York SPIDA John Fox

Review of Linear Models and Related Topics

1. Topics

- Multiple regression
- Elliptical geometry of linear least-squares regression
- Dummy-variable regression
- Regression diagnostics (as time permits)
- Implementation of linear models in R

- Principal sources:
  - Fox, *Applied Regression Analysis and Generalized Linear Models, Second Edition* (Sage, 2008)
  - Fox and Weisberg, *An R Companion to Applied Regression, Second Edition* (Sage, 2011)
  - Fox, *A Mathematical Primer for Social Statistics* (Sage, 2009)

2. Multiple Regression

- The linear multiple-regression model relates a quantitative response variable to one or several quantitative explanatory variables.
  - In its basic form, the multiple regression model specifies linear, additive relationships, but it is readily generarlized to certain kinds of nonlinear relationships, interactions, and categorical explanatory variables.

2.1 The Multiple-Regression Model

- The statistical model for multiple regression is

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

- In vector form:

$$Y_i = [1, x_{i1}, x_{i2}, \ldots, x_{ik}] \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \varepsilon_i$$

$$= \underset{(1\times k+1)}{\mathbf{x}_i'} \quad \underset{(k+1\times 1)}{\boldsymbol{\beta}} + \varepsilon_i$$

- Written as a matrix equation for n observations:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\underset{(n\times 1)}{\mathbf{y}} = \underset{(n\times k+1)}{\mathbf{X}} \quad \underset{(k+1\times 1)}{\boldsymbol{\beta}} + \underset{(n\times 1)}{\boldsymbol{\varepsilon}}$$

  - $\mathbf{X}$ is called the *model-matrix* for the regression.

▶ The assumptions underlying the model concern the errors, ε_i:

- **Linearity.** $E(\varepsilon_i) = 0$, which implies that

$$E(Y_i) = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

- **Constant Variance.** $V(\varepsilon_i) = \sigma^2_\varepsilon$, which implies that

$$V(Y_i|X_{i1}, ..., X_{ik}) = \sigma^2_\varepsilon$$

- **Normality.** $\varepsilon_i \sim N(0, \sigma^2_\varepsilon)$, which implies that

$$Y_i|X_{i1}, ..., X_{ik} \sim N(\alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}, \sigma^2_\varepsilon)$$

  The first three assumptions are illustrated in Figure 1 for a single X (simple linear regression).

- **Independence.** $\varepsilon_i, \varepsilon_j$ independent for $i \neq j$. These assumptions can be written compactly as $\boldsymbol{\varepsilon} \sim \mathbf{N}_n(\mathbf{0}, \sigma^2_\varepsilon \mathbf{I}_n)$.
- **Fixed X's or X's independent of ε.**

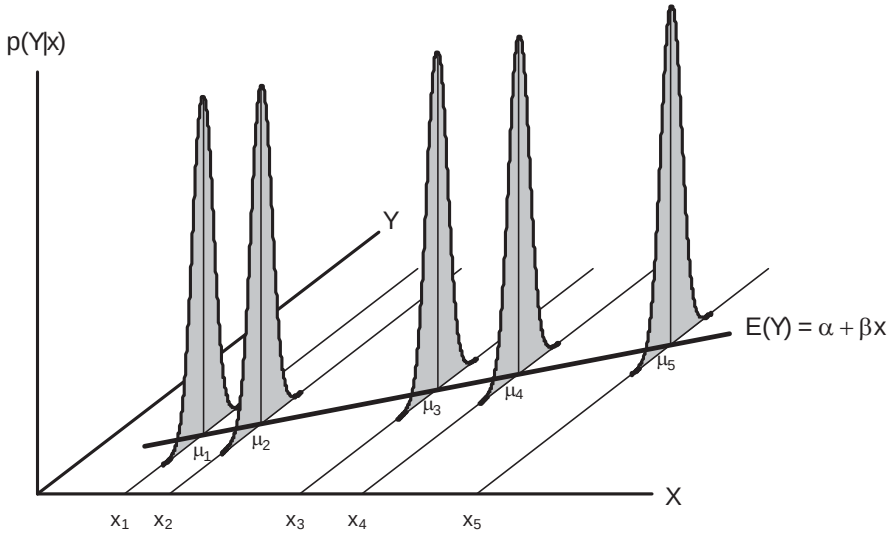

Figure 1. The assumptions of linearity, normality, and constant variance in the simple-regression model.

- Under these assumptions (or particular subsets of them), the least-squares estimators $A, B_1, ..., B_k$ of $\alpha, \beta_1, ..., \beta_k$ are
  - linear functions of the data, and hence relatively simple:

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

    with covariance matrix

$$V(\mathbf{b}) = \sigma^2_\varepsilon(\mathbf{X}'\mathbf{X})^{-1}$$

  - unbiased: $E(\mathbf{b}) = \boldsymbol{\beta}$.
  - maximally efficient among unbiased estimators;
  - maximum-likelihood estimators;
  - normally distributed.

- The slope coefficient B_j in multiple regression has sampling variance

$$V(B_j) = \frac{1}{1 - R_j^2} \times \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n}(X_{ij} - \overline{X}_j)^2}$$

  where R_j^2 is the multiple correlation from the regression of X_j on all of the other X's.
  - The second factor is essentially the sampling variance of the slope in simple regression, although the error variance σ_ε^2 is generally smaller than in simple regression.
  - The first factor — called the *variance-inflation factor* — is large when the explanatory variable X_j is strongly correlated with other explanatory variables (the problem of collinearity).
- Fitted values and residuals for the regression are given respectively by

$$\widehat{\mathbf{y}} = \{\widehat{Y}_i\} = \mathbf{Xb}$$
$$\mathbf{e} = \{E_i\} = \mathbf{y} - \widehat{\mathbf{y}}$$

2.2 Confidence Intervals and Hypothesis Tests

2.2.1 Individual Slope Coefficients

- Confidence intervals and hypothesis tests for individual coefficients closely follow the pattern of inference for means:
  - The variance of the residuals provides an unbiased estimator of σ_ε^2:

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

  - Using S_E^2, we can calculate the standard error of B_j:

$$\text{SE}(B_j) = \frac{1}{\sqrt{1 - R_j^2}} \times \frac{S_E}{\sqrt{\sum(X_{ij} - \overline{X}_j)^2}}$$

  - Confidence intervals and tests, based on the t-distribution with $n-k-1$ degrees of freedom, follow straightforwardly.

2.2.2 All Slopes

- We can also test the global or 'omnibus' null hypothesis that all of the regression slopes are zero:

$$H_0\text{: } \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

which is not quite the same as testing the separate hypotheses

$$H_0^{(1)}\text{: } \beta_1 = 0;\ H_0^{(2)}\text{: } \beta_2 = 0; ...;\ H_0^{(k)}\text{: } \beta_k = 0$$

- An F-test for the omnibus null hypothesis is given by

$$F_0 = \frac{\dfrac{\text{RegSS}}{k}}{\dfrac{\text{RSS}}{n-k-1}} = \frac{n-k-1}{k} \times \frac{R^2}{1-R^2}$$

where RegSS $= \sum(\widehat{Y}_i - \overline{Y})^2$ and RSS $= \sum E_i^2$ are, respectively, the *regression* and *residual sums of squares*, which add to TSS, the *total sum of squares*.

· Then $R^2 =$ RegSS/TSS is the squared multiple correlation.

- Under the null hypothesis, this test statistic follows an F-distribution with k and $n-k-1$ degrees of freedom.
- The calculation of the omnibus F-statistic can be organized in an *analysis-of-variance table*:

Source	*Sum of Squares*	*df*	*Mean Square*	F
Regression	RegSS	k	$\dfrac{\text{RegSS}}{k}$	$\dfrac{\text{RegMS}}{\text{RMS}}$
Residuals	RSS	$n-k-1$	$\dfrac{\text{RSS}}{n-k-1}$	
Total	TSS	$n-1$		

- When the null hypothesis is true, RegMS and RMS provide independent estimates of the error variance, so the ratio of the two mean squares should be close to one.
- When the null hypothesis is false, RegMS estimates the error variance plus a positive quantity that depends upon the β's:

$$E(F_0) \approx \frac{E(\text{RegMS})}{E(\text{RMS})} = \frac{\sigma_\varepsilon^2 + \boldsymbol{\beta}_1' \mathbf{X}^{*\prime} \mathbf{X}^* \boldsymbol{\beta}_1 / k}{\sigma_\varepsilon^2}$$

$$\frac{\sigma_\varepsilon^2 + \text{ positive quantity}}{\sigma_\varepsilon^2}$$

  where $\boldsymbol{\beta}_1 = [\beta_1, \ldots, \beta_k]'$ and $\mathbf{X}^* = \{x_{ij} - \overline{x}\}$.
- We consequently reject the omnibus null hypothesis for values of F_0 that are sufficiently larger than 1.

2.2.3 A Subset of Slopes

▶ Consider the hypothesis

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_q = 0$$

where $1 \leq q \leq k$.

