efficiency of the least squares estimator is compromised.

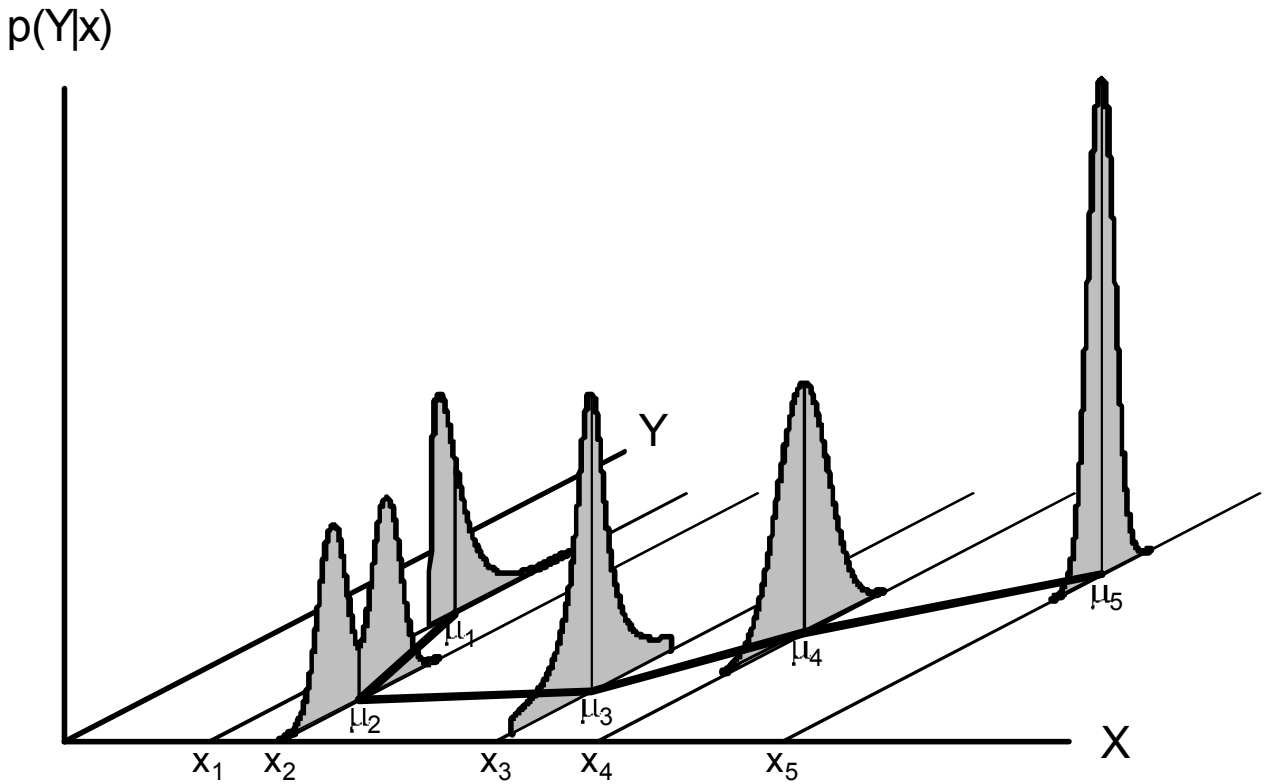


Figure 2. How the usual regression assumptions can fail.

Linear or Non-Linear

- **Linearity:** The linear regression model can be a great choice for efficiently modeling correlations in your data. However, it requires several assumptions and violations of those assumptions are common in social science data.
- **Non-Linearity:** Although we often expect that the values of Y will increase or decrease with some X, there is rarely a good reason to assume *a priori* that the relationship is linear; this problem is compounded when there are several X's.

This is not to say that linear regression analysis lacks practical use. Much of this course is devoted to the exposition of linear models. However, it is prudent to begin with an appreciation of the linear model's limitations, since their effective use in data analysis often depends on adapting to these limitations.

Naive Nonparametric Regression

Example:

Suppose we have a large random sample of employed Canadians that includes those people's hourly wages and years of education.

- We could 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 $\bar{Y}|x$, as in **Figure 4**.
  - But using 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.

