

York SPIDA

John Fox

Notes

Review of Linear Models

Copyright © 2010 by John Fox

1. Topics

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

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 generalized 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:

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

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

- Written as a matrix equation for n observations:

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

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

- \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.** $V(\varepsilon_i) = \sigma_\varepsilon^2$, which implies that

$$V(Y_i | X_{i1}, \dots, X_{ik}) = \sigma_\varepsilon^2$$

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

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

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 $\boldsymbol{\varepsilon} \sim \mathbf{N}_n(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_n)$.
- **Fixed X 's or X 's independent of ε .**

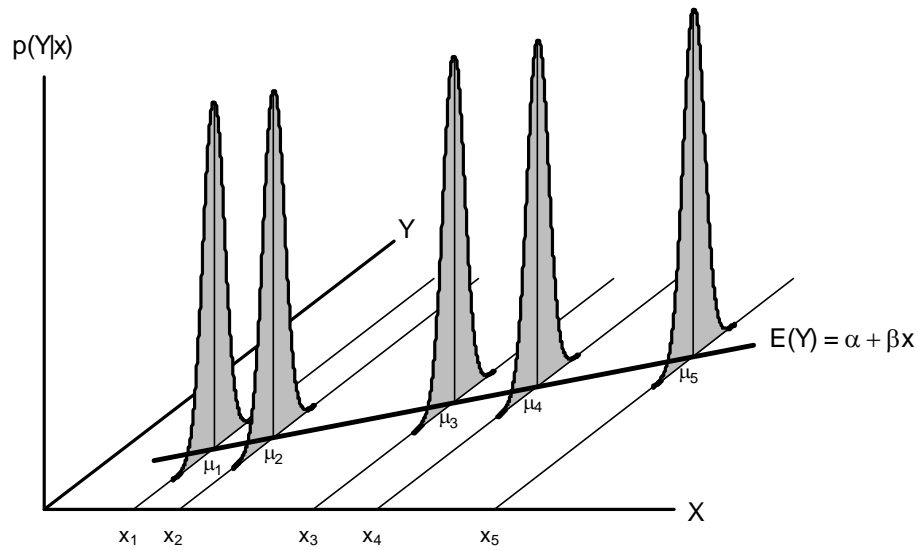


Figure 1. The assumptions of linearity, normality, and constant variance in the simple-regression model.

© 2010 by John Fox

York SPIDA

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

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

with covariance matrix

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

- 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

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

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

$$\begin{aligned}\hat{\mathbf{y}} &= \{\hat{Y}_i\} = \mathbf{X}\mathbf{b} \\ \mathbf{e} &= \{E_i\} = \mathbf{y} - \hat{\mathbf{y}}\end{aligned}$$

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 :

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

- 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:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$$

which is not quite the same as testing the separate hypotheses

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

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

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

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

- Then $R^2 = \text{RegSS}/\text{TSS}$ 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	$\frac{\text{RegSS}}{k}$	$\frac{\text{RegMS}}{\text{RMS}}$
Residuals	RSS	$n - k - 1$	$\frac{\text{RSS}}{n - k - 1}$	
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:

$$E(F_0) \approx \frac{E(\text{RegMS})}{E(\text{RMS})} = \frac{\sigma_\varepsilon^2 + \text{positive quantity}}{\sigma_\varepsilon^2}$$

- 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

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

where $1 \leq q \leq k$.

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

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

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

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

- 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, $RSS_0 \geq RSS_1$. Equivalently, $RegSS_0 \leq RegSS_1$.
 - If the null hypothesis is wrong and (some of) β_1, \dots, β_q are nonzero, then the *incremental* (or 'extra') *sum of squares* due to fitting the additional explanatory variables

$$RSS_0 - RSS_1 = RegSS_1 - RegSS_0$$
 should be large.

