Sociology 740 John Fox

Lecture Notes

3. Linear Least-Squares Regression

1. Goals:

- To review/introduce the calculation and interpretation of the least-squares regression coefficients in simple and multiple regression.
- To review/introduce the calculation and interpretation of the regression standard error and the simple and multiple correlation coefficients.
- To introduce and criticize the use of standardized regression coefficients
- Time and interest permitting: To introduce matrix arithmetic and least-squares regession in matrix form.

2. Introduction

- Despite its limitations, linear least squares lies at the very heart of applied statistics:
  - Some data are adequately summarized by linear least-squares regression.
  - The effective application of linear regression is expanded by data transformations and diagnostics.
  - The *general* linear model — an extension of least-squares linear regression — is able to accommodate a very broad class of specifications.
  - Linear least-squares provides a computational basis for a variety of generalizations (such as *generalized* linear models).
- This lecture describes the mechanics and descriptive interpretation of linear least-squares regression.

3. Simple Regression

3.1 Least-Squares Fit

- Figure 1 shows Davis's data on the measured and reported weight in kilograms of 101 women who were engaged in regular exercise.
  - The relationship between measured and reported weight appears to be linear, so it is reasonable to fit a line to the plot.
- Denoting measured weight by Y and reported weight by X, a line relating the two variables has the equation $Y = A + BX$.
  - No line can pass perfectly through all of the data points. A residual, E, reflects this fact.
  - The regression equation for the ith of the $n = 101$ observations is

$$\begin{aligned} Y_i &= A + BX_i + E_i \\ &= \widehat{Y}_i + E_i \end{aligned}$$

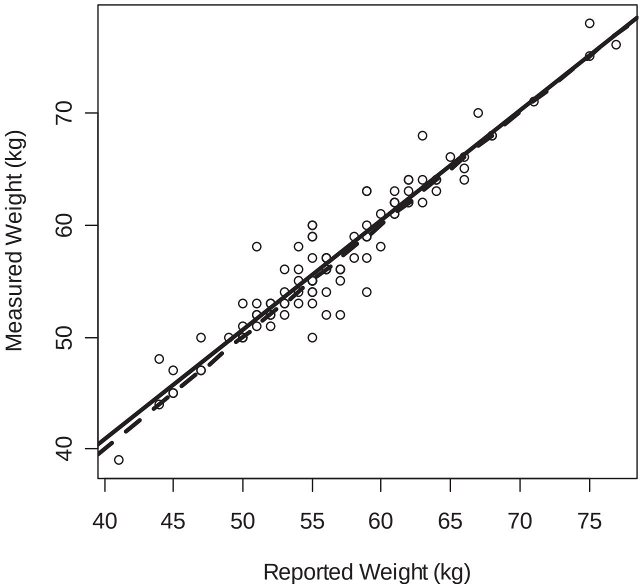

Figure 1. A least-squares line fit to Davis's data on reported and measured weight. (The broken line is the line $Y = X$.) Some points are over-plotted.

- The residual

  $$E_i = Y_i - \widehat{Y}_i = Y_i - (A + BX_i)$$

  is the signed vertical distance between the point and the line, as shown in Figure 2.

▶ A line that fits the data well makes the residuals small.

- Simply requiring that the sum of residuals, $\sum_{i=1}^{n} E_i$, be small is futile, since large negative residuals can offset large positive ones.
- Indeed, any line through the point $(\overline{X}, \overline{Y})$ has $\sum E_i = 0$.

▶ Two possibilities immediately present themselves:

- Find A and B to minimize the absolute residuals, $\sum |E_i|$, which leads to least-absolute-values (LAV) regression.
- Find A and B to minimize the squared residuals, $\sum E_i^2$, which leads to least-squares (LS) regression.

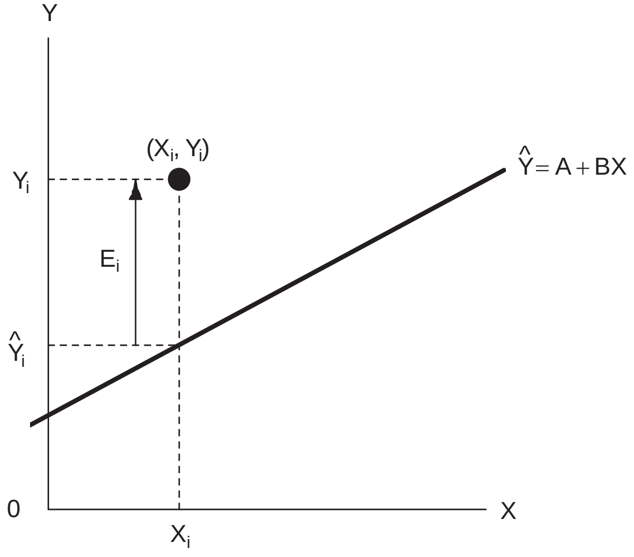

Figure 2. The residual E_i is the signed vertical distance between the point and the line.

- In least-squares regression, we seek the values of A and B that minimize:

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

- *For those with calculus:*
  - The most direct approach is to take the partial derivatives of the sum-of-squares function with respect to the coefficients:

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

- Setting these partial derivatives to zero yields simultaneous linear equations for A and B, the *normal equations* for simple regression:

$$\begin{aligned} An + B\sum X_i &= \sum Y_i \\ A\sum X_i + B\sum X_i^2 &= \sum X_i Y_i \end{aligned}$$

- Solving the normal equations produces the least-squares coefficients:

$$\begin{aligned} A &= \overline{Y} - B\overline{X} \\ B &= \frac{n\sum X_iY_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2} = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sum(X_i - \overline{X})^2} \end{aligned}$$

  - The formula for A implies that the least-squares line passes through the point-of-means of the two variables. The least-squares residuals therefore sum to zero.
  - The second normal equation implies that $\sum X_iE_i = 0$; similarly, $\sum \widehat{Y}_iE_i = 0$. These properties imply that the residuals are uncorrelated with both the X's and the $\widehat{Y}$'s.

