Chapter 7

Statistical Forecasting

7.1. BACKGROUND

Much of operational weather and long-range (seasonal, or “climate”) forecasting has a statistical basis. As a nonlinear dynamical system, the atmosphere is not perfectly predictable in a deterministic sense. Consequently, statistical methods are useful, and indeed necessary, parts of the forecasting enterprise. This chapter provides an introduction to statistical forecasting of scalar (single-number) quantities. Some methods suited to statistical prediction of vector (multiple values simultaneously) quantities, for example, spatial patterns, are presented in Sections 14.2.3, 14.3.2, and 15.4. The forecasting emphasis here is consistent with the orientation of this book, but the methods presented in this chapter are applicable to other settings as well.

Some statistical forecast methods operate without information from the fluid-dynamical forecast models that have become the mainstay of weather forecasting for lead times ranging from one day to a week or so in advance. Such pure statistical forecast methods are sometimes referred to as Classical, reflecting their prominence in the years before dynamical forecast information was available. These methods continue to be viable and useful at very short lead times (hours in advance), or very long lead times (weeks or more in advance), for which the dynamical forecast information is not available with sufficient promptness or accuracy, respectively.

Another important application of statistical methods to weather forecasting is in conjunction with dynamical forecast information. Statistical forecast equations routinely are used to postprocess and enhance the results of dynamical forecasts at operational weather forecasting centers throughout the world and are essential as guidance products to aid weather forecasters. The combined statistical and dynamical approaches are especially important for providing forecasts for quantities and locations (e.g., particular cities rather than gridpoints) not represented by the dynamical models.

The types of statistical forecasts mentioned so far are objective, in the sense that a given set of inputs always produces the same particular output. However, another important aspect of statistical weather forecasting is in the subjective formulation of forecasts, particularly when the forecast quantity is a probability or set of probabilities. Here the Bayesian interpretation of probability as a quantified degree of belief is fundamental. Subjective probability assessment forms the basis of many operationally important forecasts and is a technique that could be used more broadly to enhance the information content of operational forecasts.

7.2. LINEAR REGRESSION

Much of statistical weather forecasting is based on the procedure known as linear, least-squares regression. In this section, the fundamentals of linear regression are reviewed. Much more complete treatments can be found in standard texts such as Draper and Smith (1998) and Neter et al. (1996).

Statistical Methods in the Atmospheric Sciences. https://doi.org/10.1016/B978-0-12-815823-4.00007-9

7.2.1. Simple Linear Regression

Regression is most easily understood in the case of *simple linear regression*, which describes the linear relationship between two variables, say x and y. Conventionally the symbol x is used for the *independent*, or *predictor variable*, and the symbol y is used for the *dependent variable*, or *predictand*. The terms dependent and independent variable are in common in the literature of statistics and other disciplines, whereas the terms predictor and predictand are used primarily in the atmospheric and related sciences, having apparently been introduced by Gringorten (1949). More than one predictor ("x") variable is very often required in practical forecast problems, but the ideas for simple linear regression generalize easily to this more complex case of *multiple linear regression*. Therefore most of the important ideas about regression can be presented in the context of simple linear regression.

Essentially, simple linear regression seeks to summarize the relationship between x and y, shown graphically in their scatterplot, using a single straight line. The regression procedure chooses the line producing the least error for predictions of y given observations of x, within the (x, y) data set used to define that relationship. Exactly what is defined to be least error can depend on context, but the most usual error criterion is minimization of the sum (or, equivalently, the average) of the squared errors. It is the choice of the squared-error criterion that is the basis of the name *least-squares regression* or *ordinary least squares* (OLS) regression. Choosing the squared-error criterion is conventional not because it is necessarily best, but rather because it makes the mathematics analytically tractable. Adopting the squared-error criterion results in the line-fitting procedure being fairly tolerant of small discrepancies between the line and the points. However, the fitted line will adjust substantially to avoid very large discrepancies, and so the method is not resistant to outliers.