- The 'full' regression model, including all of the explanatory variables, may be written:

$$Y_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_q X_{iq} + \beta_{q+1} X_{i,q+1} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

- If the null hypothesis is correct, then the first q of the β's are zero, yielding the 'null' model

$$Y_i = \alpha + \beta_{q+1} X_{i,q+1} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

- The null model omits the first q explanatory variables, regressing Y on the remaining $k - q$ explanatory variables.

- An F-test of the null hypothesis is based upon a comparison of these two models:
  - RSS_1 and RegSS_1 are the residual and regression sums of squares for the full model.
  - RSS_0 and RegSS_0 are the residual and regression sums of squares for the null model.
  - Because the null model is a special case of the full model, $\text{RSS}_0 \geq \text{RSS}_1$. Equivalently, $\text{RegSS}_0 \leq \text{RegSS}_1$.
  - If the null hypothesis is wrong and (some of) $\beta_1, ..., \beta_q$ are nonzero, then the *incremental* (or *'extra'*) *sum of squares* due to fitting the additional explanatory variables
$$\text{RSS}_0 - \text{RSS}_1 = \text{RegSS}_1 - \text{RegSS}_0$$
should be large.

- The F-statistic for testing the null hypothesis is
$$F_0 = \frac{\dfrac{\text{RegSS}_1 - \text{RegSS}_0}{q}}{\dfrac{\text{RSS}_1}{n-k-1}} = \frac{n-k-1}{q} \times \frac{R_1^2 - R_0^2}{1 - R_1^2}$$
- Under the null hypothesis, this test statistic has an F-distribution with q and $n-k-1$ degrees of freedom.

2.2.4 General Linear Hypotheses

- More generally, we can test the linear hypothesis

$$H_0: \underset{(q\times k+1)}{\mathbf{L}} \underset{(k+1\times 1)}{\boldsymbol{\beta}} = \underset{(q\times 1)}{\mathbf{c}}$$

  - $\mathbf{L}$ and $\mathbf{c}$ contain pre-specified constants.
  - The *hypothesis matrix* $\mathbf{L}$ is of full row rank $q \leq k+1$.
- The test statistic

$$F_0 = \frac{(\mathbf{Lb}-\mathbf{c})'\left[\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}'\right]^{-1}(\mathbf{Lb}-\mathbf{c})}{qS_E^2}$$

  follows an F-distribution with q and $n-k-1$ degrees of freedom if H_0 is true.
- Tests of individual coefficients, of all slope coefficients, and of subsets of coefficients can all be expressed in this form.

3. Elliptical Geometry of Regression

- The material in this section was strongly influenced by Monette (1990), "Geometry of Multiple Regression and Interactive 3-D Graphics"; and Friendly, Ornstein, and Fox, (2012) "Elliptical Insights: Understanding Statistical Methods through Elliptical Geometry."

3.1 The Standard Data Ellipse

- Consider the quadratic form $(\mathbf{x}-\overline{\mathbf{x}})'\mathbf{S}_{XX}^{-1}(\mathbf{x}-\overline{\mathbf{x}})$, where $\mathbf{x}$ is a $k\times 1$ vector of explanatory-variable values, $\overline{\mathbf{x}}$ is the vector of means of the X's, and $\mathbf{S}_{XX}$ is the sample covariance matrix of the X's.
- Setting the quadratic form to 1 produces the equation of an ellipsoid—called the *standard data ellipsoid*—centred at the means of the explanatory variables.

▶ For two variables, X_1 and X_2, the standard data *ellipse* has the equation

$$\frac{n-1}{\sum x_{i1}^{*2} \sum x_{i2}^{*2} - \left(\sum x_{i1}^* x_{i2}^*\right)^2} [x_1 - \overline{x}_1, x_2 - \overline{x}_2] \times \begin{bmatrix} \sum x_{i2}^{*2} & -\sum x_{i1}^* x_{i2}^* \\ -\sum x_{i1}^* x_{i2}^* & \sum x_{i1}^{*2} \end{bmatrix} \begin{bmatrix} x_1 - \overline{x}_1 \\ x_2 - \overline{x}_2 \end{bmatrix} = 1$$

where the x_{ij}^*s are deviations from the means $\overline{x}_j$.

- The horizontal shadow of the ellipse is twice the standard deviation of X_1, and the vertical shadow is twice the standard deviation of X_2 (see Figure 2).
- Figure 3 shows data ellipses corresponding to different correlations.

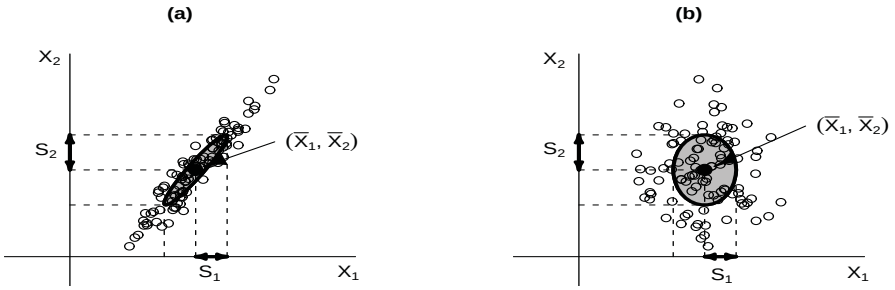

Figure 2. Scatterplot and standard data ellipse for *(a)* two highly correlated variables and *(b)* two uncorrelated variables, X_1 and X_2. In each panel, the standard ellipse is centred at the point of means $(\overline{X}_1, \overline{X}_2)$; its shadows on the axes give the standard deviations of the two variables. (The standard deviations are the half-widths of the shadows.)

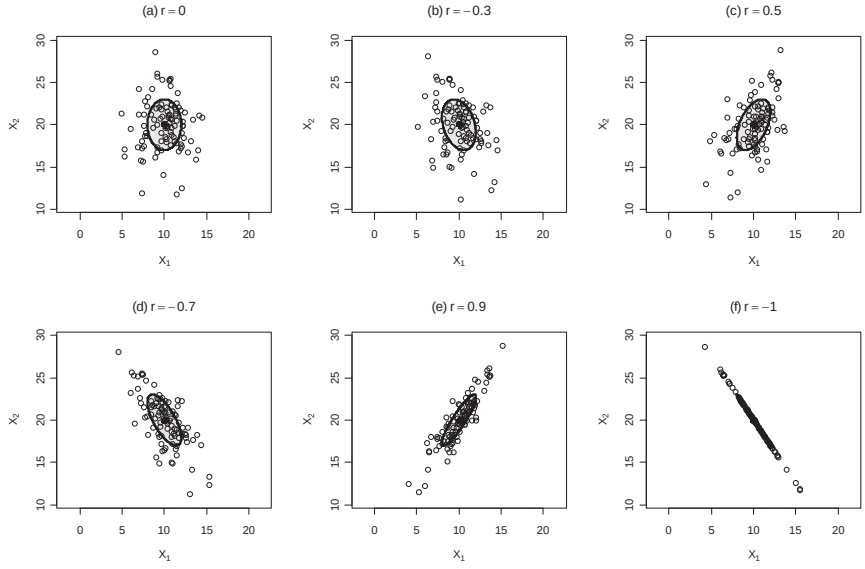

Figure 3. Scatterplots and standard data ellipses corresponding to different correlations. In each case, $\overline{X}_1 = 10$, $\overline{X}_2 = 20$, $SD(X_1) = 2$, and $SD(X_2) = 3$.

- This representation of the data is most compelling when the variables are multivariately normally distributed.
  - In this case, the means and covariance matrix of the X's are sufficient statistics for their joint distribution (that is, exhaust all the information in the data about the parameters of the distribution), and the standard data ellipsoid estimates a constant-density contour of the joint distribution.
  - Even when variables are *not* multivariate normal, the standard ellipsoid is informative because of the role of the means, variances, and covariances in least-squares regression.
- Figure 4 shows the standard data ellipse and the least-squares line for the regression of Y on X.
  - For bivariate-normal data, vertical slices of the data ellipse represent the conditional distibutions of Y fixing the value of X, and the bisectors of these slices given the conditional means, $\overline{Y}|x$.
  - As a consequence, the least-squares line goes through the points of

vertical tangency of the ellipse.

- As illustrated in Figure 5, many properties of least-squares regression are illuminated by the standard data ellipse:
  - The vertical slice in the centre of the ellipse shows the conditional variation of Y given X, that is (diregarding degrees of freedom) twice the standard deviation of the residuals, S_E.
  - Where the least-squares line intersects the ellipse gives the correlation between X and Y — actually, the correlation times the standard deviation of Y.
  - The diagram also shows the relationship between the correlation and the slope of the regression of Y on X.