▶ For Davis's data on measured weight (Y) and reported weight (X):

$$\begin{aligned} n &= 101 \\ \overline{Y} &= \frac{5780}{101} = 57.23 \\ \overline{X} &= \frac{5731}{101} = 56.74 \\ \sum(X_i - \overline{X})(Y_i - \overline{Y}) &= 4435. \\ \sum(X_i - \overline{X})^2 &= 4539. \\ B &= \frac{4435}{4539} = 0.9771 \\ A &= 57.23 - 0.9771 \times 56.74 = 1.789 \end{aligned}$$

- The least-squares regression equation is

$$\widehat{\text{Measured Weight}} = 1.79 + 0.977 \times \text{Reported Weight}$$

▶ Interpretation of the least-squares coefficients:

- $B = 0.977$: A one-kilogram increase in reported weight is associated on average with just under a one-kilogram increase in measured weight.
  - Since the data are not longitudinal, the phrase “a unit increase” here implies not a literal change over time, but rather a static comparison between two individuals who differ by one kilogram in their reported weights.

- Ordinarily, we may interpret the intercept A as the fitted value associated with $X = 0$, but it is impossible for an individual to have a reported weight equal to zero.
  - The intercept A is usually of little direct interest, since the fitted value above $X = 0$ is rarely important.
  - Here, however, if individuals’ reports are unbiased predictions of their actual weights, then we should have $\widehat{Y} = X$ — i.e., $A = 0$. The intercept $A = 1.79$ is close to zero, and the slope $B = 0.977$ is close to one.

3.2 Simple Correlation

- It is of interest to determine how closely the line fits the scatter of points.
- The standard deviation of the residuals, S_E, called the *standard error of the regression*, provides one index of fit.
  - Because of estimation considerations, the variance of the residuals is defined using $n-2$ *degrees of freedom:*
  $$S_E^2 = \frac{\sum E_i^2}{n-2}$$
  - The standard error is therefore
  $$S_E = \sqrt{\frac{\sum E_i^2}{n-2}}$$
  - Since it is measured in the units of the response variable, the standard error represents a type of 'average' residual.

- For Davis's regression of measured on reported weight, the sum of squared residuals is $\sum E_i^2 = 418.9$, and the standard error
$$S_E = \sqrt{\frac{418.9}{101-2}} = 2.05 \text{ kg.}$$
- I believe that social scientists overemphasize correlation and pay insufficient attention to the standard error of the regression.

- The *correlation coefficient* provides a *relative* measure of fit: To what degree do our predictions of Y improve when we base that prediction 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?
    - To disregard the explanatory variable is implicitly to fit the equation
    $$Y_i = A' + E_i'$$
    - We can find the best-fitting constant A' by least-squares, minimizing
    $$S(A') = \sum E_i'^2 = \sum (Y_i - A')^2$$

- The value of A' that minimizes this sum of squares is the response-variable mean, $\overline{Y}$.

- The residuals $E_i = Y_i - \widehat{Y}_i$ from the linear regression of Y on X will generally be smaller than the residuals $E'_i = Y_i - \overline{Y}$, and it is necessarily the case that
$$\sum (Y_i - \widehat{Y}_i)^2 \leq \sum (Y_i - \overline{Y})^2$$
  - This inequality holds because the 'null model,' $Y_i = A' + E'_i$ is a special case of the more general linear-regression 'model,' $Y_i = A + BX_i + E_i$, setting $B = 0$.

- We call
$$\sum E_i'^2 = \sum (Y_i - \overline{Y})^2$$
the *total sum of squares* for Y, abbreviated *TSS*, while
$$\sum E_i^2 = \sum (Y_i - \widehat{Y}_i)^2$$
is called the *residual sum of squares*, and is abbreviated *RSS*.

- The difference between the two, termed the *regression sum of squares,*
$$\textit{RegSS} \equiv \text{TSS} - \text{RSS}$$
gives the reduction in squared error due to the linear regression.

- The ratio of RegSS to TSS, the proportional reduction in squared error, defines the square of the correlation coefficient:
$$r^2 \equiv \frac{\text{RegSS}}{\text{TSS}}$$

- To find the correlation coefficient r we take the positive square root of r^2 when the simple-regression slope B is positive, and the negative square root when B is negative.

- If there is a perfect positive linear relationship between Y and X, then $r = 1$.

- A perfect negative linear relationship corresponds to $r = -1$.

- If there is no linear relationship between Y and X, then RSS $=$ TSS, RegSS $= 0$, and $r = 0$.

- Between these extremes, r gives the direction of the linear relationship between the two variables, and r^2 may be interpreted as the proportion of the total variation of Y that is 'captured' by its linear regression on X.
- Figure 3 depicts several different levels of correlation.

▶ The decomposition of total variation into 'explained' and 'unexplained' components, paralleling the decomposition of each observation into a fitted value and a residual, is typical of linear models.

- The decomposition is called the *analysis of variance* for the regression:
$$\text{TSS} = \text{RegSS} + \text{RSS}$$
- The regression sum of squares can also be directly calculated as
$$\text{RegSS} = \sum(\widehat{Y}_i - \overline{Y})^2$$

▶ It is also possible to define r by analogy with the correlation ρ between two random variables.