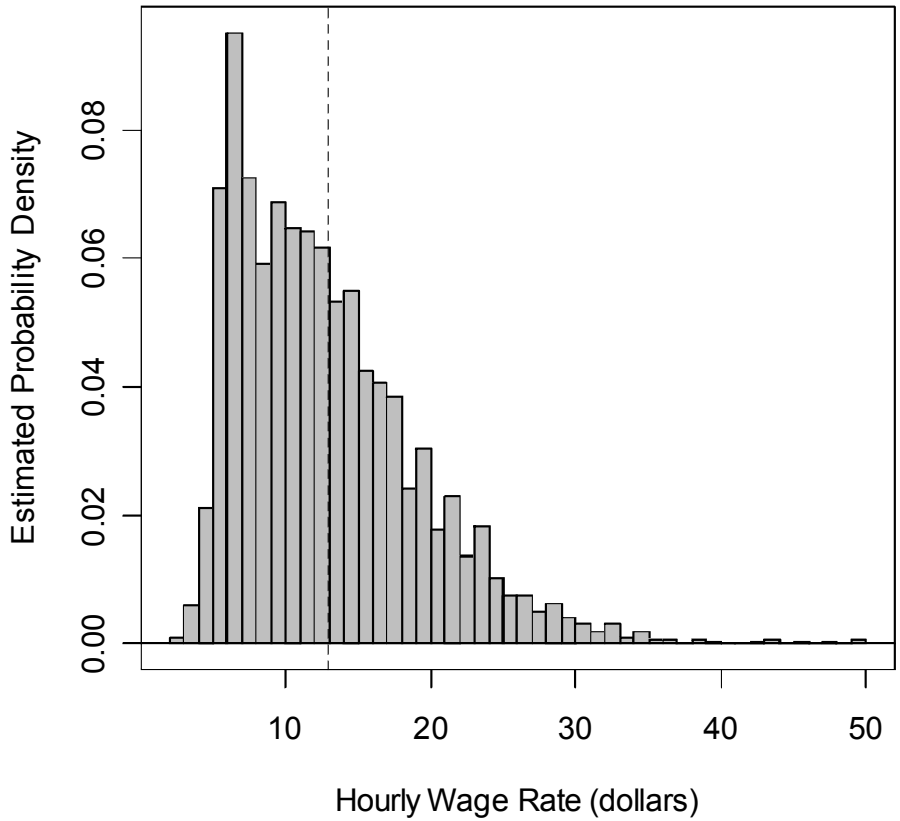


Figure 3. The conditional distribution of hourly wages for the 3384 employed Canadians in the SLID who had 12 years of education. The broken vertical line shows the conditional mean wages.

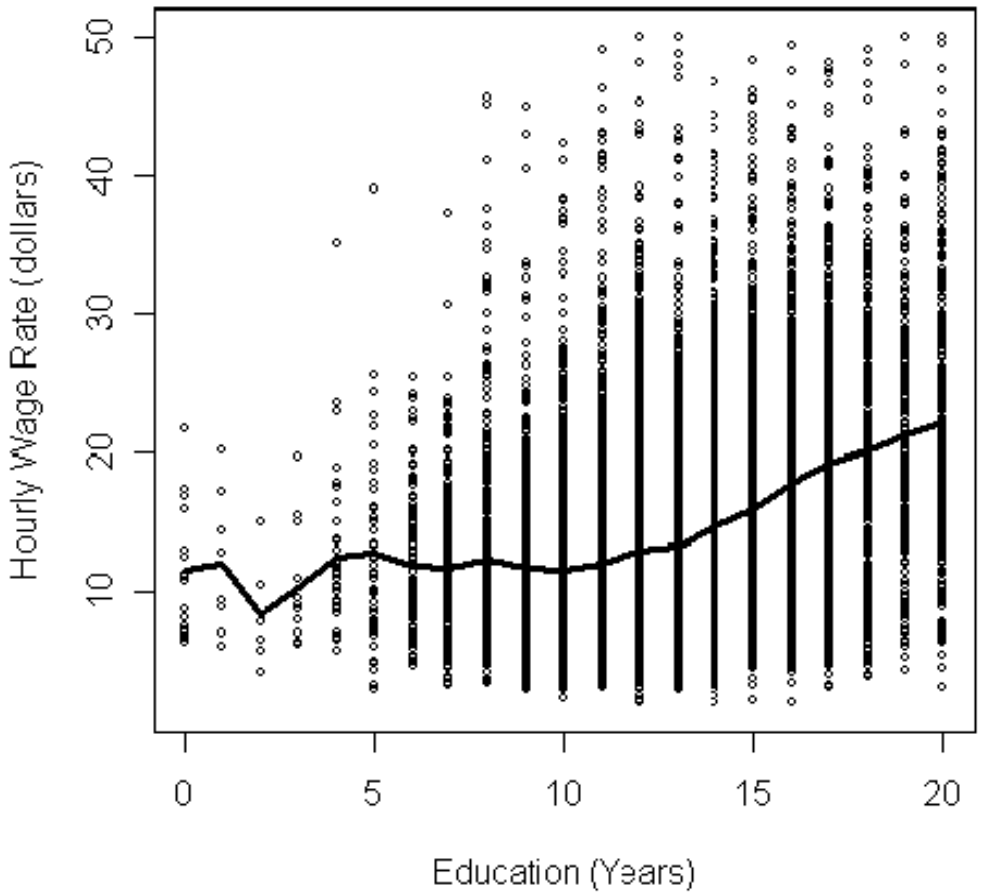


Figure 4. A scatterplot showing the relationship between hourly wages (in dollars) and education (in years) for a sample of 14,601 employed Canadians.

Imagine now that X along with Y is a continuous variable.

- For example, X is the reported weight in kg for each of a sample of individuals and Y is their measured weight, again in kg.
- 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.

$$\mu \equiv \mathbb{E}(Y|\mu) = f(x)$$

- Even if the sample is large, replicated values of X will be rare because the variable is continuous. But for a large sample we can dissect the range of X into many narrow class intervals (i.e. 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 observations (N small) 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 bins, each containing approx. 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 $\bar{Y}$ and the explanatory-variable means $\bar{X}$ in the five bins.