- The F -statistic for testing the null hypothesis is

$$\begin{aligned}
 F_0 &= \frac{\frac{RegSS_1 - RegSS_0}{q}}{\frac{RSS_1}{n - k - 1}} \\
 &= \frac{n - k - 1}{q} \times \frac{R_1^2 - R_0^2}{1 - R_1^2}
 \end{aligned}$$

- 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

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

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.”

3.1 The Standard Data Ellipse

- Consider the quadratic form $(\mathbf{x} - \bar{\mathbf{x}})' \mathbf{S}_{XX}^{-1} (\mathbf{x} - \bar{\mathbf{x}})$, where \mathbf{x} is a $k \times 1$ vector of explanatory-variable values, $\bar{\mathbf{x}}$ 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

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

- 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.

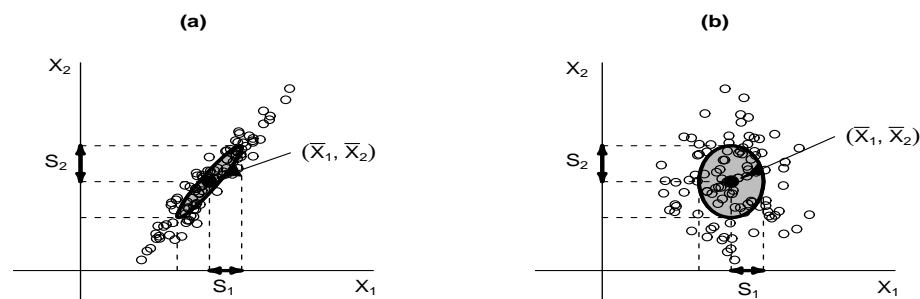


Figure 2. Scatterplot and standard data ellipse for (a) two highly correlated variables and (b) two uncorrelated variables, X_1 and X_2 . In each panel, the standard ellipse is centred at the point of means (\bar{X}_1, \bar{X}_2) ; its shadows on the axes give the standard deviations of the two variables.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

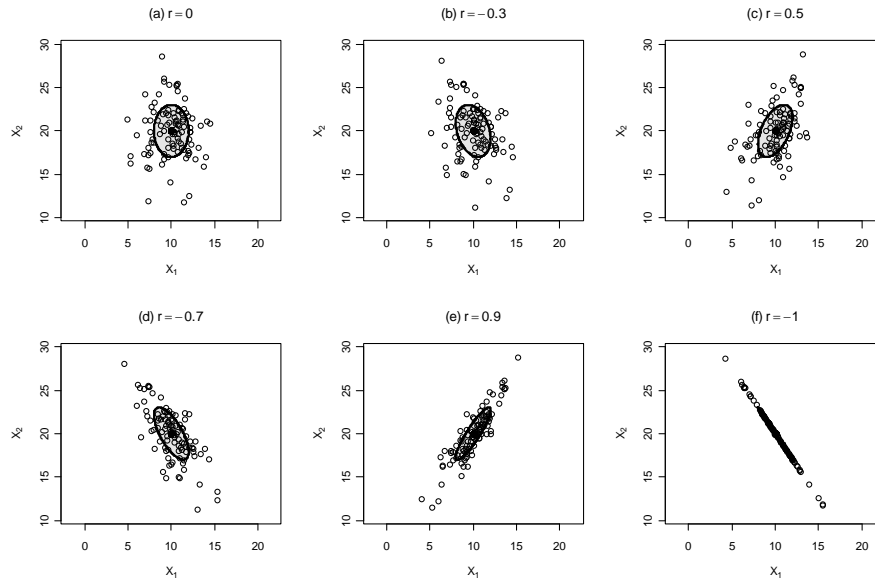


Figure 3. Scatterplots and standard data ellipses corresponding to different correlations. In each case, $\bar{X}_1 = 10$, $\bar{X}_2 = 20$, $SD(X_1) = 2$, and $SD(X_2) = 3$.

