Appendix D

Methods for Multiple Linear Regression Analysis

D.1. Introduction and Notation.

The EPOLLS model equations were derived from a multiple linear regression (MLR) analysis as described in Chapters 9 through 11. This analysis was performed using the SAS® System (version 6.08) statistical software from the SAS Institute Inc. MLR methods and associated statistical tests used to develop the EPOLLS model are documented in this appendix. This brief discussion of statistical regression methods is based on more thorough treatments given by Montgomery and Peck (1992) and Freund and Littell (1991). Notation and basic definitions associated with regression analyses include the following:

- i = index denoting different observations or case studies in a data set
- j = index denoting different regressor variables in a model
- n = number of observations in a data sample
- k = number of regressor variables in a model
- $p = k + 1$ = number of coefficients in a regression model
- u = a random variable
- μ_u = mean of the population of u
- $\overline{u} = \Sigma u_i / n$ = mean of a sample of u, giving an estimate of population mean ($\hat{\mu}_u$)
- $\sigma_u^2 = \Sigma (u_i - \mu_u)^2 / n$ = variance of the population of u, indicating variability about the mean
- $s_u^2 = \Sigma (u_i - \overline{u})^2 / (n - 1)$ = variance of a sample of u, giving an estimate of population variance ($\hat{\sigma}_u$)
- y_i = system response (dependent), which is random and observable, from i^{th} observation in sample
- x_{ij} = j^{th} regressor (independent) variable, which is constant and known, from i^{th} observation in sample
- ε_i = error (random) between a regression model and the i^{th} observation in sample
- β_j = coefficients (parameters) of a regression model
- $\hat{\theta}$ = an estimator of the true value of the parameter θ
- $\hat{\theta}_{(i)}$ = a parameter calculated with a model fit to all available data *except* the i^{th} observation
- $\hat{\theta}_{(j)}$ = a parameter calculated with a model using all regressor variables *except* the j^{th} regressor

D.2. Simple and Multiple Linear Regression.

The familiar linear regression analysis involves two coefficients, which define the slope and intercept of a line. A simple linear regression (SLR) model can be written as:

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad (D.1)$$

In fitting a line to a sample data set using SLR, the goal is to find estimates of the coefficients ($\hat{\beta}_0$ and $\hat{\beta}_1$) that minimize the error ε. Mathematically, this is done by minimizing the square of the errors between the observed and predicted values of y (hence the name *least squares regression*). With the fitted model, the predicted value of y for a given x is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \qquad (D.2)$$

Multiple linear regression (MLR) is used to fit a model with more than one regressor variable. Analogous to Equation D.1 for SLR, an MLR model can be written using matrix notation:

$$\{Y\}_{nx1} = [X]_{nxp} \{\beta\}_{px1} + \{\varepsilon\}_{nx1} \qquad (D.3)$$

where the matrix of the regressor variables [X] is called the *design matrix*. The first model coefficient (β_0) in the vector $\{\beta\}$ is the intercept of the model. Hence, the design matrix must be specified as:

$$[X]_{nxp} = \begin{bmatrix} 1.0 & x_{11} & x_{12} & \dots & x_{1k} \\ 1.0 & x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots & \dots \\ 1.0 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \qquad (D.4)$$

In fitting an MLR model, the goal is to find the "best" estimates of the coefficients $\{\beta\}$ that minimize the differences between all of the observed responses (y_i) and the corresponding model predictions ($\hat{y}_i$). In the same manner as for SLR, the coefficients are found by minimizing the sum of the squares of the errors (that is, minimize $\Sigma\varepsilon_i^2 = \Sigma(y_i - \hat{y}_i)^2$) in a least squares regression analysis. The solution for least squares estimators of the model coefficients can be written:

$$\{\hat{\beta}\} = \left([X]^T [X] \right)^{-1} [X]^T \{Y\} \qquad (D.5)$$

A related calculation defines the "hat" matrix [H]:

$$[H]_{nxn} = [X] \left([X]^T[X] \right)^{-1} [X]^T \qquad (D.6)$$

The [H] matrix is used in several model diagnostics described later in this appendix.

A multiple linear regression model is called "linear" because only linear coefficients {β} are used. Transforms of the regressor variables (such as 1/x, x^2, $x^{0.5}$, e^x, ln(x), etc.) are all permitted in an MLR model. A basic assumption in a regression analysis is that the functional form of the model is appropriate and includes all important variables. In addition, four assumptions about the errors {ε} are fundamental to a regression analysis:

(1) the mean of the errors is zero,
(2) the variances of the errors for all observations are constant,
(3) the errors are independent of each other (uncorrelated), and
(4) for some statistical tests, the errors are normally distributed.

Gross violations of these basic assumptions will yield a poor or biased model. However, if the variances of the errors {ε} are unequal and can be estimated, weighted regression schemes can sometimes be used to obtain a better model.

D.3. Category Variables.

The regressors in an MLR model are usually quantitative variables that can take on any real value. Frequently, however, the sample data can be sorted into categories. For example, data on the performance of a group of students can be sorted by gender. To consider the effect of this division in an MLR model, a *category variable* can be introduced into the design matrix [X]. The category variable x_1 might be defined as x_1=0 for females and x_1=1 for males. Instead of developing separate models for each group, a category variable allows for the fitting of a single MLR model that is simpler to use and applicable to both groups. Moreover, because all of the model coefficients are determined using the entire data set, a single model with a category variable is much better overall than separate models fit to subsets of the data.

However, adding just the category variable to the design matrix will mean, in essence, that only the zero-intercept of the model will change with the category. That is, all other regressors will have the same effect on the system response regardless of the category to which the observation belongs. To consider the effect of a category on other regressors, product or interaction terms should be added to the model. For example, to add the category variable x_1 to a model with the quantitative regressor x_j, the regression model should be:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_j + \beta_3 x_1 x_j + \varepsilon \qquad (D.7)$$

More than one category variable can be included in the design matrix. To represent categories with more than three levels (like eye color of brown, blue, or green), two category

variables are needed. For example: x_1=0, x_2=0 for brown eyes, x_1=0, x_2=1 for blue eyes, and x_1=1, x_2=1 for green eyes. A four level category would require three category variables, and so forth. However, to get a valid model, the design matrix must include all possible combinations of these category variables.

D.4. Model Quality and Significance of Regressors.