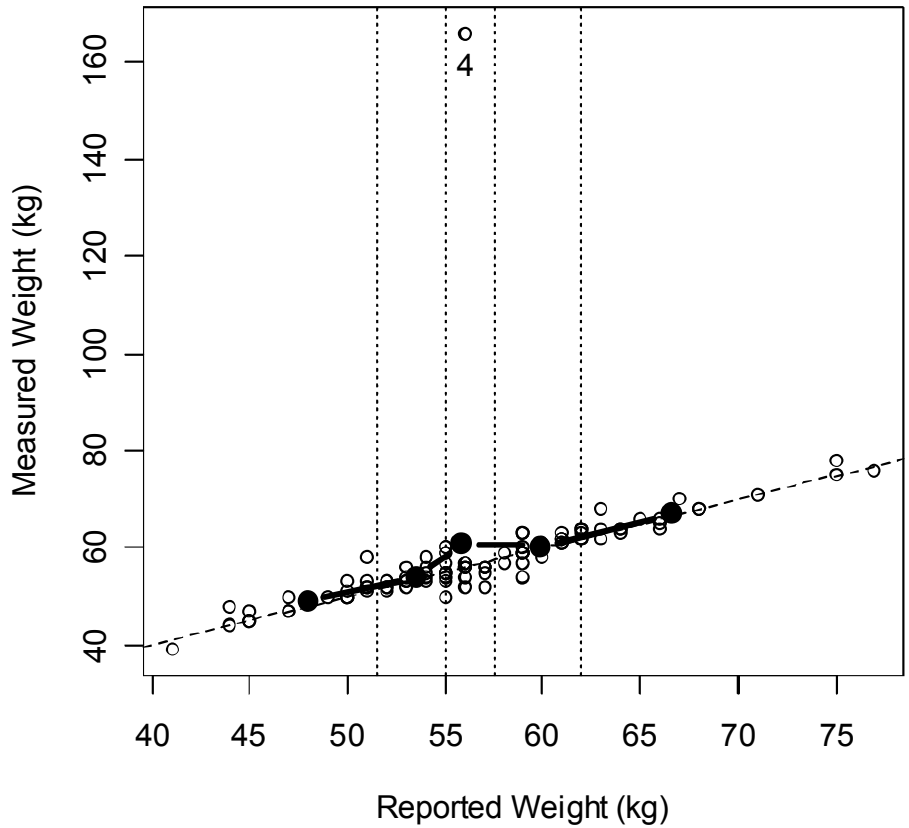


Figure 5. Naive nonparametric regression of measured on reported weight. The data are carved into fifths based on their X-values and the average Y in each fifth is calculated (the solid dots). Note the effect of the outlier (observation 4).

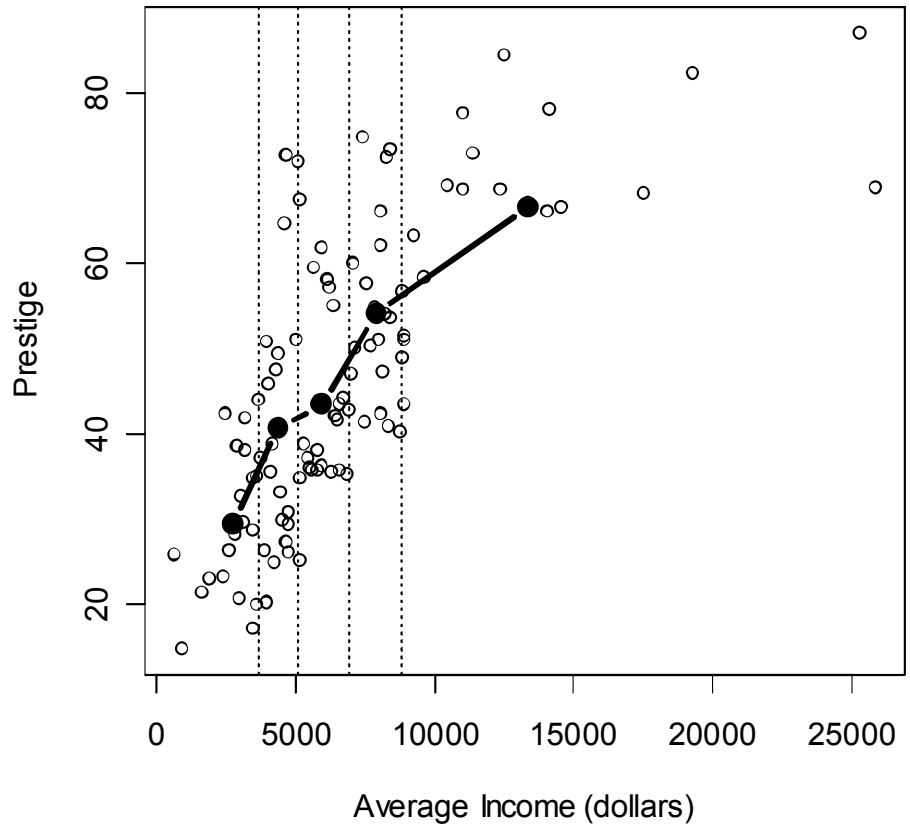


Figure 6. Naive nonparametric regression of occupational prestige on average income.

There are two sources of error in this simple procedure of binning and averaging:

- **Sampling error (variance):** The conditional sample means $\bar{Y}$ 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 ($\bar{\mu}_i$) 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**).