Figure 7.1 illustrates the situation. Given a data set of (x, y) pairs, the problem is to find the particular straight line,

$$\hat{y} = a + bx, \tag{7.1}$$

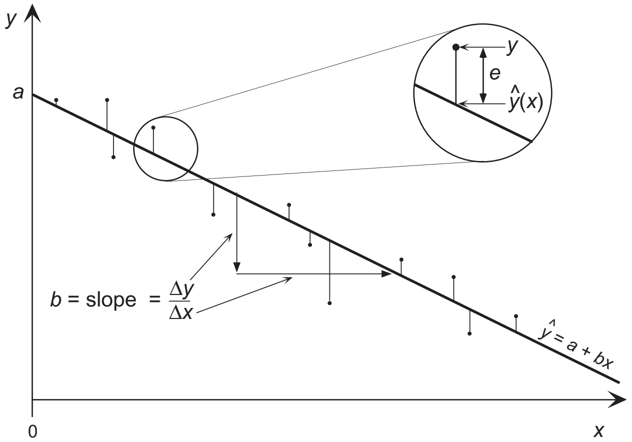


FIGURE 7.1 Schematic illustration of simple linear regression. The regression line, $\hat{y} = a + bx$, is chosen as the one minimizing some measure of the vertical differences (the residuals) between the points and the line. In least-squares regression that measure is the sum of the squared vertical distances. The inset shows a residual, e, as the difference between a data point and the regression line.

minimizing the squared vertical distances (thin lines) between it and the data points. The circumflex ("hat") accent signifies that the equation specifies a predicted value of y. The inset in Figure 7.1 indicates that the vertical distances between the data points and the line, also called errors or *residuals*, are defined as

$$e_i = y_i - \hat{y}(x_i). \tag{7.2}$$

There is a separate residual e_i for each data pair (x_i, y_i). Note that the sign convention implied by Equation 7.2 is for points above the line to be regarded as positive errors and points below the line to be negative errors. This is the usual convention in statistics, but is opposite to what often is seen in the atmospheric sciences, where forecasts smaller than the observations (the line being below the point) are regarded as having negative errors, and vice versa. However, the sign convention for the residuals is unimportant, since it is the minimization of the sum of squared residuals that defines the best-fitting line. Combining Equations 7.1 and 7.2 yields the regression equation,

$$y_i = \hat{y}_i + e_i = a + bx_i + e_i, \tag{7.3}$$

which says that the true value of the predictand is the sum of the predicted value (Equation 7.1) and the residual.

Finding analytic expressions for the least-squares intercept, a, and the slope, b, is a straightforward exercise in calculus. In order to minimize the sum of squared residuals,

$$\sum_{i=1}^{n} (e_i)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - [a + bx_i])^2, \tag{7.4}$$

it is only necessary to set the derivatives of Equation 7.4 with respect to the parameters a and b to zero and solve. These derivatives are

$$\frac{\partial \sum_{i=1}^{n} (e_i)^2}{\partial a} = \frac{\partial \sum_{i=1}^{n} (y_i - a - bx_i)^2}{\partial a} = -2 \sum_{i=1}^{n} (y_i - a - bx_i) = 0 \tag{7.5a}$$

and

$$\frac{\partial \sum_{i=1}^{n} (e_i)^2}{\partial b} = \frac{\partial \sum_{i=1}^{n} (y_i - a - bx_i)^2}{\partial b} = -2 \sum_{i=1}^{n} x_i[(y_i - a - bx_i)] = 0. \tag{7.5b}$$

Rearranging Equations 7.5 leads to the so-called *normal equations*,

$$\sum_{i=1}^{n} y_i = na + b \sum_{i=1}^{n} x_i \tag{7.6a}$$

and

$$\sum_{i=1}^{n} x_i y_i = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} (x_i)^2. \tag{7.6b}$$

Dividing Equation 7.6a by n leads to the conclusion that the fitted regression line must pass through the point located by the two sample means of x and y. Finally, solving the normal equations for the regression parameters yields

$$b = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{n\sum_{i=1}^{n}x_i y_i - \left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{n\sum_{i=1}^{n}(x_i)^2 - \left(\sum_{i=1}^{n}x_i\right)^2} \tag{7.7a}$$

and

$$a = \bar{y} - b\bar{x}. \tag{7.7b}$$

Equation 7.7a, for the slope, is proportional to the Pearson correlation coefficient between x and y, where the proportionality constant is given by the ratio of the two standard deviations

$$b = \frac{s_y}{s_x} r_{x,y}. \tag{7.8}$$

Accordingly the regression slope can be computed with a single pass through the data, using the computational form given as the second equality of Equation 7.7a. Note that, as was the case for the correlation coefficient, careless use of the computational form of Equation 7.7a can lead to roundoff errors since the numerator may be the difference between two large numbers.