© 2010 by John Fox

York SPIDA

- ▶ 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 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 distributions of Y fixing the value of X , and the bisectors of these slices given the conditional means, $\bar{Y}|x$.
 - As a consequence, the least-squares line goes through the points of vertical tangency of the ellipse.

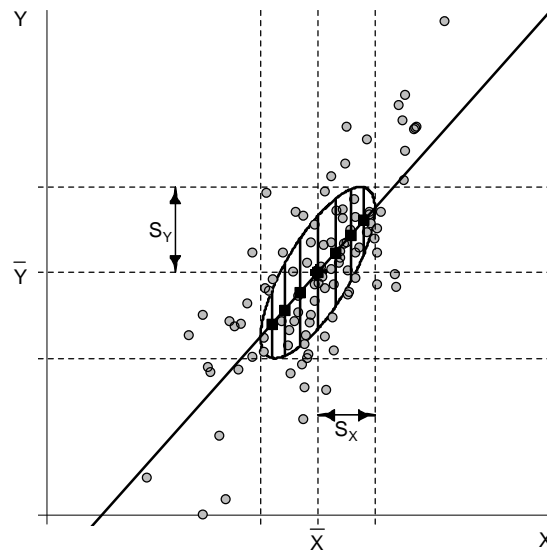


Figure 4. The least-squares line goes through the vertical bisectors and the points of vertical tangency of the standard data ellipse.

© 2010 by John Fox

York SPIDA

- ▶ 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 (deregarding 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 .

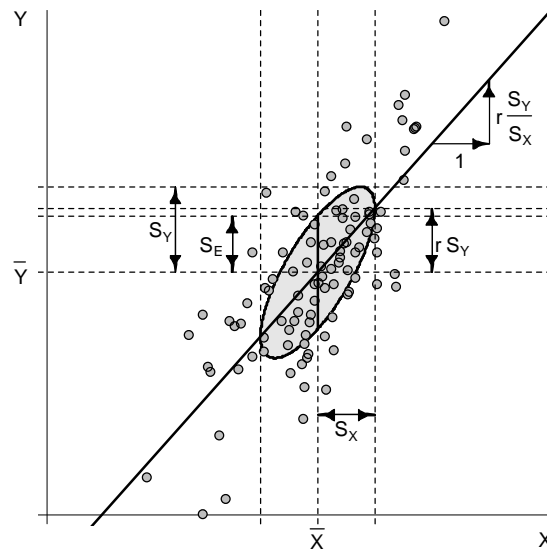


Figure 5. The standard data ellipse illuminates many characteristics of linear least-squares regression and correlation.

© 2010 by John Fox

York SPIDA

3.2 Joint Confidence Regions

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

$$F_0 = \frac{(\mathbf{b}_1 - \beta_1^{(0)})' \mathbf{V}_{11}^{-1} (\mathbf{b}_1 - \beta_1^{(0)})}{k S_E^2}$$

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 $100(1 - a)\%$ joint confidence region for the regression parameters β_1 :

$$\Pr \left[\frac{(\mathbf{b}_1 - \beta_1^{(0)})' \mathbf{V}_{11}^{-1} (\mathbf{b}_1 - \beta_1^{(0)})}{k S_E^2} \leq F_{a, k, n-k-1} \right] = 1 - a$$

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

© 2010 by John Fox

York SPIDA

- The joint confidence region for β_1 is thus

$$\text{all } \beta_1 \text{ for which } (\mathbf{b}_1 - \beta_1)' \mathbf{V}_{11}^{-1} (\mathbf{b}_1 - \beta_1) \leq k S_E^2 F_{a, k, n-k-1}$$

- 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 β_1, \dots, β_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

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

where the $x_{ij}^* = x_{ij} - \bar{x}_j$ 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.