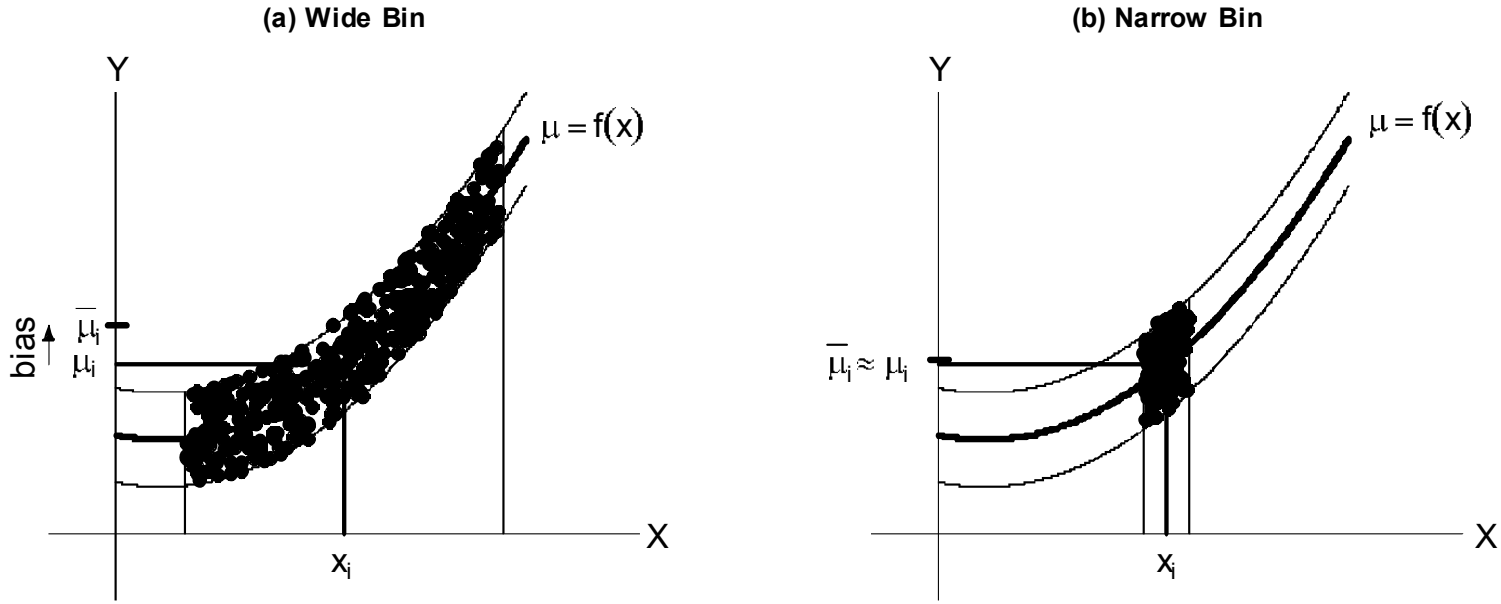


Figure 7. A narrow bin (b) generally produces less bias in estimating the regression curve than a wide bin (a).

- As is typically the case in statistical estimation, reducing bias and reducing sampling variance work in competition
  - Only if we select a very large sample can we have our cake and eat it too! (bias-variance tradeoff)
  - 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 $x \to \infty$), 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.
  - For example: suppose we have 3 discrete explanatory variables, each with ten values. Then there are $10^3 = 1000$ combinations of the 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, we would need a very large sample to calculate the conditional means of Y with sufficient precision.
  - The situation is worse when the X's are continuous, since binning the range of each X into as few as ten class intervals might introduce bias.
  - The problem grows exponentially as the number of X's increases. Statisticians refer to the intrinsic sparness of multivariate data as the 'curse of dimensionality.'

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 $\hat{Y}$ corresponding to any X-value in the range of the data – usually, at the data-valuesx_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:

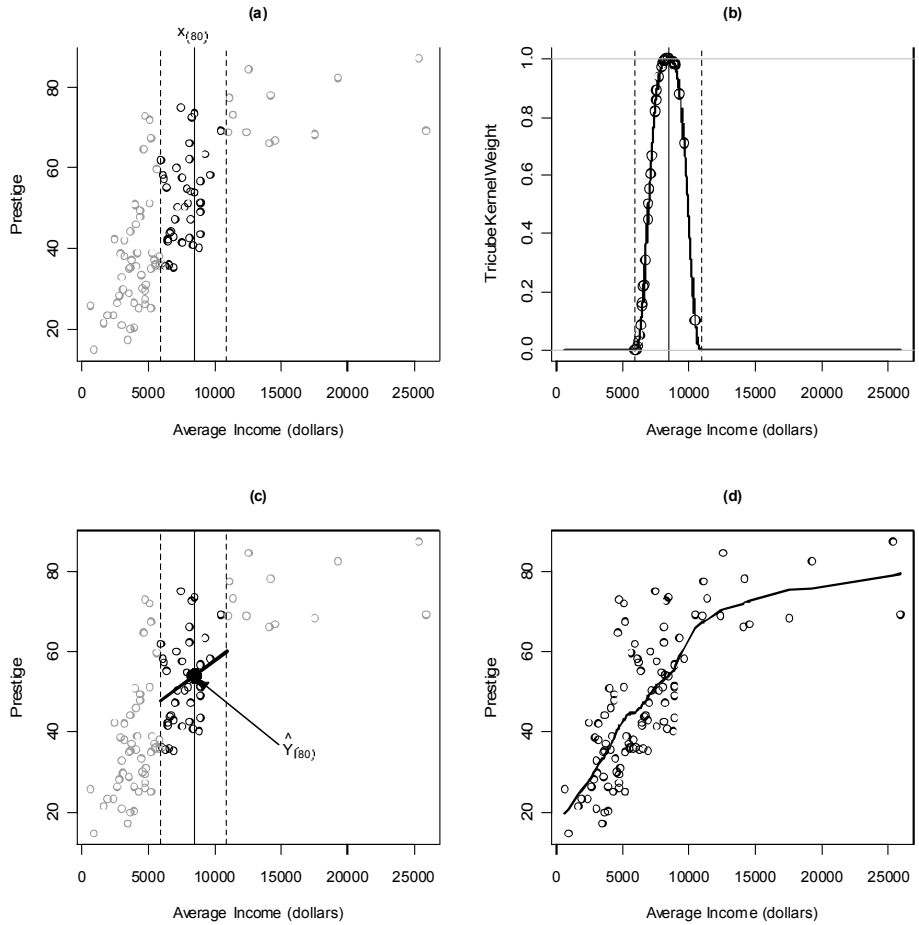


Figure 8. Local linear regression of occupational prestige on income, showing the computation of the fit at $x_{(80)}$.