The squares of the errors between the observed and predicted observations are used to evaluate how well a regression model fits the sample data. A basic identity in regression analyses is obtained by partitioning the sum of the squares of the variation in the sample data:

$$\sum_i^n (y_i - \bar{y})^2 = \sum_i^n (y_i - \hat{y}_i)^2 + \sum_i^n (\hat{y}_i - \bar{y})^2$$

$$\mathbf{SST} \quad = \quad \mathbf{SSE} \quad + \quad \mathbf{SSR} \qquad \textit{(D.8)}$$

In words, Equation D.8 means that the total variability in the observed responses [$SST=\Sigma(y_i-\bar{y})^2$] is equal to the random variability *not* explained by the model or model error [$SSE=\Sigma(y_i-\hat{y}_i)^2$] plus the systematic variability that is explained by the regression model [$SSR=\Sigma(\hat{y}_i-\bar{y})^2$]. When adjusted for the associated degrees-of-freedom, the three terms in Equation D.8 lead to the following definitions:

$MST = SST / (n-1) = \Sigma(y_i - \bar{y})^2 / (n-1)$ = total mean square of variation in observations

$MSE = SSE / (n-p) = \Sigma(y_i - \hat{y}_i)^2 / (n-p)$ = error mean square

$MSR = SSR / 1 = \Sigma(\hat{y}_i - \bar{y})^2 / 1$ = regression mean square

Significantly, the MSE value gives an unbiased estimate of the variance in the errors ε_i.

To express the quality of fit between a regression model and the sample data, the *coefficient of multiple determination* (R^2) is typically used. Ranging in value from 0.0 to 1.0, the coefficient of multiple determination is defined as:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \qquad \textit{(D.9)}$$

Higher values of R^2 (from a smaller SSE) indicate a better fit of the model to the sample observations. However, adding any regressor variable to an MLR model, even an irrelevant regressor, yields a smaller SSE and greater R^2. For this reason, R^2 by itself is not a good measure of the quality of fit.

To overcome this deficiency in R^2, an adjusted value can be used. The *adjusted coefficient of multiple determination* ($\bar{R}^2$) is defined as:

$$\bar{R}^2 = 1 - \frac{MSE}{MST} = 1 - \left(\frac{n-1}{n-p}\right)(1-R^2) \qquad (D.10)$$

Because the number of model coefficients (p) is used in computing $\bar{R}^2$, the value will not necessarily increase with the addition of any regressor. Hence, $\bar{R}^2$ is a more reliable indicator of model quality.

The *global F-test* is used to assess the overall ability of a model to explain at least some of the observed variability in the sample responses. Giving a statistical test for the significance of the regression, the global *F*-test is performed in the following steps:

(1) Null hypothesis: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$.
(2) Compute the test statistic F_0 with an analysis-of-variance table:

Source	*Degrees of Freedom*	*Sum of Squares*	*Mean Square*	F_o
Regression	k	SSR	MSR	$F_{0, k, n-p}$ = MSR / MSE
Error	n - p	SSE	MSE	
Total	n - 1	SST		

(4) From the *F* distribution, find $F_{\alpha, k, n-p}$ corresponding to the desired level of significance (α).
(5) If $F_{0, k, n-p} > F_{\alpha, k, n-p}$, reject the null hypothesis and conclude that at least one $\beta_j \neq 0$ and at least one model regressor explains some of the response variation.

A global *F*-test only indicates that at least one regressor is significant, but does not indicate which regressors are significant and which are not.

To study the significance of each regressor variable, *partial F-tests* are used. These tests look at the significance of a given regressor in the presence of all other regressor variables in the model. Hence, partial *F*-tests can be performed for each of the "k" regressors:

(1) Null hypothesis: $\beta_j = 0$.
(2) Compute the test statistic F_j with:
$F_{j, 1, n-p}$ = [SSR(full model) - SSR(full model without x_j)] / MSE = (SSR - $SSR_{(j)}$)/MSE
(4) From the *F* distribution, find $F_{\alpha, 1, n-p}$ corresponding to the desired level of significance (α).
(5) If $F_{j, 1, n-p} > F_{\alpha, 1, n-p}$, reject the null hypothesis and conclude that $\beta_j \neq 0$ and that the variable x_j is a significant regressor in the full model.

If a regression model is poorly specified, either because important regressor variables are

missing or unnecessary variables have been included, the fitted model may be biased. An indication of the model bias is given by the *Mallows statistic* (C_p), which attempts to measure the overall bias or mean square error in the estimated model parameters. The Mallows statistic is defined with:

$$C_p = \frac{SSE(p)}{MSE} - n + 2p \qquad (D.11)$$

Here, MSE is computed using all available regressors and SSE(p) is computed from a model with only "p" coefficients. Note that when p includes all available regressors, SSE(p) = SSE and $C_p = p$. A low C_p value indicates good model predictions but, more importantly, a p-term model with little bias will yield $C_p \approx p$. The C_p statistic is most useful in evaluating candidate regression models as discussed in the next section.

D.5. Selection of Regressors for Candidate Models.

The real challenge in performing a multiple linear regression analysis is to find the "best" set of regressor variables that explain the variation in the observed system responses. A model is desired that not only fits well to the observations, but also yields good predictions of future responses and only includes regressor variables that contribute significantly to the model. This process of finding the best set of regressors for an MLR model is known as *variable selection* or *model building*.

Given M regressor variables that could be included in a simple MLR model, there are 2^M possible equations that could be written with different combinations of the regressors. With a typically large pool of potential regressor variables, examination of all possible models is simply not practical. Instead, systematic methods are employed to find a subset of the regressor variables that will form an appropriate, "best" model. Actually, different variable selection procedures usually suggest different "best" models so additional analyses and judgment are required to obtain a good, reliable MLR model. Final selection of the best model can be based on the evaluations discussed in Section D.8.

Perhaps the simplest variable selection procedure involves an attempt to find a model with maximum R^2 or $\bar{R}^2$. This procedure does require fitting all possible models, but the results are ranked to easily identify the best model. All possible models with "k" regressors are evaluated and the one model giving the greatest R^2 or $\bar{R}^2$ is tabulated. The maximum R^2 or $\bar{R}^2$ found from models with one, two, three, etc. regressors are then plotted as shown in Figure D.1. Recall that R^2 always increases with the addition of more regressor variables, while $\bar{R}^2$ may eventually decrease as shown in Figure D.1. A good candidate model can be selected where $\bar{R}^2$ reaches a maximum or where the R^2 curve begins to flatten out.