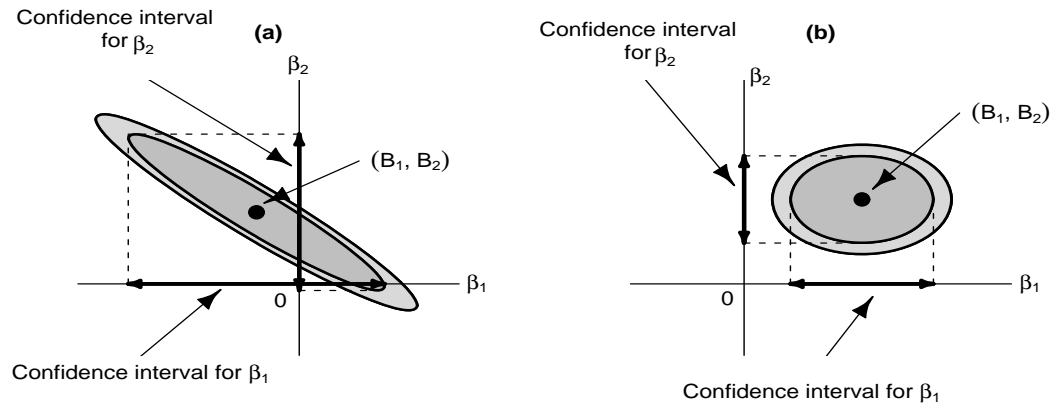


Figure 6. Illustrative joint confidence ellipses for the slope coefficients β_1 and β_2 in multiple-regression analysis. In (a), the X 's are positively correlated, producing a joint confidence ellipse that is negatively tilted. In (b), the X 's are uncorrelated, producing a joint confidence ellipse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

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

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

or, more conventionally,

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

- 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: $2S_E^2 F_{a,2,n-3}$ for the joint confidence region, and $S_E^2 F_{a,1,n-3}$ for the confidence interval.

- For a 95% region and interval, if the residual degrees of freedom $n - 3$ are large, then $2F_{.05, 2, n-3} \simeq \chi_{.05, 2}^2 = 5.99$, while $F_{.05, 1, n-3} \simeq \chi_{.05, 1}^2 = 3.84$.
- Put another way, using $5.99S_E^2$ in place of $3.84S_E^2$ produces individual intervals at approximately the $1 - \Pr(\chi_1^2 > 5.99) = .986$ (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.

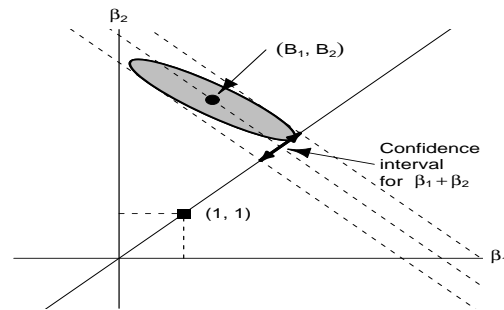


Figure 7. To find the 95% confidence interval for the linear combination of coefficients $\beta_1 + \beta_2$, find the perpendicular shadow of the confidence-interval generating ellipse on the line through the origin and the point $(1, 1)$.

- ▶ 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. 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.

4.1 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 8 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.

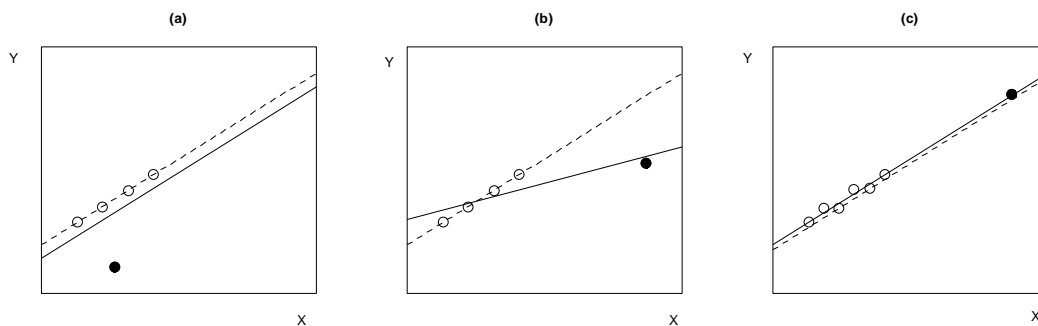


Figure 8. Unusual data in regression: (a) a low-leverage and hence un-influential outlier; (b) a high-leverage and hence influential outlier; (c) a high-leverage in-line observation. In each case, the solid line is the least-squares line for all of the data; the broken line is the least-squares line 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'):
$$\text{Influence on Coefficients} = \text{Leverage} \times \text{Discrepancy}$$