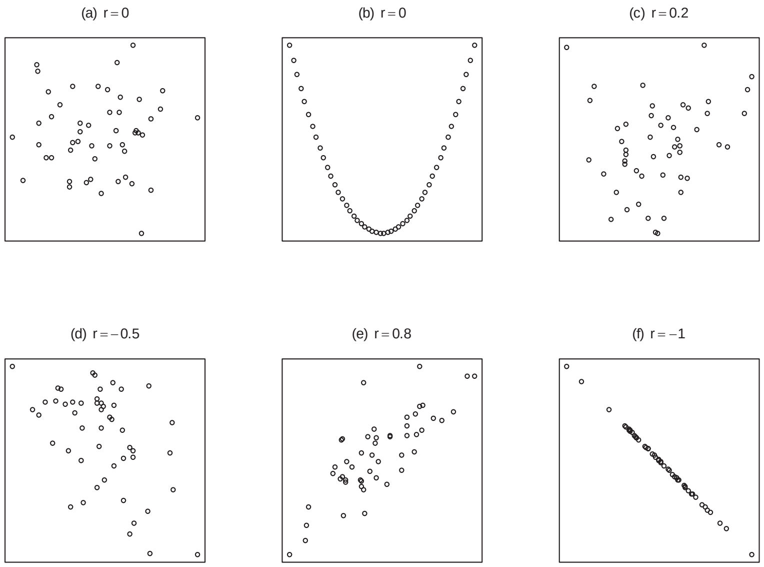

Figure 3. Scatterplots showing different correlation coefficients r. Panel (b) reminds us that r measures *linear* relationship.

- First defining the *sample covariance* between X and Y,
$$S_{XY} \equiv \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$
- we may then write
$$r = \frac{S_{XY}}{S_X S_Y} = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum(X_i - \overline{X})^2 \sum(Y_i - \overline{Y})^2}}$$
where S_X and S_Y are, respectively, the sample standard deviations of X and Y.

▶ Some comparisons between r and B:

- The correlation coefficient r is symmetric in X and Y, while the least-squares slope B is not.
- The slope coefficient B is measured in the units of the response variable per unit of the explanatory variable. For example, if dollars of income are regressed on years of education, then the units of B are dollars/year. The correlation coefficient r, however, is unitless.

- A change in scale of Y or X produces a compensating change in B, but does not affect r. If, for example, income is measured in thousands of dollars rather than in dollars, the units of the slope become $1000s/year, and the value of the slope decreases by a factor of 1000, but r remains the same.

▶ For Davis's regression of measured on reported weight,

$$\begin{aligned} \text{TSS} &= 4753.8 \\ \text{RSS} &= 418.87 \\ \text{RegSS} &= 4334.9 \\ r^2 &= \frac{4334.9}{4753.8} = .91188 \end{aligned}$$

- Since B is positive, $r = +\sqrt{.91188} = .9549$.
- The linear regression of measured on reported weight captures 91 percent of the variation in measured weight.

- Equivalently,

$$\begin{aligned} S_{XY} &= \frac{4435.9}{101-1} = 44.359 \\ S_X^2 &= \frac{4539.3}{101-1} = 45.393 \\ S_Y^2 &= \frac{4753.8}{101-1} = 47.538 \\ r &= \frac{44.359}{\sqrt{45.393 \times 47.538}} = .9549 \end{aligned}$$

4. Multiple Regression

4.1 Two Explanatory Variables

▶ The linear multiple-regression equation

$$\widehat{Y} = A + B_1X_1 + B_2X_2$$

for two explanatory variables, X_1 and X_2, describes a plane in the three-dimensional $\{X_1, X_2, Y\}$ space, as shown in Figure 4.

- The residual is the signed vertical distance from the point to the plane:

$$E_i = Y_i - \widehat{Y}_i = Y_i - (A + B_1X_{i1} + B_2X_{i2})$$

- To make the plane come as close as possible to the points in the aggregate, we want the values of A, B_1, and B_2 that minimize the sum of squared residuals:

$$S(A, B_1, B_2) = \sum E_i^2 = \sum (Y_i - A - B_1X_{i1} - B_2X_{i2})^2$$

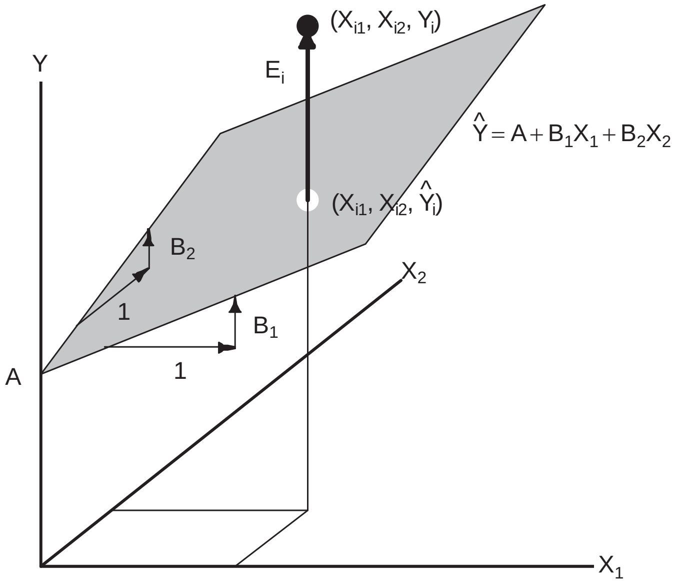

Figure 4. The multiple regression plane.