Local Regression

1. *Choose the span:* Select a fraction of the data $0 < s \leq 1$ (span of 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.
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 neighbors 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)):

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

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

Local Regression

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

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

One regression equation is fit, and one fitted value is calculated, for each $i = 1, \ldots n$, as shown in panel (c). Connecting these fitted value produces the nonparametric regression smooth (panel (d))

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 the distribution such as its mean, as a function of X's.

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 the distribution such as its mean, as a function of 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 X's. When the explanatory variables are continuous, we can proceed similarly by dissecting the X's into a large number of narrow bins.

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 the distribution such as its mean, as a function of 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 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.

Table of Contents

Linear Regression with Single Variable

How it Works

(Source: Fox Chapter 5)

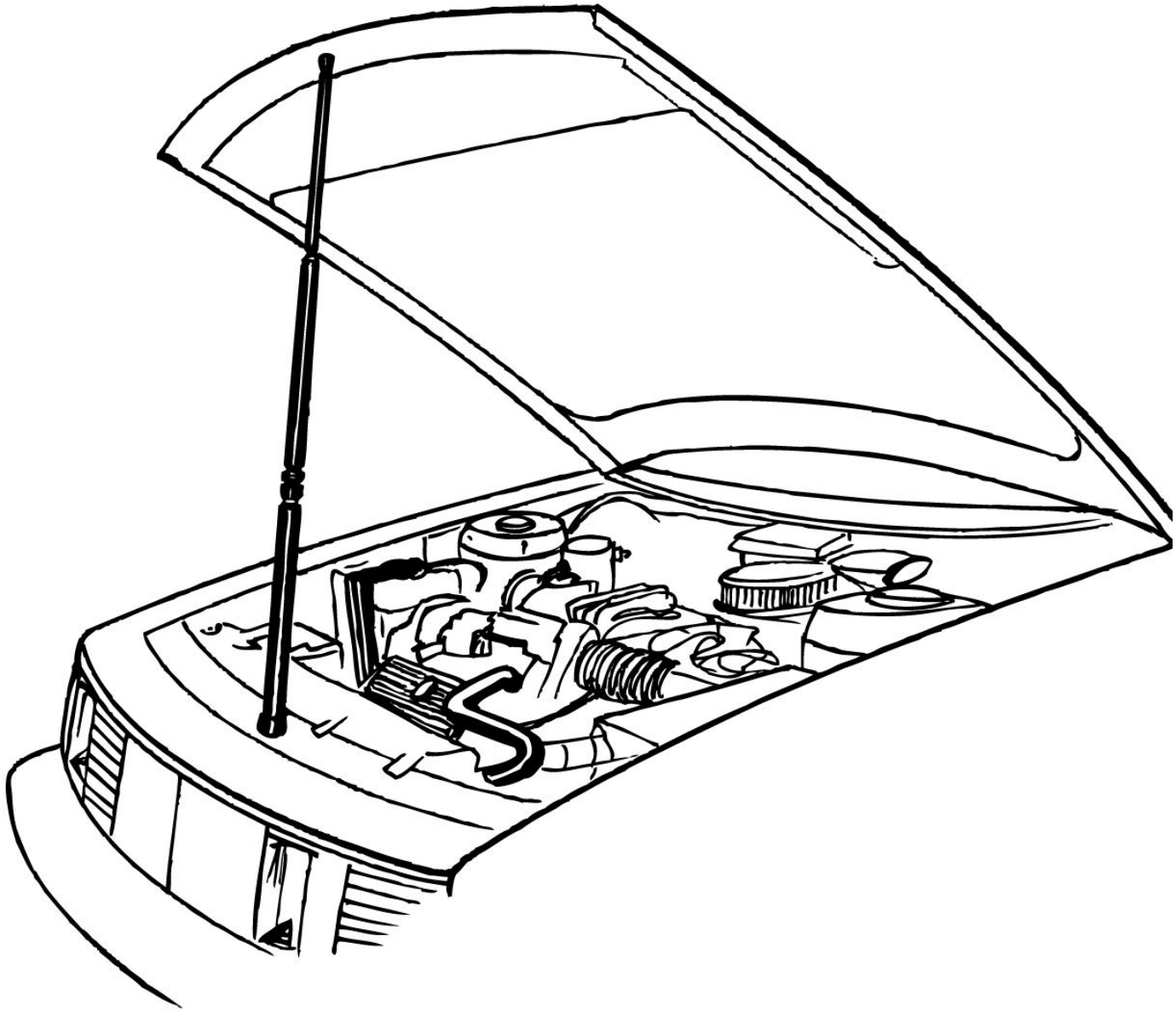
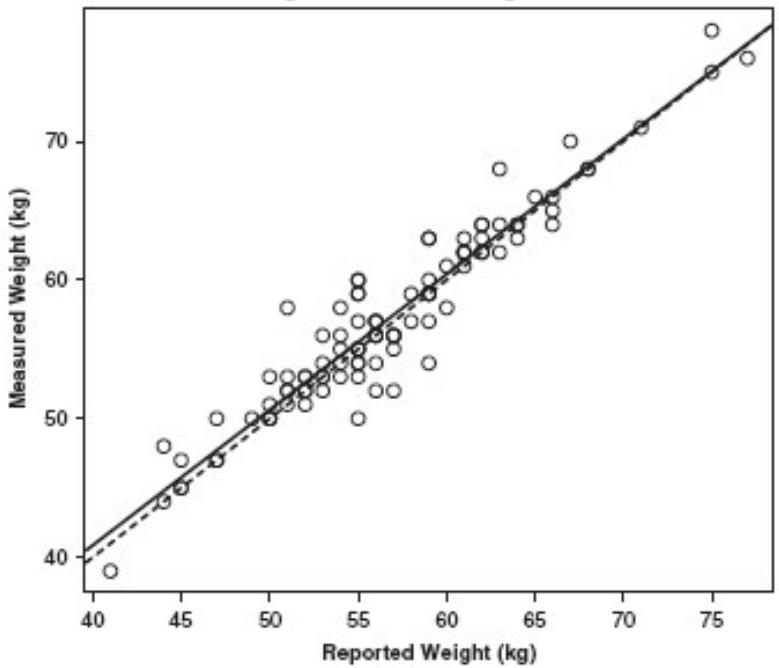