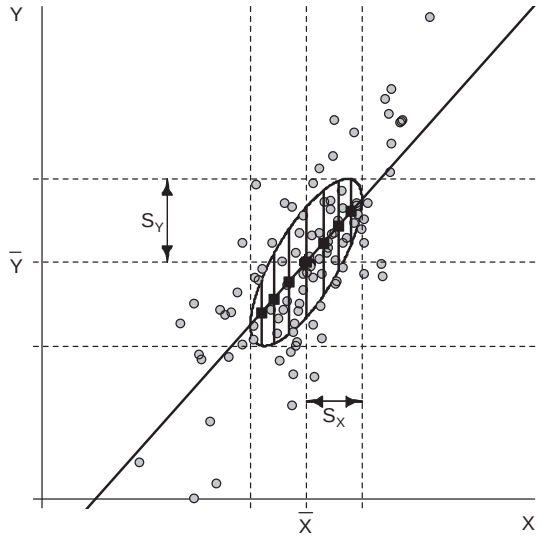

Figure 4. The least-squares line goes through the vertical bisectors and the points of vertical tangency of the standard data ellipse.

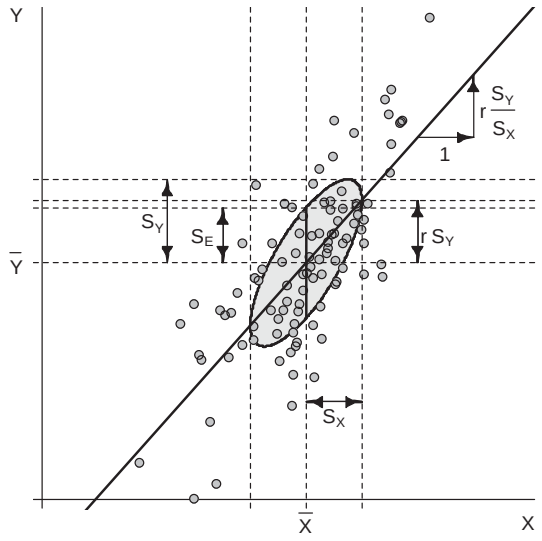

Figure 5. The standard data ellipse illuminates many characteristics of linear least-squares regression and correlation.

3.2 Joint Confidence Regions

- Consider the F-test statistic for the linear hypothesis that the slope coefficients $\boldsymbol{\beta}_1 = (\beta_1, \ldots, \beta_k)'$ in a multiple regression are all equal to particular values, $\boldsymbol{\beta}_1^{(0)}$

$$F_0 = \frac{(\mathbf{b}_1 - \boldsymbol{\beta}_1^{(0)})' \mathbf{V}_{11}^{-1} (\mathbf{b}_1 - \boldsymbol{\beta}_1^{(0)})}{k S_E^2}$$

  where $\mathbf{V}_{11}$ represents the square submatrix consisting of the entries in the k rows and k columns of $(\mathbf{X}'\mathbf{X})^{-1}$ for the slope coefficients in $\mathbf{b}_1$.

- This test can be turned around to produce a $100(1-\alpha)\%$ joint confidence region for the regression parameters $\boldsymbol{\beta}_1$:

$$\Pr\left[\frac{\left(\mathbf{b}_1 - \boldsymbol{\beta}_1^{(0)}\right)' \mathbf{V}_{11}^{-1} \left(\mathbf{b}_1 - \boldsymbol{\beta}_1^{(0)}\right)}{k S_E^2} \leq F_{\alpha,\, k,\, n-k-1}\right] = 1 - \alpha$$

  where $F_{\alpha,\, k,\, n-k-1}$ is the critical value of F with k and $n-k-1$ degrees of freedom, corresponding to a right-tail probability of α.

- The joint confidence region for $\boldsymbol{\beta}_1$ is thus

$$\text{all } \boldsymbol{\beta}_1 \text{for which } (\mathbf{b}_1 - \boldsymbol{\beta}_1)' \mathbf{V}_{11}^{-1} (\mathbf{b}_1 - \boldsymbol{\beta}_1) \leq k S_E^2 F_{\alpha,\, k,\, n-k-1}$$

  - This region represents an ellipsoid in the k dimensional parameter space ("β-space") of the slope coefficients.
  - Like a confidence interval, a joint confidence region is a portion of the parameter space constructed so that, with repeated sampling, a preselected percentage of regions will contain the true parameter values.
  - Unlike a confidence interval, however, which pertains to a *single* coefficient β_j, a joint confidence region encompasses all *combinations* of values for the parameters $\beta_1, \ldots, \beta_k$ that are *simultaneously* acceptable at the specified level of confidence.

- In the case of two explanatory variables X_1 and X_2 with slopes β_1 and β_2, the joint confidence region for the slopes takes the form of an ellipse in the $\{\beta_1, \beta_2\}$ plane centred at (B_1, B_2), with equation

$$[B_1 - \beta_1, B_2 - \beta_2] \begin{bmatrix} \sum x_{i1}^{*2} & \sum x_{i1}^* x_{i2}^* \\ \sum x_{i1}^* x_{i2}^* & \sum x_{i2}^{*2} \end{bmatrix} \begin{bmatrix} B_1 - \beta_1 \\ B_2 - \beta_2 \end{bmatrix} \leq 2 S_E^2 F_{\alpha,\, 2,\, n-3}$$

  where the $x_{ij}^* = x_{ij} - \overline{x}_j$ are deviations from the means of X_1 and X_2.
- Figure 6 shows joint-confidence ellipses for two cases: (a) in which X_1 and X_2 are highly correlated, and (b) in which X_1 and X_2 are uncorrelated.
  - The outer ellipse is drawn at a level of confidence of 95%.
  - The inner ellipse (the *confidence-interval generating ellipse*) is drawn so that its perpendicular shadows on the axes are 95% confidence intervals for the individual β's.

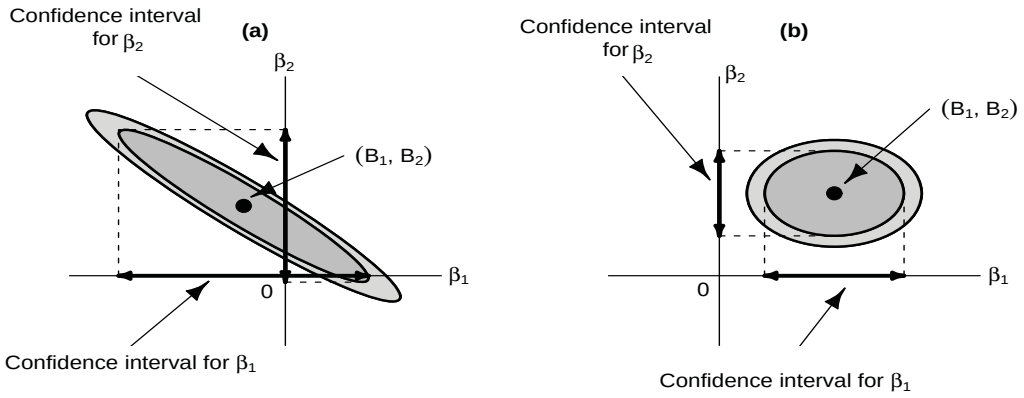

Figure 6. Illustrative joint confidence ellipses for the slope coefficients β_1 and β_2 in multiple-regression analysis. In (a), the X's are positively correlated, producing a joint confidence ellipse that is negatively tilted. In (b), the X's are uncorrelated, producing a joint confidence ellipse with axes parallel to the axes of the parameter space.

- The confidence *interval* for the individual coefficient β_1 can be written:

  all β_1 for which $(B_1 - \beta_1)^2 \frac{\sum x_{i2}^{*2}}{\sum x_{i1}^{*2} \sum x_{i2}^{*2} - \left(\sum x_{i1}^* x_{i2}^*\right)^2} \leq S_E^2 F_{\alpha,\,1,\,n-3}$

  or, more conventionally,

$$B_1 - t_{\alpha,\,n-3} \frac{S_E}{\sqrt{\frac{\sum x_{i1}^{*2}}{1 - r_{12}^2}}} \leq \beta_1 \leq B_1 + t_{\alpha,\,n-3} \frac{S_E}{\sqrt{\frac{\sum x_{i1}^{*2}}{1 - r_{12}^2}}}$$

  - The individual confidence intervals for the regression coefficients are very nearly the perpendicular "shadows" (i.e., projections) of the joint confidence ellipse onto the β_1 and β_2 axes.
  - The only slippage here is due to the right-hand-side constant: $2S_E^2 F_{\alpha,\,2,\,n-3}$ for the joint confidence region, and $S_E^2 F_{\alpha,\,1,\,n-3}$ for the confidence interval.