7.2.2. Distribution of the Residuals

Thus far, fitting the straight line has involved no statistical ideas at all. All that has been required was to define least error to mean minimum squared error. The rest has followed from straightforward mathematical manipulation of the data, namely, the (x, y) pairs. To bring in statistical ideas, it is conventional to assume that the quantities e_i are independent random variables with zero mean and constant variance. Often, the additional assumption is made that these residuals follow a Gaussian distribution.

Assuming that the residuals have zero mean is not at all problematic. In fact, one convenient property of the least-squares fitting procedure is the guarantee that

$$\sum_{i=1}^{n} e_i = 0, \tag{7.9}$$

from which it is clear that the sample mean of the residuals (dividing this equation by n) is also zero.

Imagining that the residuals can be characterized in terms of a variance is really the point at which statistical ideas begin to come into the regression framework. Implicit in their possessing a variance is the idea that the residuals scatter randomly about some mean value (Equations 4.22 or 3.6). Equation 7.9 says that the mean value around which they will scatter is zero, so it is the regression line around which the data points will scatter. We then need to imagine a series of distributions of the residuals *conditional* on the x values, with each observed residual regarded as having been drawn from one of these conditional distributions. The constant variance assumption really means that the variance of the residuals is constant in x, or that all of these conditional distributions of the residuals have the same variance. Therefore a

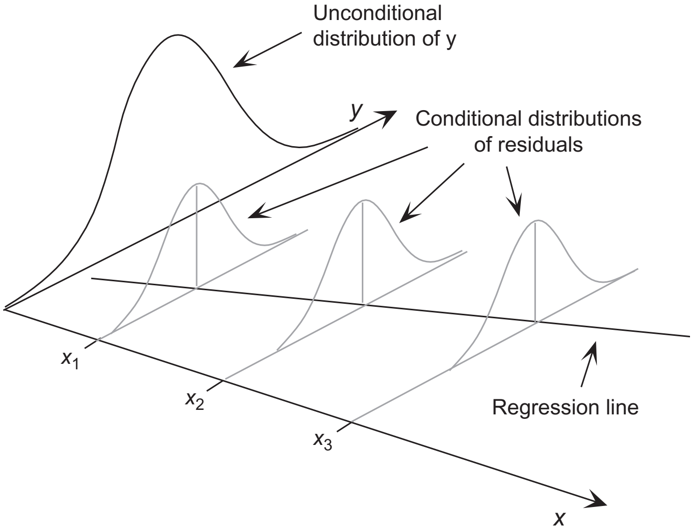

Unconditional distribution of y
y
Conditional distributions of residuals
x_1
x_2
x_3
Regression line
x

FIGURE 7.2 Schematic illustration of distributions (gray) of residuals around the regression line, conditional on these values of the predictor variable, x. The actual residuals are regarded as having been drawn from these distributions.

given residual (positive or negative, large or small) is by assumption equally likely to occur at any part of the regression line.

Figure 7.2 is a schematic illustration of the idea of a collection of conditional distributions centered on the regression line. The three small gray distributions are identical, except that their means are shifted higher or lower depending on the level of the regression line (predicted value of y) for each x. Extending this thinking slightly, it is not difficult to see that the regression equation can be regarded as specifying the conditional mean of the predictand, given a specific value of the predictor. Also shown by the large black distribution in Figure 7.2 is a schematic representation of the unconditional distribution of the predictand, y. The distributions of residuals are less spread out (have smaller variance) than the unconditional distribution of y, indicating that there is less uncertainty about y if a corresponding x value is known.

Central to the making of statistical inferences in the regression setting is estimation of this (constant) residual variance from the sample of residuals. Since the sample average of the residuals is guaranteed by Equation 7.9 to be zero, the square of Equation 3.6 becomes

$$s_e^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2, \tag{7.10}$$

where the sum of squared residuals is divided by $n-2$ because two parameters (a and b) have been estimated. Substituting Equation 7.2 then yields

$$s_e^2 = \frac{1}{n-2}\sum_{i=1}^{n} [y_i - \hat{y}(x_i)]^2. \tag{7.11}$$

Rather than compute the estimated residual variance using 7.11, however, it is more usual to use a computational form based on the relationship,

$$\text{SST} = \text{SSR} + \text{SSE}, \tag{7.12}$$

which proved in most regression texts. The notation in Equation 7.12 consists of acronyms describing the variation in the predictand, y (SST), and a partitioning of that variation between the portion represented by the regression (SSR), and the unrepresented portion ascribed to the variation of the residuals (SSE). The term SST is an acronym for sum of squares, total, which has the mathematical meaning of the sum of squared deviations of the y values around their mean,