4.1.1 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 \hat{Y} ('Y-hat') in terms of the observed values Y :

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

- Thus, the weight h_{ij} captures the contribution of observation Y_i to the fitted value \hat{Y}_j : If h_{ij} is large, then the i th observation can have a substantial impact on the j th 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 $\bar{h} = (k + 1)/n$.

- 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 :

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

- 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 9. Multivariate outliers in the X -space are thus high-leverage observations. The response-variable values are not at all involved in determining leverage.

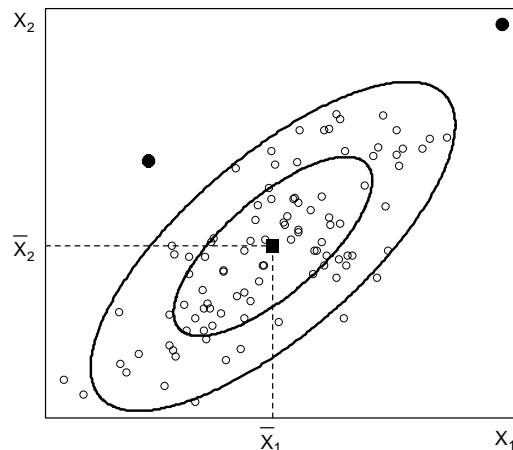


Figure 9. Contours of constant leverage in multiple regression with two explanatory variables, X_1 and X_2 . The two observations marked with solid black dots have equal hat-values.

4.1.2 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 i th 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 $t_0 = \hat{\gamma}/\text{SE}(\hat{\gamma})$. 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 $E_1^*, E_2^*, \dots, E_n^*$. (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.

4.1.3 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:

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

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 i th 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 $j = 0, 1, \dots, k$.
 - This statistic is recalculated for each observation $i = 1, \dots, n$.
 - 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.

4.1.4 Joint Influence: Added-Variable Plots

- ▶ As illustrated in Figure 10, 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*).

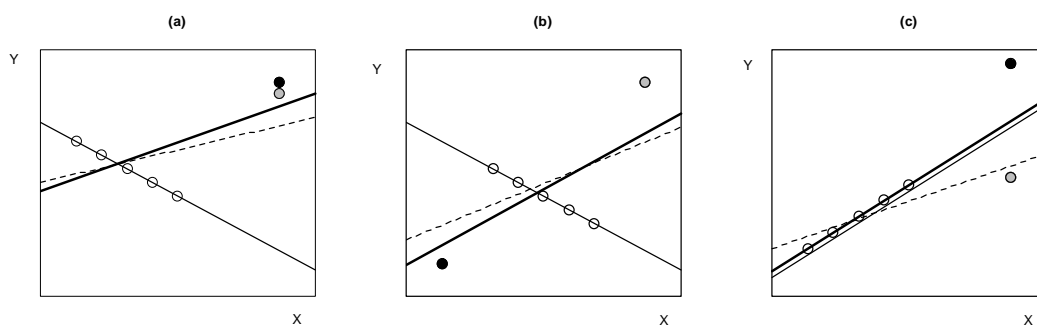


Figure 10. Jointly influential observations: (a) a pair of jointly influential points; (b) a widely separated jointly influential pair; (c) two points that offset each other's influence. In each case the heavier solid line is the least-squares line for all of the data, the broken line deletes the black 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, $X_{i1}^{(1)}$ are the residuals from the least-squares regression of X_1 on all the other X 's:

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

- 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, \dots, 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

$$SE(B_1) = \frac{S_E}{\sqrt{\sum X_i^{(1)2}}}$$

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 $j = 0, \dots, k$

4.1.5 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.