- For a 95% region and interval, if the residual degrees of freedom $n-3$ are large, then $2F_{.05,\,2,\,n-3} \simeq \chi^2_{.05,\,2} = 5.99$, while $F_{.05,\,1,\,n-3} \simeq \chi^2_{.05,\,1} = 3.84$.
- Put another way, using $5.99S_E^2$ in place of $3.84S_E^2$ produces individual intervals at approximately the $1 - \Pr(\chi^2_1 > 5.99) = .986$ (rather than .95) level of confidence (but a *joint* 95% confidence region).
- If we construct the joint confidence region using the multiplier 3.84, the resulting smaller ellipse produces shadows that give approximate 95% confidence intervals for *individual* coefficients [and a smaller *joint* level of confidence of $1 - \Pr(\chi^2_2 > 3.84) = .853$]. This *confidence-interval generating ellipse* is shown along with the joint confidence ellipse in Figure 6.

▶ The confidence-interval generating ellipse can be projected onto *any* line through the origin of the $\{\beta_1, \beta_2\}$ plane.
- Each line represents a linear combination of β_1 and β_2, and the shadow of the ellipse gives the corresponding confidence interval for that linear combination of the parameters.
- See Figure 7 for the linear combination $\beta_1 + \beta_2$; the line representing $\beta_1 + \beta_2$ is drawn through the origin and the point (1, 1), the coefficients of the parameters in the linear combination.
- Directions in which the ellipse is narrow correspond to linear combinations of the parameters that are relatively precisely estimated.

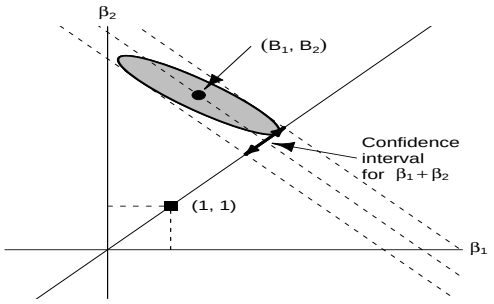

Figure 7. To find the 95% confidence interval for the linear combination of coefficients $\beta_1 + \beta_2$, find the perpendicular shadow of the confidence-interval generating ellipse on the line through the origin and the point (1, 1).

- It is illuminating to examine the relationship between the joint confidence region for the regression coefficients β_1 and β_2, and the data ellipse for X_1 and X_2.
  - The joint confidence ellipse for the slope coefficients and the standard data ellipse of the X's are, except for a constant scale factor and their respective centres, inverses of each other—that is, the confidence ellipse is (apart from its size and location) the $90°$ rotation of the data ellipse.
  - If the data ellipse is positively tilted, reflecting a *positive* correlation between the X's, then the confidence ellipse is negatively tilted, reflecting *negatively* correlated coefficient estimates.
  - Directions in which the data ellipse is relatively *thick*, reflecting a substantial amount of data, are directions in which the confidence ellipse is relatively *thin*, reflecting substantial information about the corresponding linear combination of regression coefficients.

- When the X's are strongly positively correlated (and assuming, for simplicity, that the standard deviations of X_1 and X_2 are similar), there is a great deal of information about $\beta_1 + \beta_2$ but little about $\beta_1 - \beta_2$ (as in Figure 7).

4. Dummy-Variable Regression

4.1 A Dichotomous Explanatory Variable

- The simplest case: one dichotomous and one quantitative explanatory variable.
- Assumptions:
  - Relationships are *additive* — the partial effect of each explanatory variable is the same regardless of the specific value at which the other explanatory variable is held constant.
  - The other assumptions of the regression model hold.

- The motivation for including a qualitative explanatory variable is the same as for including an additional quantitative explanatory variable:
  - to account more fully for the response variable, by making the errors smaller; and
  - to avoid a biased assessment of the impact of an explanatory variable, as a consequence of omitting another explanatory variables that is related to it.

- Figure 8 represents idealized examples, showing the relationship between education and income among women and men.
  - In both cases, the within-gender regressions of income on education are parallel. Parallel regressions imply additive effects of education and gender on income.
  - In (a), gender and education are unrelated to each other: If we ignore gender and regress income on education alone, we obtain the same slope as is produced by the separate within-gender regressions; ignoring gender inflates the size of the errors, however.
  - In (b) gender and education are related, and therefore if we regress income on education alone, we arrive at a biased assessment of the effect of education on income. The overall regression of income on education has a *negative* slope even though the within-gender regressions have positive slopes.

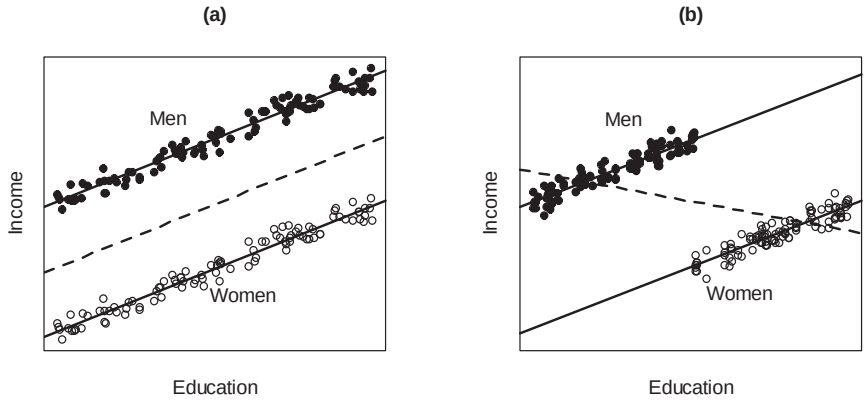

Figure 8. In both cases the within-gender regressions of income on education are parallel: in (a) gender and education are unrelated; in (b) women have higher average education than men.

- We could perform separate regressions for women and men. This approach is reasonable, but it has its limitations:
  - Fitting separate regressions makes it difficult to estimate and test for gender differences in income.
  - Furthermore, if we can assume parallel regressions, then we can more efficiently estimate the common education slope by pooling sample data from both groups.

- One way of formulating the common-slope model is

$$Y_i = \alpha + \beta X_i + \gamma D_i + \varepsilon_i$$

where D, called a *dummy-variable regressor* or an *indicator variable*, is coded 1 for men and 0 for women:

$$D_i = \begin{cases} 1 & \text{for men} \\ 0 & \text{for women} \end{cases}$$

  - Thus, for women the model becomes

$$Y_i = \alpha + \beta X_i + \gamma(0) + \varepsilon_i = \alpha + \beta X_i + \varepsilon_i$$

  - and for men

$$Y_i = \alpha + \beta X_i + \gamma(1) + \varepsilon_i = (\alpha + \gamma) + \beta X_i + \varepsilon_i$$

- These regression equations are graphed in Figure 9.

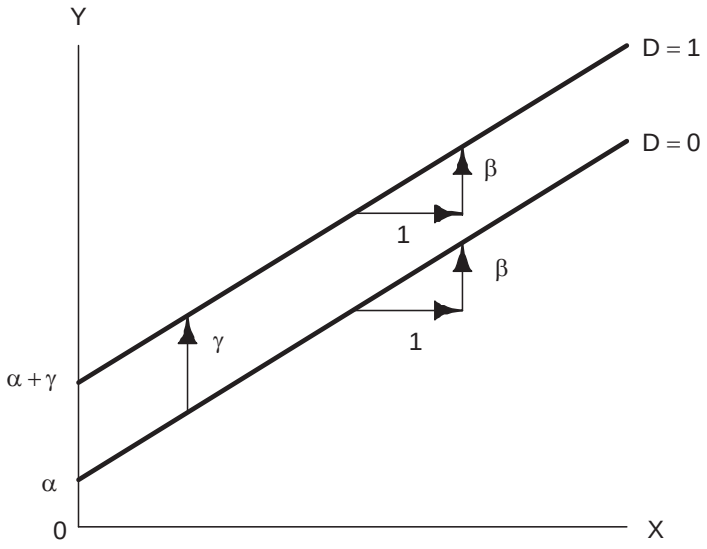

Figure 9. The parameters in the additive dummy-regression model.

4.1.1 Regressors vs. Explanatory Variables

- This is our initial encounter with an idea that is fundamental to many linear models: the distinction between *explanatory variables* and *regressors.*
  - Here, *gender* is a qualitative explanatory variable (or *factor*), with categories (also called *levels*) *male* and *female*.
  - The dummy variable D is a regressor, representing the explanatory variable gender.
  - In contrast, the quantitative explanatory variable (or *covariate*) *income* and the regressor X are one and the same.
- We will see later that an explanatory variable can give rise to several regressors, and that some regressors are functions of more than one explanatory variable.

4.1.2 How Dummy Regression Works

- Interpretation of parameters in the additive dummy-regression model:
  - γ gives the difference in intercepts for the two regression lines.
    - Because these regression lines are parallel, γ also represents the constant separation between the lines — the expected income advantage accruing to men when education is held constant.
    - If men were *dis*advantaged relative to women, then γ would be *negative*.
  - α gives the intercept for women, for whom $D = 0$.
  - β is the common within-gender education slope.

- Essentially similar results are obtained if we code D zero for men and one for women (Figure 10):
  - The sign of γ is reversed, but its magnitude remains the same.
  - The coefficient α now gives the income intercept for men.
  - It is therefore immaterial which group is coded one and which is coded zero.