$$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2. \tag{7.13}$$

This term is proportional (by the factor n–1) to the sample variance of y, and thus measures the overall variability of the predictand. The term SSR stands for the regression sum of squares or the sum of squared differences between the regression predictions and the sample mean of y,

$$\text{SSR} = \sum_{i=1}^{n} [\hat{y}(x_i) - \bar{y}]^2, \tag{7.14a}$$

which relates to the regression equation according to

$$\text{SSR} = b^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 = b^2 \left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right] = (n-1)\, b^2\, s_x^2. \tag{7.14b}$$

Equation 7.14a indicates that a regression line differing little from the sample mean of the y values will have a small slope and produce a very small SSR, whereas one with a large slope will exhibit some large differences from the sample mean of the predictand and therefore produce a large SSR.

Finally, SSE refers to the sum of squared errors, or sum of squared differences between the residuals and their mean, which is zero,

$$\text{SSE} = \sum_{i=1}^{n} e_i^2. \tag{7.15}$$

Since this differs from Equation 7.10 only by the factor of n–2, rearranging Equation 7.12 yields the computational form

$$s_e^2 = \frac{1}{n-2}\{\text{SST} - \text{SSR}\} = \frac{1}{n-2}\left\{ \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 - b^2 \left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right] \right\}. \tag{7.16}$$

An additional assumption that is implicit in the usual regression framework is that the predictor variable(s) is observed without error, so that the regression uncertainty relates only to the value of the predictand. Defining the residuals as vertical distances to the regression function, as in Figure 7.1, is consistent with this assumption. Although zero predictor uncertainty is rarely literally true, often uncertainty about the x values is much smaller than uncertainty about the predictions (or, equivalently, about the magnitudes of the residuals), in which case the assumption is reasonable from a practical perspective.

7.2.3. The Analysis of Variance Table

In practice, regression analysis is now almost universally done using computer software. A central part of the regression output from these software packages is a summary of the foregoing information in an

TABLE 7.1 Generic Analysis of Variance (ANOVA) Table for Simple Linear Regression

Source	df	SS	MS	F
Total	$n-1$	SST (7.13)		
Regression	1	SSR (7.14)	MSR = SSR/1	(F = MSR/MSE)
Residual	$n-2$	SSE (7.15)	MSE = s_e^2	

The column headings df, SS, and MS stand for degrees of freedom, sum of squares, and mean square, respectively. Regression df = 1 is particular to simple linear regression (i.e., a single predictor x). Parenthetical references are to equation numbers in the text.

analysis of variance, or ANOVA table. The ANOVA table was introduced in Section 5.5 as a vehicle for making inferences about mean responses in designed experiments. Such analyses can be viewed as special cases of regression, where the predictor variables are binary, $x \in \{0, 1\}$ (e.g., Draper and Smith, 1998), which are sometimes referred to as "dummy variables."

Usually, not all the information in an ANOVA table will be of interest, but it is such a universal form of regression output that you should understand its components. Table 7.1 outlines the arrangement of an ANOVA table for simple linear regression and indicates where the quantities described in the previous section are reported. The three rows correspond to the partition of the variation of the predictand as expressed in Equation 7.12. Accordingly, the Regression and Residual entries in the df (degrees of freedom) and SS (sum of squares) columns will sum to the corresponding entry in the Total row. Therefore the ANOVA table contains some redundant information, and as a consequence the output from some regression packages will omit the Total row entirely.

The entries in the MS (mean squared) column are given by the corresponding quotients of SS/df. For simple linear regression, the regression df = 1, and SSR = MSR. Comparing with Equation 7.16, it can be seen that the MSE (mean squared error) is the estimated sample variance of the residuals. The total mean square, left blank in Table 7.1 and in the output of most regression packages, would be SST/(n – 1), or simply the sample variance of the predictand.