- Differentiating the sum-of-squares function with respect to the regression coefficients, setting the partial derivatives to zero, and rearranging terms produces the normal equations,

$$\begin{array}{lllll} An & + B_1\sum X_{i1} & + B_2\sum X_{i2} & = & \sum Y_i \\ A\sum X_{i1} & + B_1\sum X_{i1}^2 & + B_2\sum X_{i1}X_{i2} & = & \sum X_{i1}Y_i \\ A\sum X_{i2} & + B_1\sum X_{i2}X_{i1} & + B_2\sum X_{i2}^2 & = & \sum X_{i2}Y_i \end{array}$$

- This is a system of three linear equations in three unknowns, so it usually provides a unique solution for the least-squares regression coefficients $A, B_1,$ and B_2.

- Dropping the subscript i for observations, and using asterisks to denote variables in mean-deviation form (e.g., $Y^* \equiv Y_i - \overline{Y}$),

$$
\begin{aligned}
A &= \overline{Y} - B_1\overline{X}_1 - B_2\overline{X}_2 \\[1em]
B_1 &= \frac{\sum X_1^*Y^* \sum X_2^{*2} - \sum X_2^*Y^* \sum X_1^*X_2^*}{\sum X_1^{*2} \sum X_2^{*2} - (\sum X_1^*X_2^*)^2} \\
B_2 &= \frac{\sum X_2^*Y^* \sum X_1^{*2} - \sum X_1^*Y^* \sum X_1^*X_2^*}{\sum X_1^{*2} \sum X_2^{*2} - (\sum X_1^*X_2^*)^2}
\end{aligned}
$$

- The least-squares coefficients are uniquely defined as long as

$$\sum X_1^{*2} \sum X_2^{*2} \neq (\sum X_1^*X_2^*)^2$$

that is, unless X_1 and X_2 are perfectly correlated or unless one of the explanatory variables in invariant.

- If X_1 and X_2 are perfectly correlated, then they are said to be *collinear*.

- An illustration, using Duncan's occupational prestige data and regressing the prestige of occupations (Y) on their educational and income levels (X_1 and X_2, respectively):

$$
\begin{array}{ll}
n = 45 & \sum X_1^{*2} = 38,971. \\
\overline{Y} = \frac{2146}{45} = 47.69 & \sum X_2^{*2} = 26,271. \\
\overline{X}_1 = \frac{2365}{45} = 52.56 & \sum X_1^*X_2^* = 23,182. \\
\overline{X}_2 = \frac{1884}{45} = 41.87 & \sum X_1^*Y^* = 35,152. \\
 & \sum X_2^*Y^* = 28,383.
\end{array}
$$

  - Substituting these results into the equations for the least-squares coefficients produces $A = -6.070, B_1 = 0.5459,$ and $B_2 = 0.5987$.
  - The fitted least-squares regression equation is

$$\widehat{\text{Prestige}} = -6.07 + 0.546 \times \text{Education} + 0.599 \times \text{Income}$$

▶ A central difference in interpretation between simple and multiple regression: The slope coefficients for the explanatory variables in the multiple regression are *partial* coefficients, while the slope coefficient in simple regression gives the *marginal* relationship between the response variable and a single explanatory variable.

- That is, each slope in multiple regression represents the 'effect' on the response variable of a one-unit increment in the corresponding explanatory variable *holding constant* the value of the other explanatory variable.
- The simple-regression slope effectively ignores the other explanatory variable.
- This interpretation of the multiple-regression slope is apparent in the figure showing the multiple-regression plane. Because the regression plane is flat, its slope in the direction of X_1, holding X_2 constant, does not depend upon the specific value at which X_2 is fixed.

- Algebraically, fix X_2 to the specific value x_2 and see how $\widehat{Y}$ changes as X_1 is increased by 1, from some specific value x_1 to $x_1 + 1$:
$$[A + B_1(x_1 + 1) + B_2x_2] - (A + B_1x_1 + B_2x_2) = B_1$$
- A similar result holds for X_2.

▶ For Duncan's regression:

- A unit increase in education, holding income constant, is associated on average with an increase of 0.55 units in prestige.
- A unit increase in income, holding education constant, is associated on average with an increase of 0.60 units in prestige.
- The regression intercept, $A = -6.1$, has the following literal interpretation: The fitted value of prestige is -6.1 for a hypothetical occupation with education and income levels both equal to zero. No occupations have levels of zero for both income and education, however, and the response variable cannot take on negative values.

4.2 Several Explanatory Variables

▶ For the general case of k explanatory variables, the multiple-regression equation is

$$\begin{aligned} Y_i &= A + B_1 X_{i1} + B_2 X_{i2} + \cdots + B_k X_{ik} + E_i \\ &= \widehat{Y}_i + E_i \end{aligned}$$

- It is not possible to visualize the point cloud of the data directly when $k > 2$, but it is simple to find the values of A and the B's that minimize

$$\begin{aligned} & S(A, B_1, B_2, ..., B_k) \\ & \quad = \sum_{i=1}^{n} [Y_i - (A + B_1 X_{i1} + B_2 X_{i2} + \cdots + B_k X_{ik})]^2 \end{aligned}$$

- Minimization of the sum-of-squares function produces the normal equations for general multiple regression:

$$\begin{array}{llllll} An & +B_1 \sum X_{i1} & +B_2 \sum X_{i2} & + \cdots +B_k \sum X_{ik} & = \sum Y_i \\ A \sum X_{i1} & +B_1 \sum X_{i1}^2 & +B_2 \sum X_{i1} X_{i2} & + \cdots +B_k \sum X_{i1} X_{ik} & = \sum X_{i1} Y_i \\ A \sum X_{i2} & +B_1 \sum X_{i2} X_{i1} & +B_2 \sum X_{i2}^2 & + \cdots +B_k \sum X_{i2} X_{ik} & = \sum X_{i2} Y_i \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ A \sum X_{ik} & +B_1 \sum X_{ik} X_{i1} & +B_2 \sum X_{ik} X_{i2} & + \cdots +B_k \sum X_{ik}^2 & = \sum X_{ik} Y_i \end{array}$$

- Because the normal equations are linear, and because there are as many equations as unknown regression coefficients ($k+1$), there is usually a unique solution for the coefficients $A, B_1, B_2, ..., B_k$.
- Only when one explanatory variable is a perfect linear function of others, or when one or more explanatory variables are invariant, will the normal equations not have a unique solution.
- Dividing the first normal equation through by n reveals that the least-squares surface passes through the point of means $(\overline{X}_1, \overline{X}_2, ..., \overline{X}_k, \overline{Y})$.