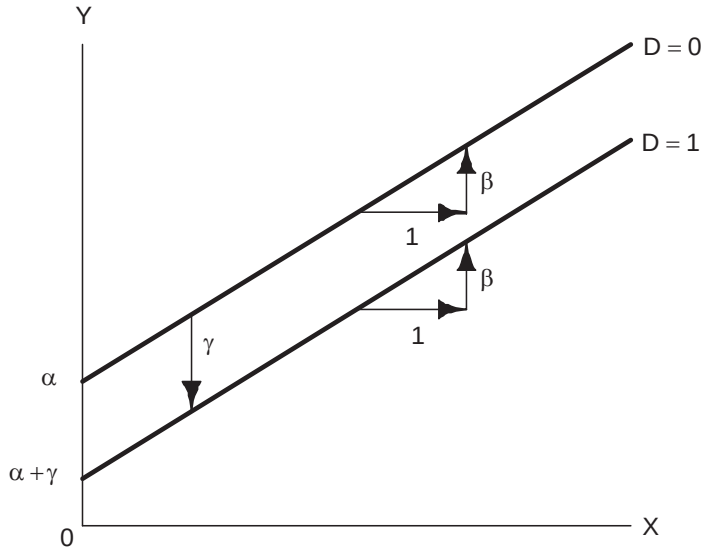

Figure 10. Parameters corresponding to the alternative coding $D = 0$ for men and $D = 1$ for women.

4.2 Polytomous Explanatory Variables

- Consider the regression of the rated prestige of occupations on their income and education levels.
  - Let us classify the occupations into three categories: (1) professional and managerial; (2) 'white-collar'; and (3) 'blue-collar'.
  - The *three*-category classification can be represented in the regression equation by introducing *two* dummy regressors:

Category	D_2	D_3
Blue Collar	0	0
White Collar	1	0
Professional & Managerial	0	1

  - The regression model is then

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_2 D_{i2} + \gamma_3 D_{i3} + \varepsilon_i$$

    where X_1 is income and X_2 is education.

  - This model describes three parallel regression planes, which can differ in their intercepts (see Figure 11):

$$\begin{aligned}
\text{Blue Collar: } Y_i &= \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \\
\text{White Collar: } Y_i &= (\alpha + \gamma_2) + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \\
\text{Professional: } Y_i &= (\alpha + \gamma_3) + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i
\end{aligned}$$

    - α gives the intercept for blue-collar occupations.
    - γ_2 represents the constant vertical distance between the regression planes for white-collar and blue-collar occupations.
    - γ_3 represents the constant vertical difference between the parallel regression planes for professional and blue-collar occupations (fixing the values of education and income).
  - Blue-collar occupations are coded 0 for both dummy regressors, so 'blue collar' serves as a *baseline* category to which the other occupational categories are compared.

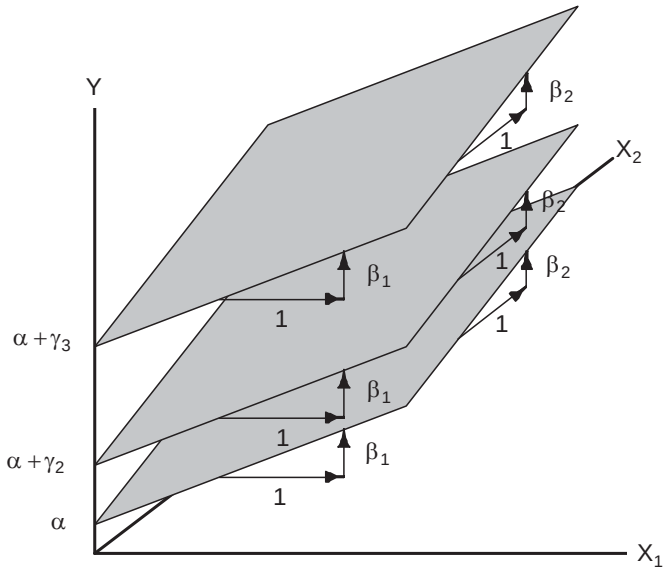

Figure 11. The additive dummy-regression model showing three parallel regression planes.

- The choice of a baseline category is usually arbitrary, for we would fit the same three regression planes regardless of which of the three categories is selected for this role.

▶ Because the choice of baseline is arbitrary, we want to test the null hypothesis of no partial effect of occupational type,

$$H_0\colon \gamma_2 = \gamma_3 = 0$$

but the individual hypotheses $H_0\colon \gamma_2 = 0$ and $H_0\colon \gamma_3 = 0$ are of less interest.

- The hypothesis $H_0\colon \gamma_2 = \gamma_3 = 0$ can be tested by the incremental-sum-of-squares approach, removing D_2 and D_3 from the model.

- For a polytomous explanatory variable with m categories, we code $m-1$ dummy regressors.
  - One simple scheme is to select the first category as the baseline, and to code $D_{ij} = 1$ when observation i falls in category j, and 0 otherwise, for $j = 2, \ldots, m$:

Category	D_2	D_3	$\cdots$	D_m
1	0	0	$\cdots$	0
2	1	0	$\cdots$	0
.	.	.		.
.	.	.		.
.	.	.		.
m	0	0	$\cdots$	1

  - To test the hypothesis that the effects of a qualitative explanatory variable are nil, delete its dummy regressors from the model and compute an incremental F-test.

4.3 Modeling Interactions

- Two explanatory variables *interact* in determining a response variable when the partial effect of one depends on the value of the other.
  - Additive models specify the absence of interactions.
  - If the regressions in different categories of a qualitative explanatory variable are not parallel, then the qualitative explanatory variable interacts with one or more of the quantitative explanatory variables.
  - The dummy-regression model can be modified to reflect interactions.
- Consider the hypothetical data in Figure 12 (and contrast these examples with those shown in Figure 8, where the effects of gender and education were additive):
  - In (a), gender and education are independent, since women and men have identical education distributions.
  - In (b), gender and education are related, since women, on average, have higher levels of education than men.

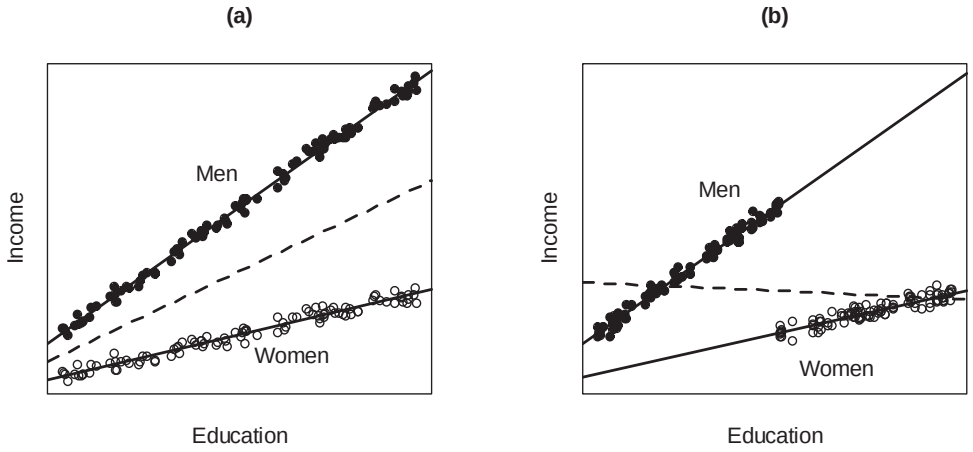

Figure 12. In both cases, gender and education interact in determining income. In (a) gender and education are independent; in (b) women on average have more education than men.

- In both (a) and (b), the within-gender regressions of income on education are not parallel — the slope for men is larger than the slope for women.
  - Because the effect of education varies by gender, education and gender interact in affecting income.
- It is also the case that the effect of gender varies by education. Because the regressions are not parallel, the relative income advantage of men changes with education.
  - *Interaction is a symmetric concept — the effect of education varies by gender, and the effect of gender varies by education.*

- These examples illustrate another important point: *Interaction* and *correlation* of explanatory variables are empirically and logically distinct phenomena.
  - Two explanatory variables can interact *whether or not* they are related to one-another statistically.
  - Interaction refers to the manner in which explanatory variables combine to affect a response variable, not to the relationship between the explanatory variables themselves.

4.3.1 Constructing Interaction Regressors

- We could model the data in the example by fitting separate regressions of income on education for women and men.
  - A combined model facilitates a test of the gender-by-education interaction, however.
  - A properly formulated unified model that permits different intercepts and slopes in the two groups produces the same fit to the data as separate regressions.
- The following model accommodates different intercepts and slopes for women and men:

$$Y_i = \alpha + \beta X_i + \gamma D_i + \delta(X_i D_i) + \varepsilon_i$$

  - Along with the dummy regressor D for gender and the quantitative regressor X for education, I have introduced the *interaction regressor* XD.