4.2 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.

4.2.1 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 11, $Y^{(p)}$ reveals the essential unity of the family of powers and roots:

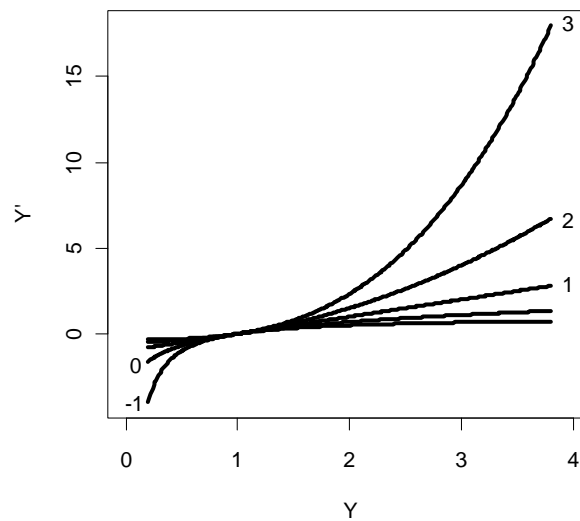


Figure 11. The Box-Cox family of modified power transformations, $Y^{(p)} = (Y^p - 1)/p$, for values of $p = -1, 0, 1, 2, 3$. When $p = 0$, $Y^{(p)} = \log_e Y$.

- 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:

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

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.

4.3 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 \hat{Y}) is generally unsatisfactory, because the plot will be ‘tilted’
 - There is a built-in linear correlation between Y and E , since $Y = \hat{Y} + E$.
 - The least-squares fit insures that the correlation between \hat{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 \hat{Y} .
- Following a suggestion by Tukey, 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.

4.4 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 $E(Y|X_1, \dots, X_k)$ 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 $\{Y_i, X_{i1}, \dots, X_{ik}\}$.

4.4.1 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 12 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 $Y = \alpha + \beta\sqrt{X} + \varepsilon$.

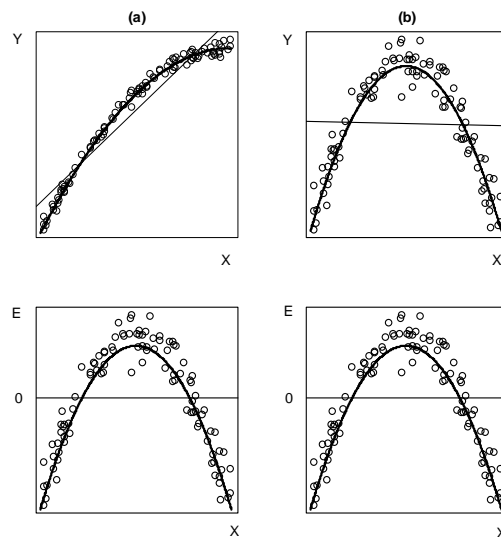


Figure 12. The residual plots of E versus X (bottom) are identical, even though the regression of Y on X in (a) is monotone while that in (b) is non-monotone.

- 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 j th 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 $E_i^{(j)}$ versus X_j .
- By construction, the multiple-regression coefficient B_j is the slope of the simple linear regression of $E_i^{(j)}$ on X_j , but nonlinearity may be apparent in the plot as well.

4.4.2 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 13.
- 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)].