Figure 5.1 Scatterplot of Davis's data on the measured and reported weight of 101 women. The solid line gives the least-squares fit; the broken line is $Y = X$. Because weight is given to the nearest kilogram, both variables are discrete, and some points are overplotted.



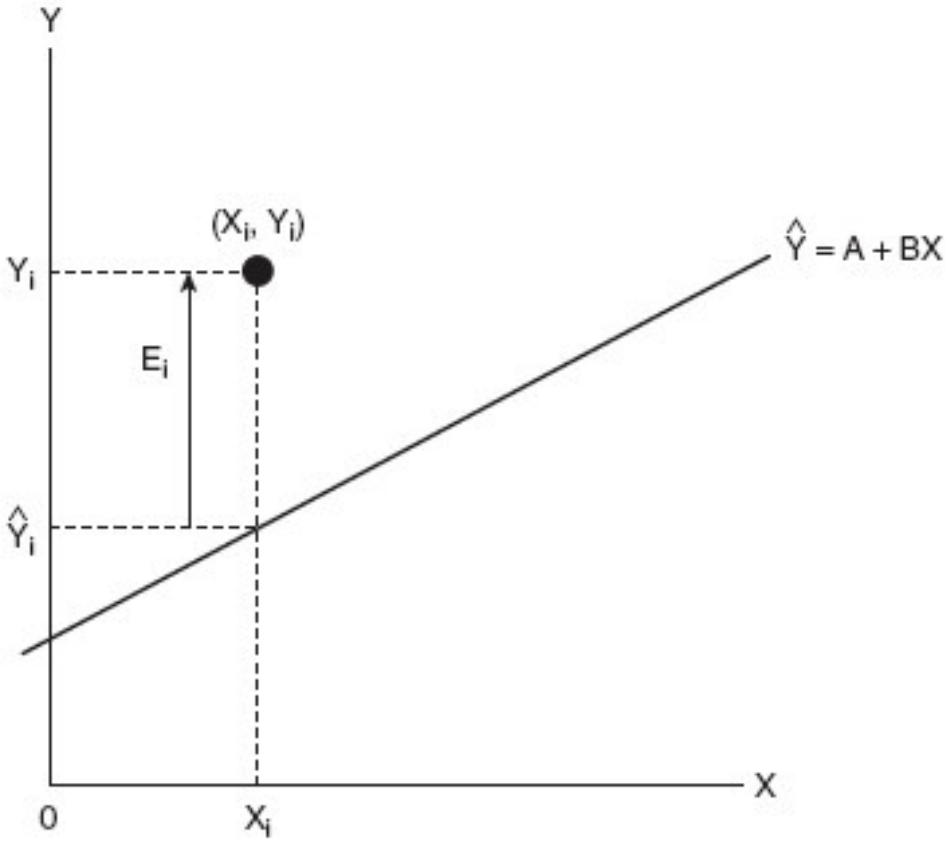
Least Squares Fit

Recall, the line relating Y and X can be written as $Y = A + BX$. But because no line can pass perfectly through all the data points, we need to account for the *fitted value* ($\hat{Y}_i$) and the *residual* (E_i) corresponding to each observation i.

$$Y_i = A + BX_i + Ei = \hat{Y}_i + Ei$$

The essential geometry is in **Fig. 5.2**. It shows that the residual is the vertical distance between the point and the fitted line. $E_i = Y_i - \hat{Y}_i = Y_i - (A + BX_i)$. A line that fits the data well makes the residuals small across all the observations. But what is meant by "small"?

Figure 5.2 Linear regression of Y on X, showing the residual E_i for the ith observation.



Least Squares Fit

A mathematically tractable option to find the "smallest" discrepancy between all the the data points and the best fit line is to minimize the sum of the squared residuals.

Any line through the means of the variables $(\bar{X}, \bar{Y})$ has $\sum E_i = 0$. Such a line satisfies $\bar{Y} = A + B\bar{X}$. Subtracting this off of our previous equation gives:

$$Y_i - \bar{Y} = A + B(X_i - \bar{X}) + Ei$$

Then we sum over all the observations:

$$\sum_{i=1}^{n} E_i = \sum(Y_i - \bar{Y}) - B\sum(X_i - \bar{X}) = 0 - B \times 0 = 0$$

We use the *least square criterion*: that is, find the A (intercept) and B (slope) to minimize the sum of squared residuals, $\sum E_i^2$, over all the observations.