- The interaction regressor is the *product* of the other two regressors: XD is a function of X and D, but it is not a *linear* function, avoiding perfect collinearity.
- For women,

$$\begin{aligned} Y_i &= \alpha + \beta X_i + \gamma(0) + \delta(X_i \cdot 0) + \varepsilon_i \\ &= \alpha + \beta X_i + \varepsilon_i \end{aligned}$$

- and for men,

$$\begin{aligned} Y_i &= \alpha + \beta X_i + \gamma(1) + \delta(X_i \cdot 1) + \varepsilon_i \\ &= (\alpha + \gamma) + (\beta + \delta) X_i + \varepsilon_i \end{aligned}$$

▶ These regression equations are graphed in Figure 13:

- α and β are the intercept and slope for the regression of income on education among women.
- γ gives the *difference* in intercepts between the male and female groups
- δ gives the *difference* in slopes between the two groups.

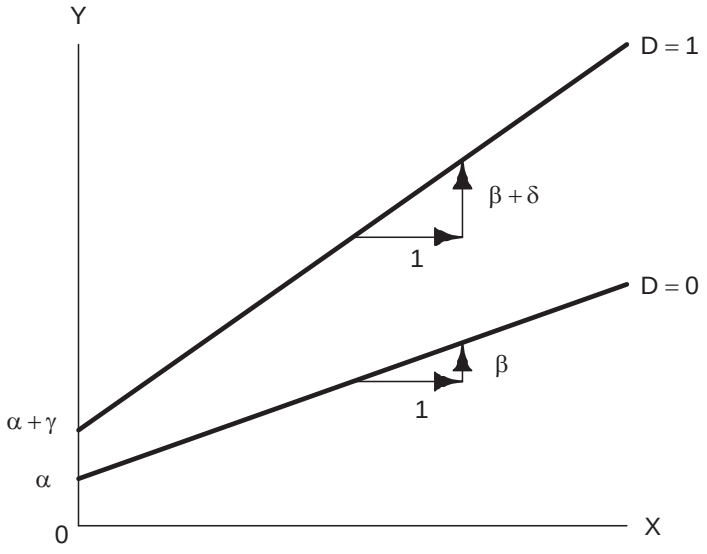

Figure 13. The parameters in the dummy-regression model with interaction.

  - To test for interaction, we can test the hypothesis H_0: $\delta = 0$.

- In the additive, no-interaction model, γ represented the unique partial effect of gender, while the slope β represented the unique partial effect of education.
  - In the interaction model, γ is no longer interpretable as the unqualified income difference between men and women of equal education — γ is now the income difference at $X = 0$.
  - Likewise, in the interaction model, β is not the unqualified partial effect of education, but rather the effect of education among women.
    - The effect of education among men ($\beta + \delta$) does not appear directly in the model.
- Extension to polytomous factors is straight-forward.

4.4 The Principle of Marginality

- The separate partial effects, or *main effects*, of education and gender are *marginal* to the education-by-gender interaction.
- In general, we neither test nor interpret main effects of explanatory variables that interact.
  - If we can rule out interaction either on theoretical or empirical grounds, then we can proceed to test, estimate, and interpret main effects.
- It does not generally make sense to specify and fit models that include interaction regressors but that delete main effects that are marginal to them.
  - Such models — which violate the *principle of marginality* — are interpretable, but they are not broadly applicable.

- Consider the model

$$Y_i = \alpha + \beta X_i + \delta(X_i D_i) + \varepsilon_i$$

  - As shown in Figure 14 (a), this model describes regression lines for women and men that have the same intercept but (potentially) different slopes, a specification that is peculiar and of no substantive interest.

- Similarly, the model

$$Y_i = \alpha + \gamma D_i + \delta(X_i D_i) + \varepsilon_i$$

  graphed in Figure 14 (b), constrains the slope for women to 0, which is needlessly restrictive.

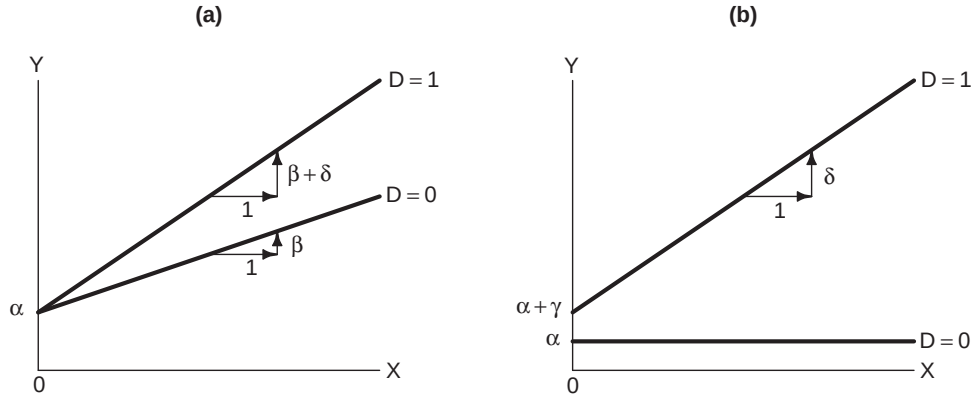

Figure 14. Two models that violate the principle of marginality, by including the interaction regressor XD but (a) omitting D or (b) omitting X.

5. Regression Diagnostics

- Linear statistical models make strong assumptions about the structure of data, which often do not hold in applications.
- For example, the method of least-squares is very sensitive to the structure of the data, and can be markedly influenced by one or a few unusual observations.
- We could abandon linear models and least-squares estimation in favor of nonparametric regression and robust estimation.
- Alternatively, we can use "diagnostic" methods to detect problems and to suggest solutions.

5.1 Unusual Data

- Unusual data are problematic in linear models fit by least squares because they can unduly influence the results of the analysis, and because their presence may be a signal that the model fails to capture important characteristics of the data.
- Some central distinctions are illustrated in Figure 15 for the simple regression model $Y = \alpha + \beta X + \varepsilon$.
  - In simple regression, an *outlier* is an observation whose response-variable value is conditionally unusual given the value of the explanatory variable.
  - In contrast, a univariate outlier is a value of Y or X that is unconditionally unusual; such a value may or may not be a regression outlier.

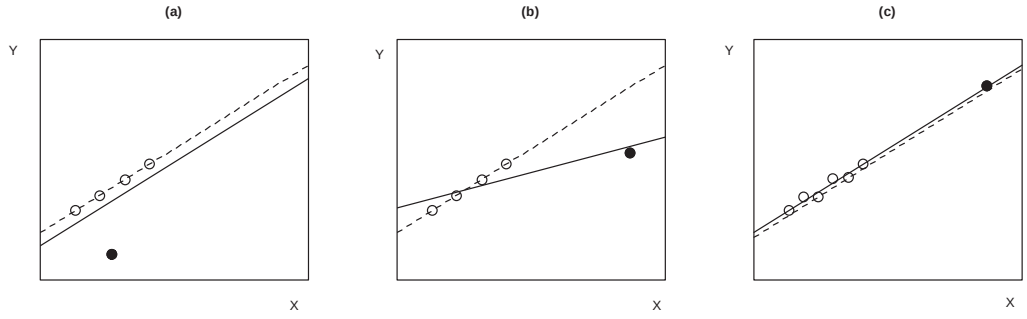

Figure 15. Unusual data in regression: (a) a low-leverage and hence uninfluential outlier; (b) a high-leverage and hence influential outlier; (c) a high-leverage in-line observation. In each case, the solid line is the least--squares line for all of the data; the broken line is the least-squares line with the unusual observation omitted.

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

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

  Influence on Coefficients = Leverage $\times$ Discrepancy

5.1.1 Assessing Leverage: Hat-Values

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

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

- Thus, the weight h_{ij} captures the contribution of observation Y_i to the fitted value $\widehat{Y}_j$: If h_{ij} is large, then the ith observation can have a substantial impact on the jth fitted value.

▶ Properties of the hat-values:

- $h_{ii} = \sum_{j=1}^{n} h_{ij}^2$, and so the hat-value $h_i \equiv h_{ii}$ summarizes the potential influence (the leverage) of Y_i on *all* of the fitted values.
- $1/n \leq h_i \leq 1$
- The average hat-value is $\overline{h} = (k+1)/n$.

- Belsley, Kuh, and Welsch suggest that hat-values exceeding about twice the average (or, in small samples, three times the average) hat-value are noteworthy.
- In simple-regression analysis, the hat-values measure distance from the mean of X:

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

- In multiple regression, h_i measures distance from the centroid (point of means) of the X's, taking into account the correlational and variational structure of the X's, as illustrated for $k = 2$ in Figure 16. Multivariate outliers in the X-space are thus high-leverage observations. The response-variable values are not at all involved in determining leverage.