A *stepwise selection* procedure relies on partial *F*-tests to find a group of significant regressor variables. The "best" model is found by adding or eliminating regressors in steps. First, partial *F*-statistics are computed for every potential regressor and the one variable giving the highest F_j is inserted into the model. Next, partial *F*-statistics are computed for all of the remaining regressors and the one yielding the highest F_j, in the presence of the first-selected regressor, is added to the model. However, no regressor is added to the model at this step unless F_j exceeds a specified threshold value. Next, all variables in the model are evaluated with partial *F*-tests to see if each one is still significant. In this step, any regressor that is no longer significant, according to the specified threshold value of F_j, is dropped from the model. The selection procedure continues in steps as new regressors are added to the model and then any variables that are no longer significant are dropped. The stepwise selection procedure stops when no other potential regressor yields a partial F greater than the threshold and all regressors in the model remain significant. The threshold or cut-off partial F values, for addition to or elimination from the model, are specified in terms of a level of significance (α). One disadvantage of the stepwise selection procedure is that not all possible combinations of regressor variables are considered for the model. Also, since one final equation is produced in a stepwise procedure, other equally good models may go unrecognized.

The Mallows statistic, defined in the previous section, can also be used to find a good set of regressors for an MLR model. As done in the selection procedure based on maximum R^2 or $\overline{R}^2$, all possible models with "k" regressors are evaluated. Here, from all possible models with one, two, three, etc. regressors, the one model giving the lowest C_p is tabulated. The results are then plotted as shown in Figure D.2. Recall that $C_p = p$ when all available regressor variables are used in the model; therefore, the trend line in Figure D.2 converges to the $C_p = p$ line. A good candidate model can be selected from the C_p plot by remembering that good model predictions are indicated by a low C_p value and low model bias is indicated by $C_p \approx p$.

D.6. Tests for Multicollinearity.

The problem of *multicollinearity* exists when two or more model regressors are strongly correlated or *linearly* dependent. In building a regression model, the regressor variables in the design matrix are assumed to be independent of one another. However, when the system behavior is poorly understood, the selected MLR model might include several regressors that each measure, to some degree, similar components of the system response. In this situation, columns of the design matrix may be linearly correlated to a sufficient degree as to create a multicollinearity problem. When significant multicollinearity exists in the data, the mathematical solution used to fit the regression model is unstable. That is, small changes in the regressor values can cause large changes in the parameter estimates and yield an unrealistic model. Multicollinearity is especially problematic if the fitted MLR model is then used to make future predictions. Note that the regressor variables do not have to be totally independent of one

another, and some degree of correlation within the design matrix is tolerable. Techniques for ascertaining potential problems with multicollinearity are outlined in this section. When serious multicollinearity is detected, the problem can often be eliminated by discarding one or more regressor variables from the model.

Possible problems due to multicollinearity may be detected during fitting of a regression model. Common indications of multicollinearity include:

(1) Parameter estimates ($\hat{\beta}_j$) with signs that defy prior knowledge (i.e., a model coefficient with a negative sign when a positive sign is expected).
(2) Models with large R^2, or high significance in a global *F*-test, but in which none of the model variables are significant in partial *F*-tests.
(3) Different model selection procedures yield very different models.
(4) Standard errors of the regression coefficients that are large, with respect to the parameter estimates, indicating poor precision in the estimates. The *standard error* of a coefficient is calculated as $se(\hat{\beta}_j) = (MSE \cdot C_{jj})^{0.5}$ where C_{jj} is the j^{th} diagonal element of $([X]^T[X])^{-1}$.

A simple form of multicollinearity is caused by pair-wise correlation between any two regressor variables. This can be detected by inspection of the correlation matrix [r] for the regressor values in the design matrix. The *empirical correlation* between regressors x_j and x_m, giving one element of [r], is computed with:

$$r_{x_j x_m} = r_{jm} = \frac{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)\ (x_{im} - \bar{x}_m)}{\sqrt{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 \cdot \sum_{i=1}^{n} (x_{im} - \bar{x}_m)^2}} \qquad (D.12)$$

The greater the linear dependence between x_j and x_m, the closer $|r_{jm}|$ will be to one (obviously, the diagonals of the correlation matrix (r_{jj}) are equal to one). As a general rule, multicollinearity may be a problem if, for the off-diagonal terms of the correlation matrix, $|r_{jm}| \geq 0.9$. However, the pair-wise correlation coefficients will not indicate multicollinearity problems arising from linear dependencies between combinations of more than two regressors.

Multicollinearity can also be detected from the eigenvalues or characteristic roots of the correlation matrix [r]. For a model with k regressors, there will be k eigenvalues, λ_j. The ratio of the maximum over the minimum eigenvalues of [r] defines the model *condition number*:

$$\kappa = \frac{\lambda_{max}}{\lambda_{min}} \qquad (D.13)$$

As a general rule, $\kappa < 100$ indicates no serious problem, $100 < \kappa < 1000$ indicates moderate to strong multicollinearity, and $\kappa > 1000$ indicates a severe problem with multicollinearity. In addition, the *condition index* associated with each regressor variable x_j is defined as:

$$\kappa_j = \frac{\lambda_{max}}{\lambda_j} \qquad (D.14)$$

The number of $\kappa_j > 1000$ indicates the number of linear dependencies in the design matrix.

Another indication of potential multicollinearity is obtained from *variance inflation factors* (VIFs). The VIF associated with regressor x_j is computed with:

$$VIF_j = \left(1 - R^2_{(j)}\right)^{-1} \qquad (D.15)$$

where $R_{(j)}^2$ is the coefficient of multiple determination (R^2) from a regression of x_j on all other k-1 regressors in the model. Hence, as more of the variation in x_j can be explained by the other regressor variables, $R_{(j)}^2$ will approach one and the VIF_j will increase. Large values of VIF_j indicate possible multicollinearity associated with regressor x_j. In general, a $VIF_j \geq 5$ indicates a possible multicollinearity problem, while a $VIF_j \geq 10$ indicates that multicollinearity is almost certainly a problem.

D.7. Tests for Influential Observations.