Least Squares Fit is an Optimization Problem

The least squares fit is an optimization problem for the data at hand where we are assuming the functional form is a line and the goal is to choose parameters (intercept A and slope B) to make the discrepancy between the line and the data as small as possible.

How do we find A and B to minimize this expression over all the i observations?

$$\begin{aligned} S(A,B) &= \sum_{i=1}^{n} E_i^2 \\ &= \sum_{i=1}^{n} (Y_i - A - BX_i)^2 \end{aligned}$$

The most direct approach is to take the partial derivatives with respect to A and B, set them equal to zero, and then solve the system of simultaneous linear equations for the least squares parameters A and B.

$$\frac{\partial S(A,B)}{\partial A} = \sum (-1)(2)(Y_i - A - BX_i) = 0$$
$$\frac{\partial S(A,B)}{\partial B} = \sum (-X_i)(2)(Y_i - A - BX_i) = 0$$

Least Squares Fit is an Optimization Problem

We then solve for all n observations:

$$An + B\sum X_i = \sum Y_i$$

$$A\sum X_i + B\sum X_i^2 = \sum X_i Y_i$$

Rearranging and substituting yields:

$$A = \bar{Y} - B\bar{X}$$

$$B = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

Least Squares Fit is an Optimization Problem

What does each part of the solution actually mean?

$$A = \bar{Y} - B\bar{X}$$

$$B = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

Intercept: the least squares line passes through the point of means of the two variables.

Least Squares Fit is an Optimization Problem

What does each part of the solution actually mean?

$$A = \bar{Y} - B\bar{X}$$

$$B = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

Slope: Take the difference of each value of Y from the mean Y multiplied by the difference of each X from the mean X. Then a sum of all those products, divided by the sum of squared deviations of X.

This tells us the overall average change in Y over change in X.

Least Squares Fit is an Optimization Problem

$$B = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

Intuition: how much do the two variables (X and Y) go together (covariance) divided by how much does the independent variable (X) vary (variance).

Evaluating Least Squares Fit

How well did we do with our solution to the optimization problem?

We often want to know how closely the line fits the scatter of points. The *residual standard error* for degrees of freedom $n - 2$ is:

$$S_E = \sqrt{\frac{\sum E_i^2}{n-2}}$$

Evaluating Least Squares Fit

We also want to know to what degree our predictions of Y improve when we base those predictions on the linear relationship between Y and X. A relative index of fit requires a baseline – how well can Y be predicted if X is disregarded? Recall, our predicted values are $\hat{Y}_i$. Then we can define:

The Total Sum of Squares (TSS):

$$\sum E_i'^2 = \sum (Y_i - \bar{Y})^2$$

where E_i' is for the *null model*, where we didn't specify any relationship between Y and X.

The Residual Sum of Squares (RSS)

$$\sum E_i^2 = \sum (Y_i - \hat{Y})^2$$

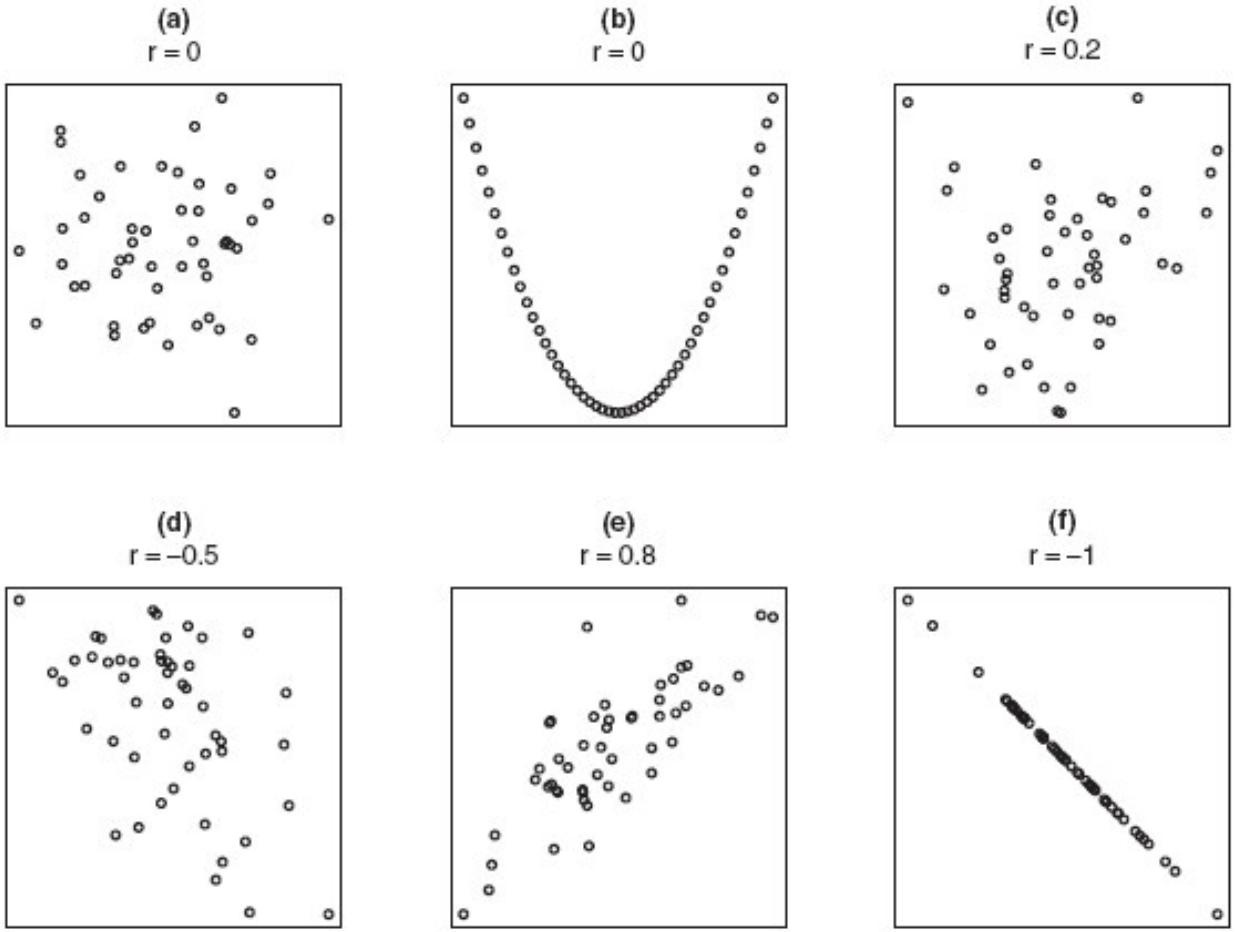
The difference between these two quantities tells us the reduction in squared error attributed to the regression fit. It is called the *regression sum of squares*: $RegSS \equiv TSS - RSS$. The ratio of $RegSS$ to TSS is the proportional reduction in squared error:

$$r^2 \equiv \frac{RegSS}{TSS}$$

Evaluating Least Squares Fit

Some examples of different model fits with correlation coefficient $r = \sqrt{\frac{RegSS}{TSS}}$

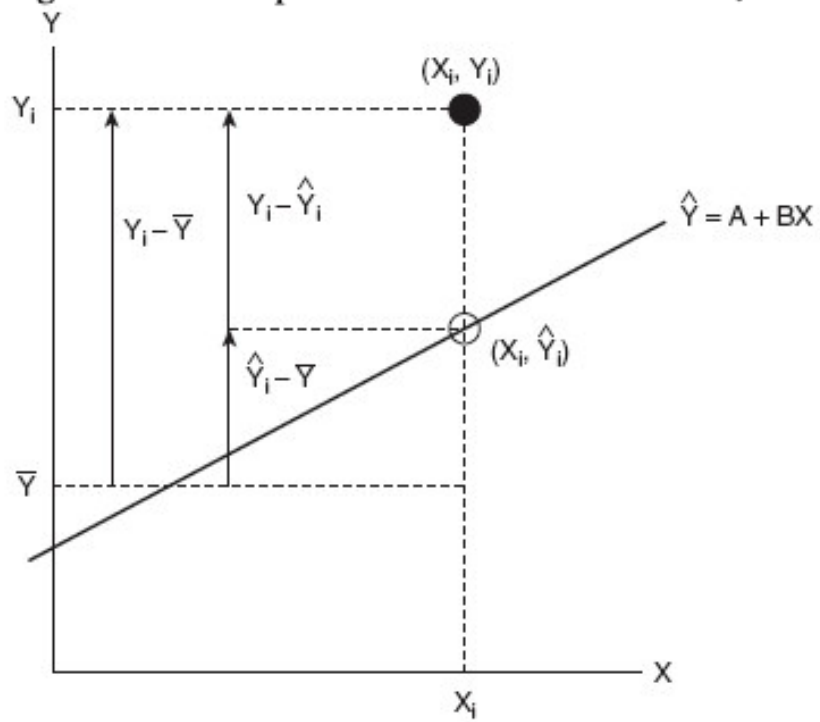
Figure 5.4 Scatterplots illustrating different levels of correlation: $r = 0$ in both (a) and (b), $r = .2$ in (c), $r = -.5$ in (d), $r = .8$ in (e), and $r = -1$ in (f). All the data sets have $n = 50$ observations. Except in panel (b), the data were generated by sampling from bivariate normal distributions.



Evaluating Least Squares Fit

The total variation can be decomposed into “explained” and “unexplained” components, paralleling the decomposition of each observation into a fitted value and a residual. Fig 5.5 illustrates. This decomposition is called the *analysis of variance* for the regression: TSS = RegSS-RSS.

Figure 5.5 Decomposition of the total deviation $Y_i - \overline{Y}$ $Y_i - \overline{Y}$ into components $Y_i - \hat{Y}_i$ $Y_i - \hat{Y}_i$ and $\hat{Y}_i - \overline{Y}$ $\hat{Y}_i - \overline{Y}$.



5.1.2 Simple Correlation

Linear Regression with Single Variable

Summary:

- Some data are adequately summarized by linear least square regression. The distribution of the data is a key factor.
- LS is fundamentally an optimization problem: choose parameters that define a line (A and B) such that the discrepancy (residual) between the line and the data are minimized. Our preferred criteria is minimizing the sum of squared residuals. We used a tiny bit of calculus in this process.
- We need ways of summarizing how good the fit was including whether the model we created did better than the null in explaining variation in the data. Importantly, the r^2 was defined.

As we extend the least squares regression model to multiple variables, the calculations become very cumbersome, as does our optimization procedure. We can make this easier on ourselves by using matrix algebra. That's for next time....