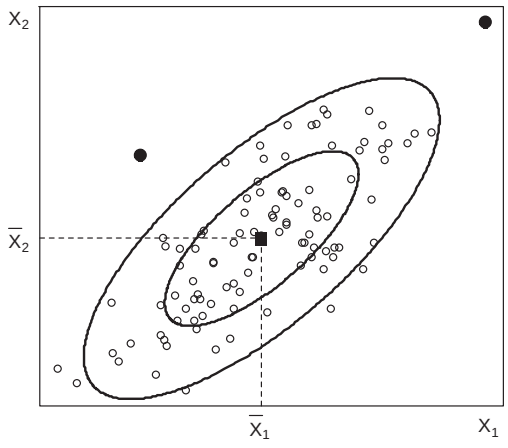

Figure 16. Contours of constant leverage in multiple regression with two explanatory variables, X_1 and X_2. The two observations marked with solid black dots have equal hat-values.

5.1.2 Detecting Outliers: Studentized Residuals

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

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

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

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

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

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

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

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

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

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

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

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

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

- Thus

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

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

- In most applications we want to look for *any* outliers that may occur in the data; we can in effect refit the mean-shift model n times, producing studentized residuals $E_1^*, E_2^*, ..., E_n^*$. (It is not literally necessary to perform n auxiliary regressions.)
  - Usually, our interest then focuses on the largest absolute E_i^*, denoted $E_{\max}^*$.
  - Because we have picked the biggest of n test statistics, it is not legitimate simply to use t_{n-k-2} to find a p-value for $E_{\max}^*$.
- One solution to this problem of simultaneous inference is to perform a *Bonferroni adjustment* to the p-value for the largest absolute E_i^*: Let $p' = \Pr(\, t_{n-k-2} > E_{\max}^*)$.
  - Then the Bonferroni p-value for testing the statistical significance of $E_{\max}^*$ is $p = 2np'$.
  - Note that a much larger $E_{\max}^*$ is required for a statistically significant result than would be the case for an ordinary individual t-test.

- Another approach is to construct a quantile-comparison plot for the studentized residuals, plotting against either the t or normal distribution.

5.1.3 Measuring Influence

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

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

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

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

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

  - In effect, the first term in the formula for Cook's D is a measure of discrepancy, and the second is a measure of leverage.
  - We look for values of D_i that are substantially larger than the rest..
  - Work by Chatterjee and Hadi implies that $D_i > 4/(n-k-1)$ are noteworthy.

5.1.4 Joint Influence: Added-Variable Plots

- As illustrated in Figure 17, subsets of observations can be *jointly influential* or can offset each other's influence.
  - Influential subsets or multiple outliers can often be identified by applying single-observation diagnostics, such as Cook's D and studentized residuals, sequentially.
  - It can be important to refit the model after deleting each point, because the presence of a single influential value can dramatically affect the fit at other points, but the sequential approach is not always successful.
- Although it is possible to generalize deletion statistics to subsets of several points, the very large number of subsets usually renders this approach impractical.
- An attractive alternative is to employ graphical methods, and a particularly useful influence graph is the *added-variable plot* (also called a *partial-regression plot* or an *partial-regression leverage plot*).

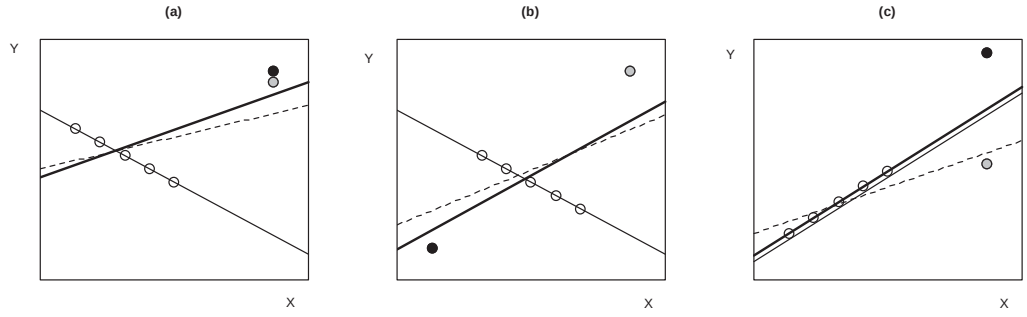

Figure 17. Jointly influential observations: (a) a pair of jointly influential points; (b) a widely separated jointly infuential pair; (c) two points that offset each other's influence. In each case the heavier solid line is the least-squares line for all of the data, the broken line deletes the black point, and the lighter solid line deletes both the gray and the black points.

- Let $Y_i^{(1)}$ represent the residuals from the least-squares regression of Y on all of the X's with the exception of X_1:

$$Y_i = A^{(1)} + B_2^{(1)}X_{i2} + \cdots + B_k^{(1)}X_{ik} + Y_i^{(1)}$$

- Likewise, $X_i^{(1)}$ are the residuals from the least-squares regression of X_1 on all the other X's:

$$X_{i1} = C^{(1)} + D_2^{(1)}X_{i2} + \cdots + D_k^{(1)}X_{ik} + X_i^{(1)}$$

- The notation emphasizes the interpretation of the residuals $Y^{(1)}$ and $X^{(1)}$ as the parts of Y and X_1 that remain when the effects of $X_2, ..., X_k$ are 'removed.'
- The residuals $Y^{(1)}$ and $X^{(1)}$ have the following interesting properties:

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

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

- Plotting $Y^{(1)}$ against $X^{(1)}$ permits us to examine leverage and influence on B_1. Because of properties 1–3, this plot also provides a visual impression of the precision of estimation of B_1.
- Similar added-variable plots can be constructed for the other regression coefficients:
Plot $Y^{(j)}$ versus $X^{(j)}$ for each $j = 0, ..., k$

5.1.5 Influence on Other Regression "Outputs"

- I have focussed on influence of observations on regression coefficients, but it is possible to consider influence on other regression "outputs" such as correlations and coefficient standard errors.
  - For example, an in-line (i.e., non-outlying) high-leverage observation serves to increase the precision — or, perhaps, apparent precision — of estimation, e.g., by increasing the variation of one or more explanatory variables or by decreasing collinearity among them.
  - In contrast, an outlier at a low-leverage point decreases the precision of estimation of the regression coefficients by inflating the standard error of the regression.
  - In both of these cases, the observation in question may not exert much influence at all on the values of the coefficients.

5.2 Non-Normally Distributed Errors

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

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

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

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

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

5.2.1 Transformations: The Family of Powers and Roots

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

$$Y \rightarrow Y^p$$

  - If p is negative, then the transformation is an inverse power: $Y^{-1} = 1/Y$, and $Y^{-2} = 1/Y^2$.
  - If p is a fraction, then the transformation represents a root: $Y^{1/3} = \sqrt[3]{Y}$ and $Y^{-1/2} = 1/\sqrt{Y}$.
- It is sometimes convenient to define the family of power transformations in a slightly more complex manner (called the *Box-Cox family*):

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

- Since $Y^{(p)}$ is a linear function of Y^p, the two transformations have the same essential effect on the data, but, as is apparent in Figure 18, $Y^{(p)}$ reveals the essential unity of the family of powers and roots:

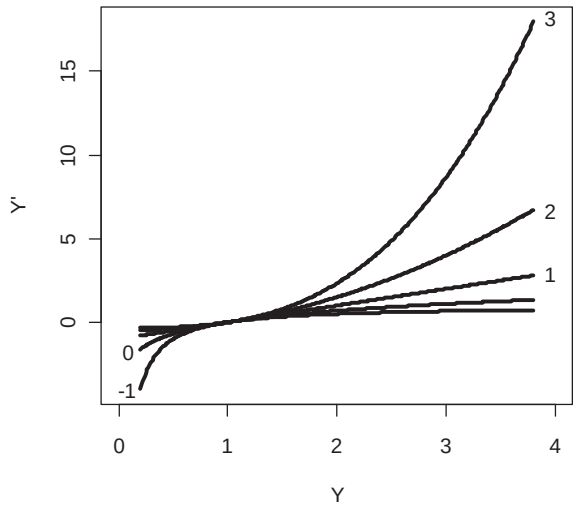

Figure 18. The Box-Cox familily of modified power transformations, $Y^{(p)} = (Y^p - 1)/p$, for values of $p = -1, 0, 1, 2, 3$. When $p = 0$, $Y^{(p)} = \log_e Y$.

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

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

- The transformations $Y^{(p)}$ are 'matched' above $Y = 1$ both in level and slope.
- The power transformation $Y^0 = 1$ is useless, but the very useful $\log$ transformation is a kind of 'zeroth' power:

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

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

▶ 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.

▶ Power transformations require that all of the data are positive; to be effective, the ratio of largest to smallest value cannot be too small.