In addition to multicollinearity, another common problem in regression analyses are models adversely affected by influential observations. Three types of influential observations are illustrated in Figure D.3 for a simple linear regression model. *Outliers*, defined as observations outside the general trend of the data, are a familiar type of influential observation. When an observation does fall within the trend of the data, but is found beyond the range of the other regressors, the resulting influential observation is called a *high leverage point.* When a high leverage point is also an outlier, that single data point can have a large impact on the regression model and is called a *highly influential observation.* In this section, methods are presented for detecting influential observations.

The presence of influential observations can be detected by computing the PRESS (prediction error sum of squares) statistic defined as:

$$PRESS = \sum_{i=1}^{n} (y_i - \hat{y}_{(i)})^2 \qquad (D.16)$$

where $\hat{y}_{(i)}$ is a prediction of the i^{th} observed response made from a model regressed on all of the available data *except* the i^{th} observation. The PRESS statistic is then compared with the sum of the square of the errors, $SSE = \Sigma(y_i - \hat{y}_i)^2$. If PRESS is much larger than SSE, influential observations may exist.

Outliers can be detected from *studentized residuals* that are defined for the i^{th} observation with:

$$r_i = \frac{y_i - \hat{y}_i}{\sqrt{MSE\ (1 - h_{ii})}} \qquad (D.17)$$

where h_{ii} is the i^{th} diagonal element of the hat matrix [H] defined in Equation D.6. When working with a sufficient number of observations, such that (n-p-1) > 20, a $|r_i| > 2.0$ indicates that the i^{th} observation might be an outlier. Similarly, a $|r_i| > 2.5$ is a strong indicator of a likely outlier.

In addition, high leverage points can be detected directly from the diagonals of the hat matrix, h_{ii}. As a general rule, an $h_{ii} > 2p/n$ indicates that the i^{th} observation is a possible high leverage point.

One diagnostic test for highly influential observations uses the *DFFITS* statistic. For the i^{th} observation, this statistic is defined as:

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{MSE_{(i)}\ h_{ii}}} \qquad (D.18)$$

where $\hat{y}_{(i)}$ and $MSE_{(i)}$ are based on a model regressed on all of the available data *except* the i^{th} observation. A possible highly influential observation is indicated by $|DFFITS_i| > 2(p/n)^{0.5}$.

Another test for highly influential observations is based on the effects of the i^{th} observation on the estimated model coefficients. The *DFBETAS* statistic, which is calculated for each j^{th} regressor variable and each i^{th} observation, is defined as:

$$DFBETAS_{ij} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{MSE_{(i)} \left([X]^T[X]\right)^{-1}_{jj}}} \qquad (D.19)$$

where $\hat{\beta}_{j(i)}$ and $MSE_{(i)}$ are computed from a regression on all available data *except* the i^{th} observation. As a general rule, a possible highly influential observation is indicated by $|DFBETAS_{ij}| > 2/n^{0.5}$. In practice, a highly influential observation will cause consistently high DFBETAS for most of the regressor variables.

A third test for highly influential observations is based on *Cook's Distance* defined as:

$$D_i = \frac{r_i^2}{p}\left(\frac{h_{ii}}{1 - h_{ii}}\right) \qquad (D.20)$$

where r_i is the studentized residual defined in Equation D.17. Values of D_i much larger than all others indicates that the i^{th} observation may be highly influential.

All influential observations should be investigated for correctness and accuracy. When using the statistical tests described here, the cutoff values for indicating an influential observation should be used only as a guideline. In practice, those observations giving the strongest indications in these tests should be investigated first. Moreover, a data point should not be discarded simply because it is an influential observation. A single influential data point can sometimes illuminate an important trend in the system response.

D.8. Evaluation of Final Model.

In addition to the diagnostics for multicollinearity and influence data, graphical methods are useful for evaluating the performance of a regression model. In a simple linear regression model, the adequacy of a linear equation can be easily visualized with a scatter plot of the observed data (x_i, y_i). However, with a multiple linear regression model, plots of the observed responses (y_i) versus each regressor variable (x_{ij}) are of little value. Because the system response is a function of multiple regressors, plots of the observed response versus individual variables often fail to indicate a linear relationship and can be very misleading in evaluating an otherwise good MLR model. Other plots that are more useful in visualizing the performance of an MLR model are described in this section.

A scatter plot of the predicted response versus the observed response ($\hat{y}$ versus y), as shown in Figure D.4, gives a simple indication of model performance. Any model that can explain most of the variation in the observed responses will produce a plot with points clustered around a 45° line. Better models yield less scatter about this $\hat{y}$=y line. Moreover, the scatter of points about the $\hat{y}$=y line should remain roughly constant with magnitude. That is, a poor model that is less accurate at larger values of $\hat{y}$ will produce increasing scatter with larger values of y.

A scatter plot of the residuals, as shown in Figure D.5, is also useful in evaluating a regression model. Here, the model residuals or errors ($e = \hat{y} - y$) are plotted against the model predictions ($\hat{y}$). Residual plots are used to visually verify some of the basic assumptions underlying an MLR analysis, as discussed previously in Section D.2. Namely, the residuals (errors) between the model predictions and observed responses should have a mean of zero and a constant variance. Hence, the scatter in the residuals should be fairly uniform and centered about e = 0. A good regression model will produce a scatter in the residuals that is roughly constant with $\hat{y}$ as shown in Figure D.5a. Unsatisfactory models yield a scatter in the residuals that changes with $\hat{y}$; three common examples are shown in (b), (c), and (d) of Figure D.5. Models producing an unsatisfactory scatter in the residuals can often be improved by transforming y to stabilize the variance in the residuals. For example, a model might be re-defined in terms of ln(y), $y^{0.5}$, $y^{-0.5}$, or 1/y. However, such transformations necessitate bias-reducing adjustments when the model predictions are de-transformed, as discussed further in the next section.