▶ To illustrate the solution of the normal equations, let us return to the Canadian occupational prestige data, regressing the prestige of the occupations on average education, average income, and the percent of women in each occupation.

- The various sums, *sums of squares*, and sums of products that are required are given in the following table:

Variable	Prestige	Education	Income	Percent Women
Prestige	*253,618.*			
Education	55,326.	*12,513.*		
Income	37,748,108.	8,121,410.	*6,534,383,460.*	
Percent Women	131,909.	32,281.	14,093,097.	*187,312.*
Sum	4777.	1095.	693,386.	2956.

- Substituting these values into the normal equations and solving for the regression coefficients produces

$$\begin{aligned} A &= -6.7943 \\ B_1 &= 4.1866 \\ B_2 &= 0.0013136 \\ B_3 &= -0.0089052 \end{aligned}$$

- The fitted regression equation is

$$\begin{aligned}\widehat{\text{Prestige}} &= -6.794 + 4.187 \times \text{ Education} \\ &\quad + 0.001314 \times \text{ Income} \\ &\quad - 0.008905 \times \text{ Percent Women}\end{aligned}$$

- In interpreting the regression coefficients, we need to keep in mind the units of each variable:
- Prestige scores are arbitrarily scaled, and range from a minimum of 14.8 to a maximum of 87.2 for these 102 occupations; the hinge-spread of prestige is 24.4 points.
- Education is measured in years, and hence the impact of education on prestige is considerable — a little more than four points, on average, for each year of education, holding income and gender composition constant.

- Despite the small absolute size of its coefficient, the partial effect of income is also substantial — about 0.001 points on average for an additional dollar of income, or one point for each $1,000.
- The impact of gender composition, holding education and income constant, is very small — an average decline of about 0.01 points for each one-percent increase in the percentage of women in an occupation.

4.3 Multiple Correlation

▶ As in simple regression, the standard error in multiple regression measures the ‘average’ size of the residuals.

- As before, we divide by degrees of freedom, here $n-(k+1)=n-k-1$ to calculate the variance of the residuals; thus, the standard error is

$$S_E = \sqrt{\frac{\sum E_i^2}{n-k-1}}$$

- Heuristically, we ‘lose’ $k+1$ degrees of freedom by calculating the $k+1$ regression coefficients, $A, B_1, ..., B_k$.

- For Duncan’s regression of occupational prestige on the income and educational levels of occupations, the standard error is

$$S_E = \sqrt{\frac{7506.7}{45-2-1}} = 13.37$$

– The response variable here is the percentage of raters classifying the occupation as good or excellent in prestige; an average error of 13 is substantial.

▶ The sums of squares in multiple regression are defined as in simple regression:

$$\begin{aligned}\text{TSS} &= \sum (Y_i - \overline{Y})^2 \\ \text{RegSS} &= \sum (\widehat{Y}_i - \overline{Y})^2 \\ \text{RSS} &= \sum (Y_i - \widehat{Y}_i)^2 = \sum E_i^2\end{aligned}$$

- The fitted values $\widehat{Y}_i$ and residuals E_i now come from the multiple-regression equation.

- We also have a similar decomposition of variation:

$$\text{TSS} = \text{RegSS} + \text{RSS}$$

- The least-squares residuals are uncorrelated with the fitted values and with each of the X’s.

- The squared multiple correlation R^2 represents the proportion of variation in the response variable captured by the regression:

$$R^2 \equiv \frac{\text{RegSS}}{\text{TSS}}$$

  - The multiple correlation coefficient is the positive square root of R^2, and is interpretable as the simple correlation between the fitted and observed Y-values.

- For Duncan's regression,

$$\begin{aligned} \text{TSS} &= 43,687. \\ \text{RegSS} &= 36,181. \\ \text{RSS} &= 7506.7 \end{aligned}$$

  - The squared multiple correlation is

$$R^2 = \frac{36,181}{43,688} = .8282$$

    indicating that more than 80 percent of the variation in prestige among the 45 occupations is accounted for by its linear regression on the income and educational levels of the occupations.

4.4 Standardized Regression Coefficients

- Social researchers often wish to compare the coefficients of different explanatory variables in a regression analysis.
  - When the explanatory variables are commensurable, comparison is straightforward.
  - Standardized regression coefficients permit a limited assessment of the relative effects of *incommensurable* explanatory variables.
- Imagine that the annual dollar income of wage workers is regressed on their years on education, years of labor-force experience, and some other explanatory variables, producing the fitted regression equation

$$\widehat{\text{Income}} = A + B_1 \times \text{ Education } + B_2 \times \text{ Experience } + \cdots$$

  - Since education and experience are measured in years, the coefficients B_1 and B_2 are both expressed in dollars/year, and can be directly compared.