5.3 Non-Constant Error Variance

- Although the least-squares estimator is unbiased and consistent even when the error variance is not constant, its efficiency is impaired, and the usual formulas for coefficient standard errors are inaccurate.
  - Non-constant error variance is sometimes termed 'heteroscedasticity.'
- Because the regression surface is k-dimensional, and imbedded in a space of $k + 1$ dimensions, it is generally impractical to assess the assumption of constant error variance by direct graphical examination of the data.
- It is common for error variance to increase as the expectation of Y grows larger, or there may be a systematic relationship between error variance and a particular X.
  - The former situation can often be detected by plotting residuals against fitted values;
  - the latter by plotting residuals against each X.

- Plotting residuals against Y (as opposed to $\widehat{Y}$) is generally unsatisfactory, because the plot will be 'tilted'
  - There is a built-in linear correlation between Y and E, since $Y = \widehat{Y} + E$.
  - The least-squares fit insures that the correlation between $\widehat{Y}$ and E is zero, producing a plot that is much easier to examine for evidence of non-constant spread.
- Because the *residuals* have unequal variances even when the variance of the *errors* is constant, it is preferable to plot studentized residuals against fitted values.
- It often helps to plot $|E_i^*|$ or E_i^{*2} against $\widehat{Y}$.
- Following a suggestion byTukey, one can alternatively construct a *spread-level plot*, graphing log absolute studentized residuals against log fitted values (as long as all of the fitted values are positive).

- Descending the ladder of powers and roots can eliminate a positive association between residual spread and the level of the response.

5.4 Nonlinearity

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

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

5.4.1 Component+Residual Plots

- Although it is useful in multiple regression to plot Y against each X, these plots can be misleading, because our interest centres on the *partial* relationship between Y and each X, controlling for the other X's, not on the *marginal* relationship between Y and an individual X, ignoring the other X's.
- Plotting residuals or studentized residuals against each X is frequently helpful for detecting departures from linearity.
  - As Figure 19 illustrates, however, residual plots cannot distinguish between monotone and non-monotone nonlinearity.
    - The distinction is important because monotone nonlinearity frequently can be 'corrected' by simple transformations.
    - Case (a) might be modeled by $Y = \alpha + \beta\sqrt{X} + \varepsilon$.

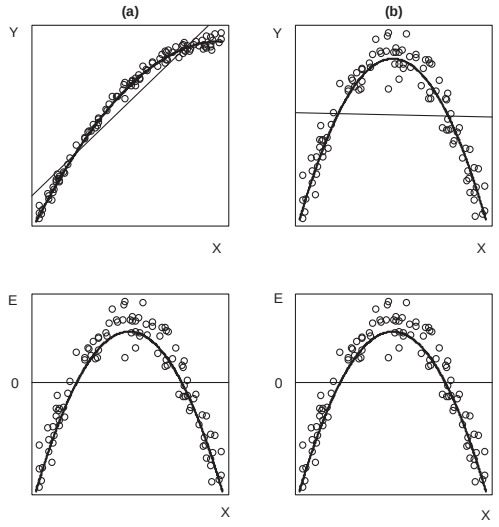

Figure 19. The residual plots of E versus X (bottom) are identical, even though the regression of Y on X in (a) is monotone while that in (b) is non-monotone.

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

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

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

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

5.4.2 The Bulging Rule

▶ 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 20.
- 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)].

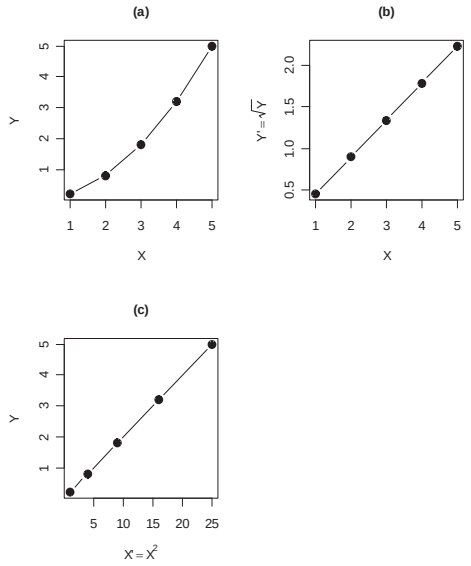

Figure 20. Transformating a nonlinear relationship (a) to linearity, (b) or (c).

- A power transformation works here because the relationship between Y and X is both monotone and simple. In Figure 21:
  - the curve in (a) is simple and monotone;
  - in (b) monotone, but not simple;
  - in (c) simple but not monotone.
    - In (c), we could fit a quadratic model, $\widehat{Y} = a + b_1X + b_2X^2$.
- Figure 22 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).
  - In multiple regression, we generally prefer to transform an X (and to leave Y alone).

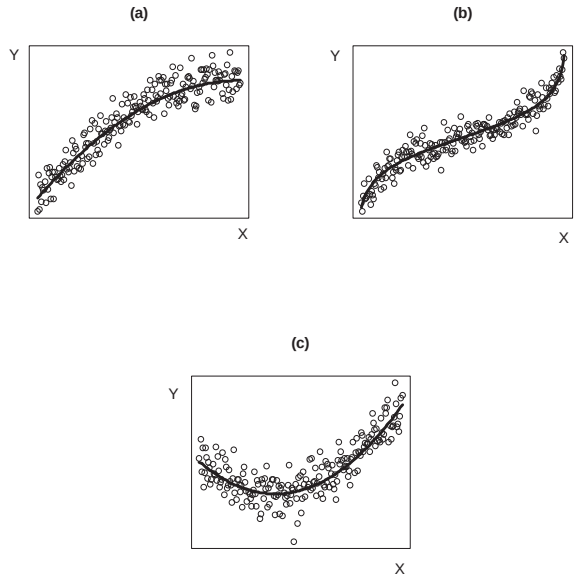

Figure 21. (a) A simple monotone relationship. (b) A monotone relationship that is not simple. (c) A simple nonmonotone relationship.

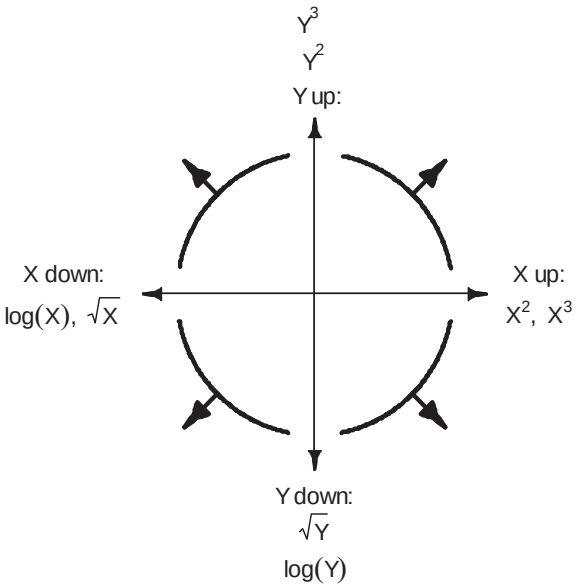

Figure 22. Mosteller and Tukey's bulging rule for selecting linearizing transformations.

6. Implementation of Linear Models in R

- The `lm()` function (with important arguments):

```
lm(formula, data, subset, weights, na.action, contrasts)
```

where:

- `formula` is a model formula, specifying the regression equation to be fit (see below).
- `data` is an optional data frame containing the data for the model, which otherwise are located on the search path if `lm()` is called from the command prompt.
- `subset` is an optional specification (e.g., in the form of a logical vector or a vector of positive or negative subscripts) of the subset of observations to which the model is to be fit.
- `weights` is an optional vector of weights for weighted-least-squares (WLS) estimation.

- `na.action` is an optional function to handle missing data; defaults to `na.omit` (unless the global `na.action` option is changed).
- `contrasts` is an optional list specifying contrast functions for specific factors in the model, which otherwise are taken from the factors themselves (if they have `contrasts` attributes) or from the global `contrasts` option, which defaults to `contr.treatment` (dummy coding) for factors and `contr.poly` (orthogonal-polynomial coding) for ordered factors.

- A model formula is of the form

```
lhs ~ rhs
```

where `lhs` is an expression evaluating to the response variable [e.g., `income`, `log(income)`], and `rhs` specifies the "terms" in the right-hand side of the model using operators in the following table [e.g., `poly(age, 2) + gender*(education + experience)`]:

Expression	*Interpretation*	*Example*
`A + B`	include both `A` and `B`	`income + education`
`A - B`	exclude `B` from `A`	`a*b*d - a:b:d`
`A:B`	all interactions of `A` and `B`	`type:education`
`A*B`	`A + B + A:B`	`type*education`
`B %in% A`	`B` nested within `A`	`education %in% type`
`A/B`	`A + B %in% A`	`type/education`
`A^k`	all effects crossed up to order `k`	`(a + b + d)^2`

- The arithmetic operators therefore have special meaning on the right-hand side of a model formula.
- To do arithmetic on the right-hand side of a formula, it is necessary to "protect" the operation within a function call [e.g., `log(income + 1)` or `I(income^2)`, where `I()` is the i*dentity function*].
- We say that "`lhs` is modeled as `rhs`" or that "`lhs` is regressed on `rhs`."