A third graphical method for evaluating a multiple linear regression model is based on the idea of *partial residuals*. Designated with $e^{y}_{i(j)}$, the partial residual of y for x_j is defined as:

$$e^{y}_{i(j)} = y_i - \hat{y}_{i(j)} \qquad (D.21)$$

where $\hat{y}_{i(j)}$ is a prediction of y_i from a regression model using all of the regressors *except* x_j. Similarly, the partial residual of x_j is designated with $e^{x}_{i(j)}$ and defined as:

$$e^{x}_{i(j)} = x_{ij} - \hat{x}_{i(j)} \qquad (D.22)$$

where $\hat{x}_{i(j)}$ is a prediction of the regressor x_{ij} from a regression of x_j on all the other regressor variables. Hence, the partial residual $e^{y}_{i(j)}$ represents the variation in y_i not explained by a model that excludes the regressor x_j, and the partial residual $e^{x}_{i(j)}$ represents the variation in x_j that can not be explained by the other regressor variables. Plotting $e^{y}_{i(j)}$ against $e^{x}_{i(j)}$ in a *partial regression plot* more clearly shows the influence of x_j on y in the presence of all other regressors in the model.

Partial regression plots, shown in Figure D.6, can be generated for every regressor variable

in a model. If the regressor x_j is linearly related to the model response, a plot of the partial residuals $e^y_{i(j)}$ and $e^x_{i(j)}$ will cluster about a line that passes through the origin at a slope equal to β_j (the coefficient corresponding to x_j in the full model). Moreover, a stronger linear relationship between y and x_j will be evidenced by a narrower clustering of the partial residuals. For example, less scatter is seen in the partial regression plot in Figure D.6a than the plot in Figure D.6b; this indicates a stronger relationship with the regressor x_1 than for the regressor x_2. Influence data and their effect on the fit of the model can also be spotted on a partial regression plot as shown in Figure D.6c. Even more useful, the need for linear transformations of a given regressor variable (such as ln(x), 1/x, $x^{0.5}$, x^2, etc.) may be suggested by a partial regression plot like that in Figure D.6d.

Finally, it is often important to evaluate the ability of a fitted regression model to predict future events. The best way to do this is to gather additional, new data and compare these observed responses with predictions from the model. However, when this is not possible, the data used to fit the model can be split and a cross validation performed. A good regression model can be fit to part of the original data set and still accurately predict the other observations. To perform a double cross-validation:

(1) Partition the data into two subsets (say, A and B) with an equal number of observations in each. Assigning individual observations to subset A or B must be done randomly.
(2) Using the same model form, fit the model using the data from subset A. Use this model to predict the observations in subset B.
(3) Compute the prediction $R^2_{p,A}$ for the model fit to subset A, as defined in Equation D.23 below.
(4) Similarly, fit the model to subset B and use this to predict the observations in subset A.
(5) Compute the prediction $R^2_{p,B}$ for the model fit to subset B.

A good model will produce high values of R^2_p for both subsets and these values will be approximately equal ($R^2_{p,A} \approx R^2_{p,B}$).

The *prediction* R^2_p for a model fit to subset A ($R^2_{p,A}$) is computed with:

$$R^2_{p,A} = 1 - \frac{\sum_{i=1}^{n_B} (y_{iB} - \hat{y}_{iA})^2}{\sum_{i=1}^{n_B} (y_{iB} - \bar{y}_B)^2} \qquad (D.23)$$

where n_B and $\bar{y}_B$ are the number and mean of the observed responses (y_{iB}) in the random subset B. Using the model fit to subset A, predictions of the observations in subset B are made to give $\hat{y}_{iA}$. The prediction R^2_p for a model fit to subset B ($R^2_{p,B}$) is computed the same way, with the use of subsets A and B reversed.

D.9. Predictions from a Regression Model.

Often, the purpose of developing a regression model is to allow predictions of future events. In equation form, the predicted system response ($\hat{y}$), for a given set of regressor values (x), is computed with:

$$\hat{y} = [x]\{\hat{\beta}\} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k \qquad (D.24)$$

where $\hat{\beta}$ denotes the fitted regression coefficients.

As mentioned in the previous section, the observed system responses are sometimes transformed to stabilize the variance of the model errors, based on an examination of a residuals plot. For example, suppose a model for predicting some response "θ" is desired, and the regression analysis indicates the need to transform θ by taking the square root and fitting the model to $\theta^{0.5}$. Predictions of $\hat{\theta}$ would then be obtained by squaring the model prediction. Unfortunately, the de-transformed prediction is substantially biased and will consistently under-predict the value of θ. To alleviate this problem, Miller (1984) suggests bias-reducing adjustment factors for logarithmic, square root, and inverse transformations of the dependent variable. For a square-root transformation, where a regression model is fit to $y_i^{0.5}$, the low-bias prediction of $\hat{y}$ given by Miller can be written as:

$$\hat{y} = \left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k\right)^2 + MSE \qquad (D.25)$$

where MSE is used as an estimate of the variance of the errors in the fitted model.

Because a regression model is simply an equation fit to a database of observed responses, a regression model should not be trusted to make predictions outside the range of the regressor variables used in fitting the model. Hence, the first step in using a regression model should be to verify that the prediction does not require extrapolation beyond the range of the regressor variables in the original data set. This can be done most simply by referring to histograms of the original regressor variables. However, with more than two or three model variables, it is possible to be within the range of each regressor yet still extrapolate beyond the combined range of the variables. Referred to as *hidden extrapolation,* this problem is illustrated in Figure D.7 for a model with two regressor variables. Hidden extrapolation is especially problematic if the regression model is unstable due to problems of multicollinearity. An indication of possible hidden extrapolation can be made by computing h_0:

$$h_0 = [x] \left([X]^T [X] \right)^{-1} [x]^T \qquad (D.26)$$

where [X] is the design matrix used to fit the model and [x] = [1 x_1 x_2 ... x_k] is the row matrix

of values at which the model prediction is to be made. The computed value of h_0 is then compared with h_{max}, which is the maximum of the diagonal elements of the hat matrix [H] from Equation D.6. As a general rule, extrapolation is indicated when $h_0 > h_{max}$.

Finally, based on how well the model fits the available data, it is possible to construct a *prediction interval*. For a given set of the regressor values, the actual response is believed to fall within the prediction interval (1-α)% of the time. The upper and lower bounds of the prediction interval are defined by:

$$\textit{Prediction Interval for } \hat{y} = \hat{y}' \pm (t_{\alpha/2,n-p})\sqrt{MSE\left(1+[x]([X]^T[X])^{-1}[x]^T\right)} \quad (D.27)$$

where $t_{\alpha/2,\ n-p}$ is the tail area of the t-distribution at $\alpha/2$ for n-p degrees of freedom. The value $\hat{y}'$ is the predicted response at [x], but without the bias-reducing adjustment used with de-transformations of y (see Equation D.25). For a model specified with a square-root transformation, $\hat{y}' = ([x]\{\hat{\beta}\})^2$. Note that the prediction interval given in Equation D.27 applies to the model prediction of a single system response. The *confidence interval*, for the predicted mean of multiple responses for a given set of regressor values, is narrower.