- More commonly, explanatory variables are measured in different units.
  - In the Canadian occupational prestige regression, for example, the coefficient for education is expressed in points (of prestige) per year; the coefficient for income is expressed in points per dollar; and the coefficient of gender composition in points per percent-women.
    - The income coefficient (0.001314) is much smaller than the education coefficient (4.187) not because income is a much less important determinant of prestige, but because the unit of income (the dollar) is small, while the unit of education (the year) is relatively large.
    - If we were to re-express income in \$1000s, then we would multiply the income coefficient by 1000.

- Standardized regression coefficients rescale the B's according to a measure of explanatory-variable spread.
  - We may, for example, multiply each regression coefficient by the hinge-spread of the corresponding explanatory variable. For the Canadian prestige data:

	H-spread $\times B_j$		
Education:	4.28×4.187	$=$	17.92
Income:	4131×0.001314	$=$	5.4281
Gender:	48.68×-0.008905	$=$	-0.4335

  - For other data, where the variation in education and income may be different, the relative impact of the two variables may also differ, even if the regression coefficients are unchanged.
  - The following observation should give you pause: If two explanatory variables are commensurable, and if their hinge-spreads differ, then performing this calculation is, in effect, to adopt a rubber ruler.

- It is much more common to standardize regression coefficients using the standard deviations of the explanatory variables rather than their hinge-spreads.
  - The usual practice standardizes the response variable as well, but this is inessential:

$$\begin{aligned}
Y_i &= A + B_1 X_{i1} + \cdots + B_k X_{ik} + E_i \\
\overline{Y} &= A + B_1 \overline{X}_1 + \cdots + B_k \overline{X}_k \\
Y_i - \overline{Y} &= B_1(X_{i1} - \overline{X}_1) + \cdots + B_k(X_{ik} - \overline{X}_k) + E_i \\[1em]
\frac{Y_i - \overline{Y}}{S_Y} &= \left(B_1 \frac{S_1}{S_Y}\right) \frac{X_{i1} - \overline{X}_1}{S_1} \\
&\quad + \cdots + \left(B_k \frac{S_k}{S_Y}\right) \frac{X_{ik} - \overline{X}_k}{S_k} + \frac{E_i}{S_Y} \\[1em]
Z_{iY} &= B_1^* Z_{i1} + \cdots + B_k^* Z_{ik} + E_i^*
\end{aligned}$$

- $Z_Y \equiv (Y - \overline{Y})/S_Y$ is the standardized response variable, linearly transformed to a mean of zero and a standard deviation of one.
- $Z_1, ..., Z_k$ are the explanatory variables, similarly standardized.
- $E^* \equiv E/S_Y$ is the transformed residual which, note, *does not* have a standard deviation of one.
- $B_j^* \equiv B_j(S_j/S_Y)$ is the *standardized partial regression coefficient* for the jth explanatory variable.
- The standardized coefficient is interpretable as the average change in Y, in standard-deviation units, for a one standard-deviation increase in X_j, holding constant the other explanatory variables.

▶ For the Canadian prestige regression,

$$\begin{array}{llcr} \text{Education:} & 4.187 \times 2.728/17.20 & = & 0.6639 \\ \text{Income:} & 0.001314 \times 4246/17.20 & = & 0.3242 \\ \text{Gender:} & -0.008905 \times 31.72/17.20 & = & -0.01642 \end{array}$$

- Because both income and gender composition have substantially non-normal distributions, however, the use of standard deviations here is difficult to justify.

▶ A common misuse of standardized coefficients is to employ them to make comparisons of the effects of the *same* explanatory variable in two or more samples drawn from populations with different spreads.

5. Least-Squares Regression Using Matrices (time permitting)

5.1 Basic Definitions

▶ A *matrix* is a rectangular table of numbers or of numerical variables; for example

$$\underset{(4\times 3)}{\mathbf{X}} = \begin{bmatrix} 1 & -2 & 3 \\ 4 & -5 & -6 \\ 7 & 8 & 9 \\ 0 & 0 & 10 \end{bmatrix}$$

or, more generally,

$$\underset{(m\times n)}{\mathbf{A}} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

- The first matrix is of *order* 4 by 3; the second of order m by n.

▶ a one-column matrix is called a *column vector*,

$$\underset{(m\times 1)}{\mathbf{a}} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix}$$

▶ Likewise, a matrix with just one row is called a *row vector*,

$$\mathbf{b}' = [b_1 \; b_2 \; \cdots \; b_n]$$

▶ The transpose of a matrix reverses rows and columns:

$$\underset{(4\times 3)}{\mathbf{X}} = \begin{bmatrix} 1 & -2 & 3 \\ 4 & -5 & -6 \\ 7 & 8 & 9 \\ 0 & 0 & 10 \end{bmatrix} \quad \underset{(3\times 4)}{\mathbf{X}'} = \begin{bmatrix} 1 & 4 & 7 & 0 \\ -2 & -5 & 8 & 0 \\ 3 & -6 & 9 & 10 \end{bmatrix}$$

- A square matrix has equal numbers of rows and columns:

$$\underset{(3\times 3)}{\mathbf{B}} = \begin{bmatrix} -5 & 1 & 3 \\ 2 & 2 & 6 \\ 7 & 3 & -4 \end{bmatrix}$$

  - The diagonal of a (generic) square matrix $\mathbf{A}$ of order n consists of the entries $a_{11}, a_{22}, \ldots, a_{nn}$.
  - For example, the diagonal of $\mathbf{B}$ consists of the entries -5, 2 and -4.