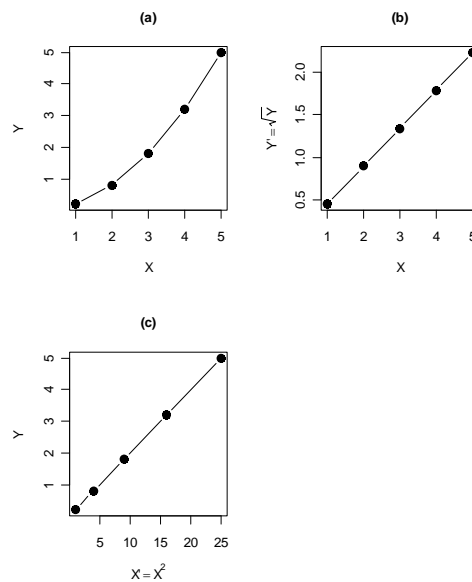


Figure 13. Transforming a nonlinear relationship (a) to linearity, (b) or (c).

- ▶ A power transformation works here because the relationship between Y and X is both monotone and simple. In Figure 14:
 - 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, $\hat{Y} = a + b_1X + b_2X^2$.
- ▶ Figure 15 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).

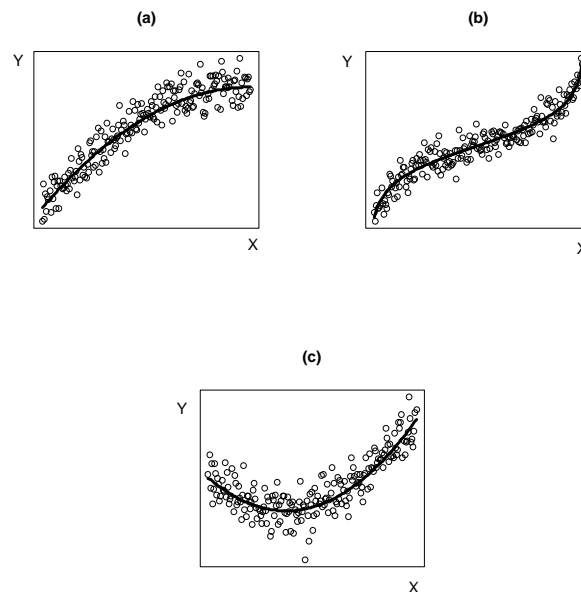


Figure 14. (a) A simple monotone relationship. (b) A monotone relationship that is not simple. (c) A simple nonmonotone relationship.

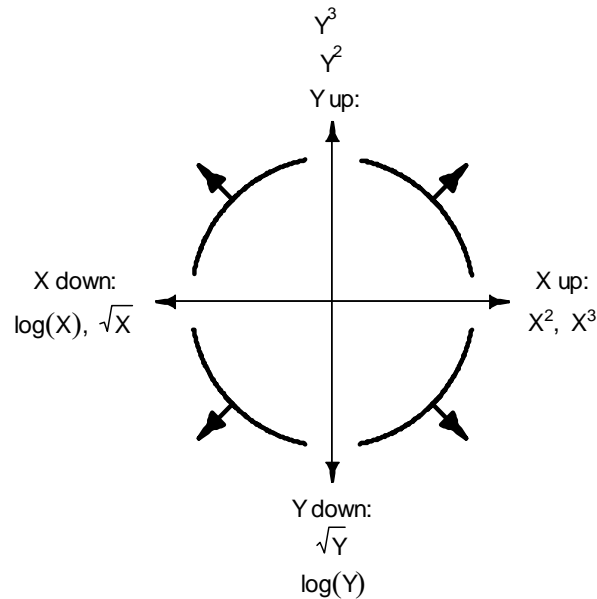


Figure 15. Mosteller and Tukey's bulging rule for selecting linearizing transformations.

© 2010 by John Fox

York SPIDA

5. 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.

© 2010 by John Fox

York SPIDA

- `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

$$\text{lhs} \sim \text{rhs}$$

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)`]:

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

- 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., $\log(\text{income} + 1)$ or $I(\text{income}^2)$, where $I()$ is the *identity function*].
- We say that “lhs is modeled as rhs” or that “lhs is regressed on rhs.”