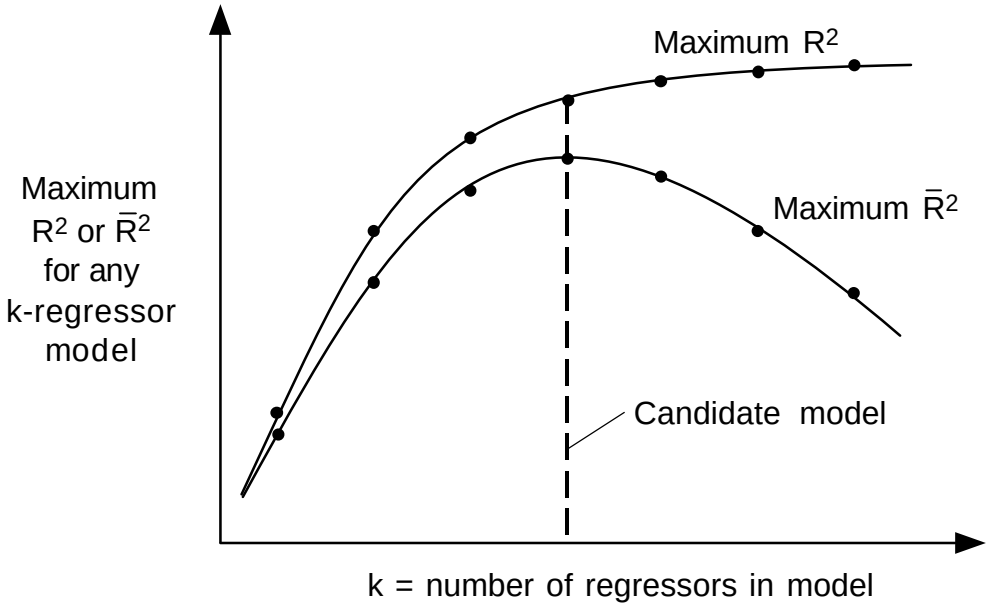

Figure D.1. Selection of a candidate model based on maximum R^2 or $\bar{R}^2$.

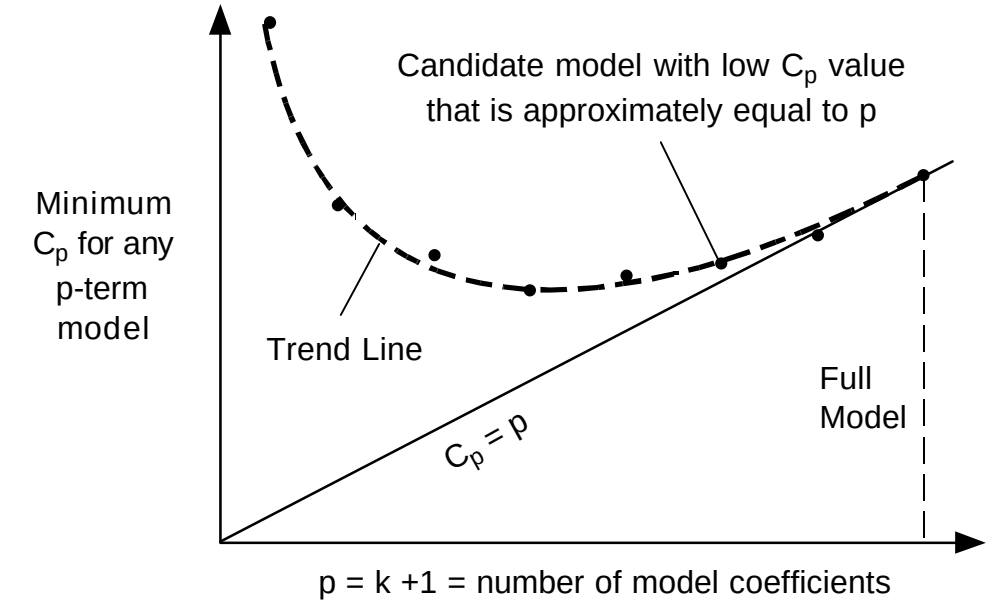

Figure D.2. Selection of a candidate model from a C_p plot.

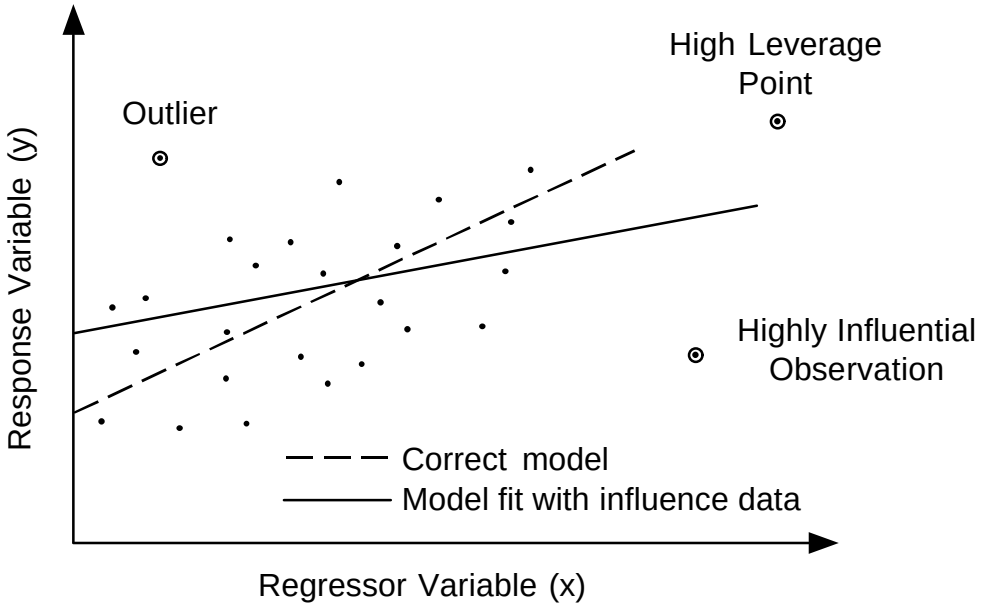

Figure D.3. Definition of influential observations in a simple linear regression model.