- A symmetric square matrix is equal to its transpose; thus $\mathbf{B}$ is not symmetric, but the following matrix is:

$$\mathbf{C} = \begin{bmatrix} -5 & 1 & 3 \\ 1 & 2 & 6 \\ 3 & 6 & -4 \end{bmatrix}$$

5.2 Matrix Arithmetic

- To be added or subtracted, matrices must be of the same order; matrices are added, subtracted, and negated element-wise; for

$$\underset{(2\times 3)}{\mathbf{A}} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$$

  and

$$\underset{(2\times 3)}{\mathbf{B}} = \begin{bmatrix} -5 & 1 & 2 \\ 3 & 0 & -4 \end{bmatrix}$$

  we have

$$\underset{(2\times 3)}{\mathbf{C}} = \mathbf{A} + \mathbf{B} = \begin{bmatrix} -4 & 3 & 5 \\ 7 & 5 & 2 \end{bmatrix}$$

$$\underset{(2\times 3)}{\mathbf{D}} = \mathbf{A} - \mathbf{B} = \begin{bmatrix} 6 & 1 & 1 \\ 1 & 5 & 10 \end{bmatrix}$$

$$\underset{(2\times 3)}{\mathbf{E}} = -\mathbf{A} = \begin{bmatrix} -1 & -2 & -3 \\ -4 & -5 & -6 \end{bmatrix}$$

- The *inner product* of two vectors (each with n entries), say $\underset{(1\times n)}{\mathbf{a}'}$ and $\underset{(n\times 1)}{\mathbf{b}}$, denoted $\mathbf{a}'\cdot\mathbf{b}$, is a number formed by multiplying corresponding entries of the vectors and summing the resulting products:

$$\mathbf{a}'\cdot\mathbf{b}=\sum_{i=1}^{n}a_ib_i$$

  - For example,

$$[2,0,1,3]\cdot\begin{bmatrix}-1\\6\\0\\9\end{bmatrix}=2(-1)+0(6)+1(0)+3(9)=25$$

- Two matrices $\mathbf{A}$ and $\mathbf{B}$ are *conformable for multiplication* in the order given (i.e., $\mathbf{AB}$) if the number of *columns* of the left-hand factor ($\mathbf{A}$) is equal to the number of *rows* of the right-hand factor ($\mathbf{B}$).
  - Thus $\mathbf{A}$ and $\mathbf{B}$ are conformable for multiplication if $\mathbf{A}$ is of order $(m\times n)$ and $\mathbf{B}$ is of order $(n\times p)$.

  - For example,

$$\underset{(2\times 3)}{\begin{bmatrix}1&2&3\\4&5&6\end{bmatrix}}\underset{(3\times 3)}{\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}}$$

are conformable for multiplication, but

$$\underset{(3\times 3)}{\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}}\underset{(2\times 3)}{\begin{bmatrix}1&2&3\\4&5&6\end{bmatrix}}$$

are not.

- Let $\mathbf{C}=\mathbf{AB}$ be the matrix product; and let $\mathbf{a}_i'$ be the ith *row* of $\mathbf{A}$ and $\mathbf{b}_j$ be the jth *column* of $\mathbf{B}$.
  - Then $\mathbf{C}$ is a matrix of order $(m\times p)$ in which

$$c_{ij}=\mathbf{a}_i'\cdot\mathbf{b}_j=\sum_{k=1}^{n}a_{ik}b_{kj}$$

- Some examples:

$$\underset{(2\times 3)}{\begin{bmatrix} & \Longrightarrow & \\ 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}} \underset{(3\times 3)}{\begin{bmatrix} & 1 & 0 & 0 \\ \Downarrow & 0 & 1 & 0 \\ & 0 & 0 & 1 \end{bmatrix}}$$

$$= \underset{(2\times 3)}{\begin{bmatrix} 1(1)+2(0)+3(0), & 1(0)+2(1)+3(0), & 1(0)+2(0)+3(1) \\ 4(1)+5(0)+6(0), & 4(0)+5(1)+6(0), & 4(0)+5(0)+6(1) \end{bmatrix}}$$

$$= \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 0 & 3 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 4 & 5 \\ 8 & 13 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 9 & 12 \\ 5 & 8 \end{bmatrix}$$

▶ A matrix that has ones down the main diagonal and zeroes elsewhere is called an identity matrix, and is symbolized by $\mathbf{I}$; the identity matrix of order 3 is symbolized by $\mathbf{I}_3$.

- As in the previous example, multiplying a matrix by an identity matrix recovers the original matrix.

▶ In scalar algebra, division is essential to the solution of simple equations.

- For example,

$$6x = 12$$
$$x = \frac{12}{6} = 2$$

or, equivalently,

$$\frac{1}{6} \times 6x = \frac{1}{6} \times 12$$
$$x = 2$$

where $\frac{1}{6} = 6^{-1}$ is the scalar inverse of 6.

- ▶ In matrix algebra, there is no direct analog of division, but most square matrices have a *matrix inverse*.
  - The inverse of a square matrix $\mathbf{A}$ is a square matrix of the same order, written $\mathbf{A}^{-1}$, with the property that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$.
  - If a square matrix has an inverse, then the matrix is termed *nonsingular*; a square matrix without an inverse is termed *singular*.

- For example, the inverse of the nonsingular matrix

$$\begin{bmatrix} 2 & 5 \\ 1 & 3 \end{bmatrix}$$

is the matrix

$$\begin{bmatrix} 3 & -5 \\ -1 & 2 \end{bmatrix}$$

as we can verify:

$$\begin{bmatrix} 2 & 5 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 3 & -5 \\ -1 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \checkmark$$

$$\begin{bmatrix} 3 & -5 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} 2 & 5 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \checkmark$$

- In scalar algebra, only the number 0 has no inverse, but there are singular *nonzero* matrices.
  - Let us hypothesize that $\mathbf{B}$ is the inverse of the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

  - But

$$\mathbf{AB} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} \\ 0 & 0 \end{bmatrix} \neq \mathbf{I}_2$$

- Matrix inverses can be used to solve systems of linear simultaneous equations.
  - For example:

$$\begin{aligned} 2x_1 + 5x_2 &= 4 \\ x_1 + 3x_2 &= 5 \end{aligned}$$

  - Writing these equations as a matrix equation,

$$\begin{bmatrix} 2 & 5 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 4 \\ 5 \end{bmatrix}$$

$$\underset{(2\times2)}{\mathbf{A}}\,\underset{(2\times1)}{\mathbf{x}} = \underset{(2\times1)}{\mathbf{b}}$$

  - Then

$$\begin{aligned} \mathbf{A}^{-1}\mathbf{Ax} &= \mathbf{A}^{-1}\mathbf{b} \\ \mathbf{Ix} &= \mathbf{A}^{-1}\mathbf{b} \\ \mathbf{x} &= \mathbf{A}^{-1}\mathbf{b} \end{aligned}$$

  - For the example

$$\begin{aligned} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} &= \begin{bmatrix} 2 & 5 \\ 1 & 3 \end{bmatrix}^{-1} \begin{bmatrix} 4 \\ 5 \end{bmatrix} \\ &= \begin{bmatrix} 3 & -5 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} 4 \\ 5 \end{bmatrix} \\ &= \begin{bmatrix} -13 \\ 6 \end{bmatrix} \end{aligned}$$

5.3 Least-Squares Regression

- In *scalar* (i.e., single-number) form, multiple linear regression is written as

$$Y_i = A + B_1X_{i1} + B_2X_{i2} + \cdots + B_kX_{ik} + E_i$$

  - There are n such equations, one for each observation $i = 1, 2, \ldots, n$.

- We can write these n equations as a single matrix equation:

$$\underset{(n\times1)}{\mathbf{y}} = \underset{(n\times k+1)}{\mathbf{X}}\underset{(k+1\times1)}{\mathbf{b}} + \underset{(n\times1)}{\mathbf{e}}$$

  - $\mathbf{y} = (Y_1, Y_2, \ldots, Y_n)'$ is a vector of observations on the response variable
  - $$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1k} \\ 1 & X_{21} & \cdots & X_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \cdots & X_{nk} \end{bmatrix}$$
    called the *model* (or *design*) *matrix*, contains the values of the explanatory variables, with an initial column of ones for the regression constant (called the *constant regressor*).
  - $\mathbf{b} = (A, B_1, \ldots, B_k)'$ contains the regression coefficient
  - $\mathbf{e} = (E_1, E_2, \ldots, E_n)'$ is a vector of residuals.

- The residual sum of squares is $S(\mathbf{b}) = \mathbf{e}'\mathbf{e}$.
- Minimizing the residual sum of squares leads to the least-squares normal equations in matrix form:

$$\underset{(k+1\times k+1)}{\mathbf{X}'\mathbf{X}}\underset{(k+1\times 1)}{\mathbf{b}} = \underset{(k+1\times 1)}{\mathbf{X}'\mathbf{y}}$$

- This is a system of $k+1$ linear equations in the $k+1$ unknown regression coefficients $\mathbf{b}$.
- The coefficient matrix for the system of equation,

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & \sum X_{i1} & \sum X_{i2} & \cdots & \sum X_{ik} \\ \sum X_{i1} & \sum X_{ik}^2 & \sum X_{i1}X_{i2} & \cdots & \sum X_{i1}X_{ik} \\ \sum X_{i2} & \sum X_{i2}X_{i1} & \sum X_{i2}^2 & \cdots & \sum X_{i2}X_{ik} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum X_{ik} & \sum X_{ik}X_{i1} & \sum X_{ik}X_{i2} & \cdots & \sum X_{ik}^2 \end{bmatrix}$$

contains sums of squares and cross-products among the columns of the model matrix.

- The right-hand-side vector, $\mathbf{X}'\mathbf{y} = [\sum Y_i, \sum X_{i1}Y_i, \sum X_{i2}Y_i, \dots, \sum X_{ik}Y_i]'$ contains sums of cross-products between each column of the model matrix and the vector of responses.
- The sums of squares and products $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{y}$ can be calculated directly from the data.

▶ $\mathbf{X}'\mathbf{X}$ is nonsingular if no explanatory variable is a perfect linear function of the others.

- Then the solution of the normal equations is

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

6. Summary

▶ In simple linear regression, the least-squares coefficients are given by

$$\begin{aligned} B &= \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sum(X_i - \overline{X})^2} \\ A &= \overline{Y} - B\overline{X} \end{aligned}$$

▶ The least-squares coefficients in multiple linear regression are found by solving the normal equations for the intercept A and the slope coefficients $B_1, B_2, ..., B_k$.

▶ The least-squares residuals, E, are uncorrelated with the fitted values, $\widehat{Y}$, and with the explanatory variables, $X_1, ..., X_k$.

▶ The linear regression decomposes the variation in Y into ‘explained’ and ‘unexplained’ components: TSS $=$ RegSS $+$ RSS.

- The standard error of the regression,
$$S_E = \sqrt{\frac{\sum E_i^2}{n-k-1}}$$
gives the ‘average’ size of the regression residuals.
- The squared multiple correlation,
$$R^2 = \frac{\text{RegSS}}{\text{TSS}}$$
indicates the proportion of the variation in Y that is captured by its linear regression on the X’s.
- By rescaling regression coefficients in relation to a measure of variation — e.g., the hinge-spread or standard deviation — standardized regression coefficients permit a limited comparison of the relative impact of incommensurable explanatory variables.

- The least-squares regression coefficients can be calculated in matrix form as
$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{y}$$