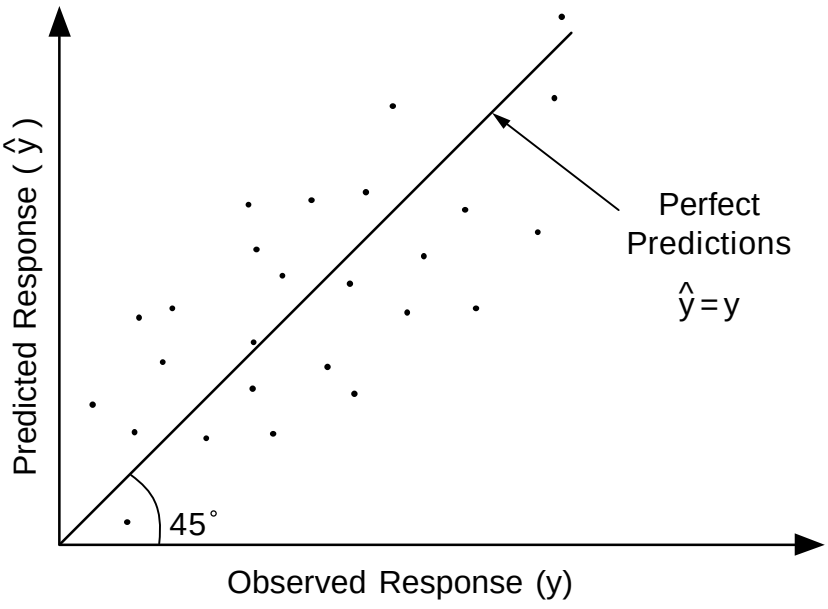

Figure D.4. Evaluation of final model with a $\hat{y}$-y scatter plot.

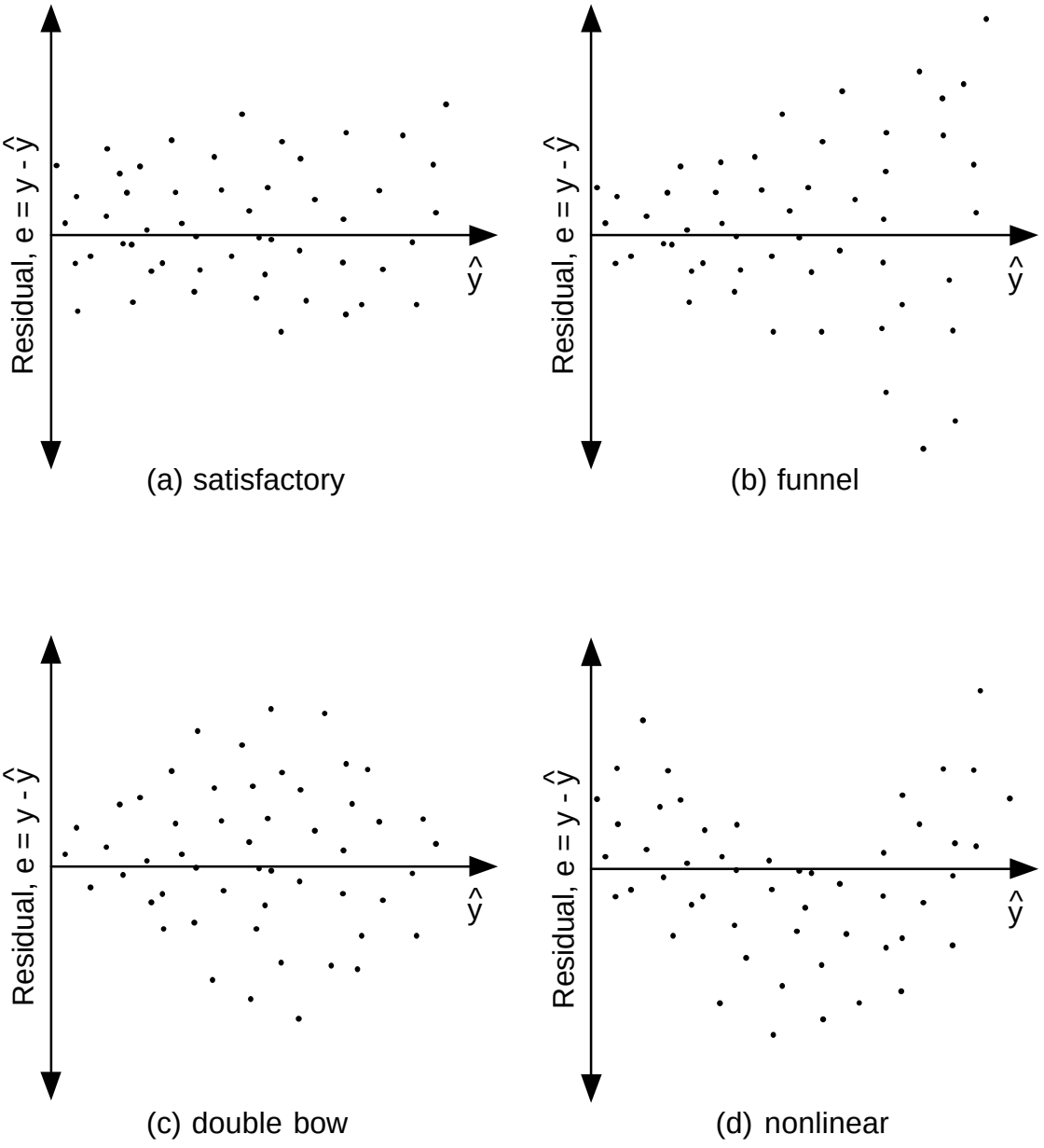

Figure D.5. Common patterns in residual plots used to evaluate MLR models (after Montgomery and Peck 1992).

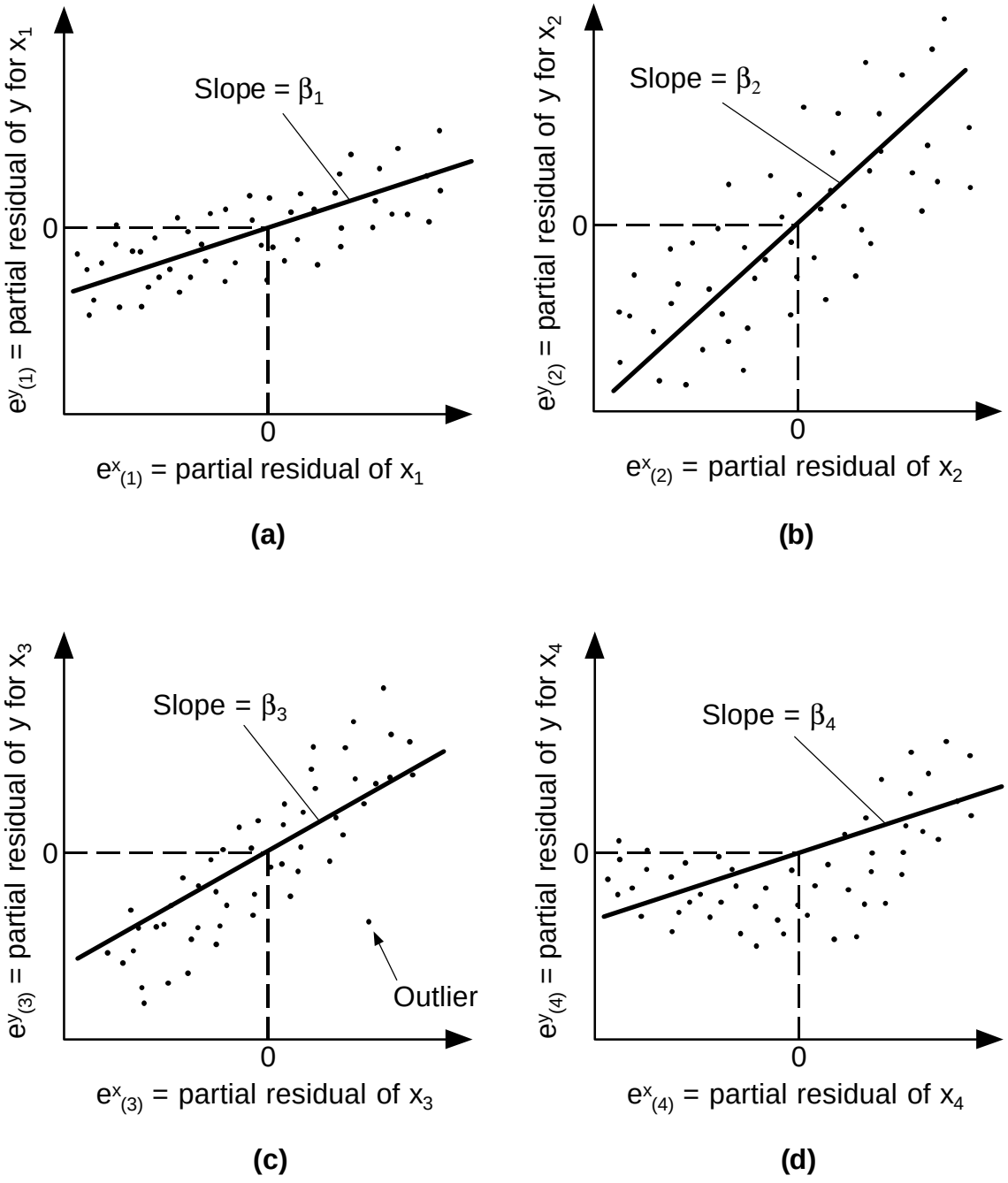

Figure D.6. Partial regression plots used to evaluate MLR models.

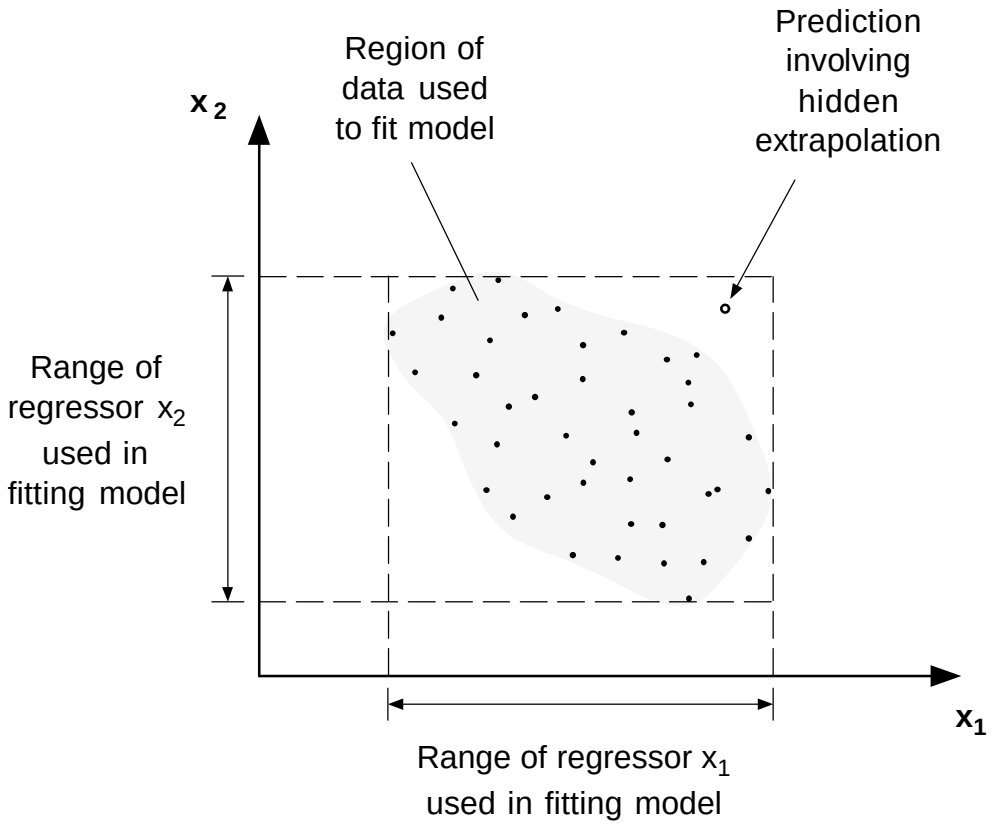

Figure D.7. Illustration of hidden extrapolation in a model with two regressor variables (after Montgomery and Peck 1992).