7.2.4. Goodness-of-Fit Measures

The ANOVA table also presents (or provides sufficient information to compute) three related measures of the fit of a regression, or the correspondence between the regression line and a scatterplot of the data. The first of these is the MSE. From the standpoint of forecasting, the MSE is perhaps the most fundamental of the three measures, since it indicates the variability of, or the uncertainty about, the observed y values (the quantities being forecast) around the forecast regression line. As such, it directly reflects the average accuracy of the resulting forecasts. Referring again to Figure 7.2, since MSE = s_e^2 this quantity indicates the degree to which the distributions of residuals cluster tightly (small MSE), or spread widely (large MSE) around a regression line. In the limit of a perfect linear relationship between x and y, the regression line coincides exactly with all the point pairs, the residuals are all zero, SST will equal SSR, SSE will be zero, and the variance of the residual distributions is also zero. In the opposite limit of absolutely no linear relationship between x and y, the regression slope will be zero, the SSR will be zero, SSE will equal SST, and the MSE will very nearly equal the sample variance of the predictand itself. In this unfortunate case, the three conditional distributions in Figure 7.2 would be indistinguishable from the unconditional distribution of y.

The relationship of the MSE to the strength of the regression fit is also illustrated in Figure 7.3. Panel (a) shows the case of a reasonably good regression, with the scatter of points around the regression line being fairly small. Here SSR and SST are nearly the same. Panel (b) shows an essentially useless regression, for values of the predictand spanning the same range as in panel (a). In this case the SSR is nearly zero since the regression has nearly zero slope, and the MSE is essentially the same as the sample variance of the y values themselves.

The second usual measure of the fit of a regression is the *coefficient of determination*, or R^2. This can be computed from

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}, \tag{7.17}$$

which is often also displayed as part of standard regression output. The SSR is nearly equal to SST if each predicted value is close to its respective y, so that the corresponding residual is near zero. Therefore MSE and R^2 are different but related ways of expressing the closeness of or discrepancy between SST and SSR. The R^2 can be interpreted as the proportion of the variation of the predictand (proportional to SST) that is described or accounted for by the regression (SSR). Sometimes we see this concept expressed as the proportion of variation "explained," although this claim is misleading: a regression analysis can quantify the nature and strength of a relationship between two variables, but can say nothing about which variable (if either) causes the other. This is the same caveat that was offered in the discussion of the correlation coefficient in Chapter 3. For the case of simple linear regression, the square root of the coefficient of determination is exactly (the absolute value of) the Pearson correlation between x and y.

For a perfect regression, SSR = SST and SSE = 0, so $R^2 = 1$. For a completely useless regression, SSR = 0 and SSE = SST, so that $R^2 = 0$. Again, Figure 7.3b shows something close to this latter case. Comparing Equation 7.14a, the least-squares regression line is almost indistinguishable from the sample mean of the predictand, so SSR is very small. In other words, little of the variation in y can be ascribed to the regression so the proportion SSR/SST is nearly zero.

The third commonly used measure of the strength of the regression is the F ratio, generally given in the last column of the ANOVA table. The ratio MSR/MSE increases with the strength of the regression, since a strong relationship between x and y will produce a large MSR and a small MSE. Assuming that the residuals are independent and follow the same Gaussian distribution, and under the null hypothesis of no real linear

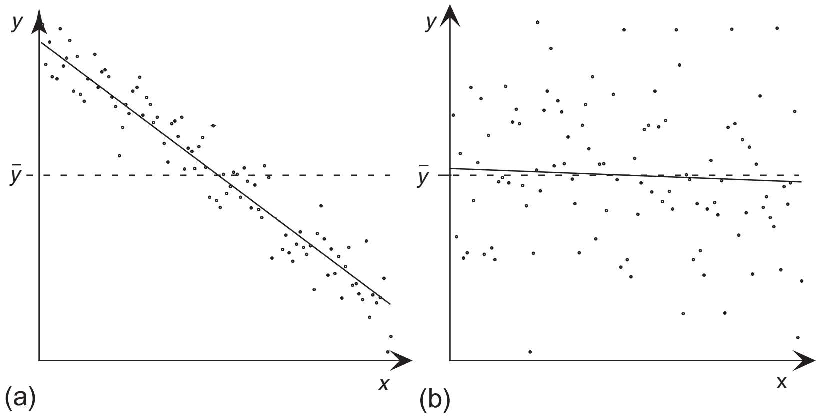


FIGURE 7.3 Illustration of the distinction between a fairly good regression relationship (a) and an essentially useless relationship (b). The points in panel (a) cluster closely around the regression line (solid), indicating small MSE, and the line deviates strongly from the average value of the predictand (dashed), producing a large SSR. In panel (b) the scatter around the regression line is large, and the regression line is almost indistinguishable from the mean of the predictand.

relationship, the sampling distribution of the F ratio has a known parametric form. Analogously to Equation 5.45, this distribution forms the basis of a test that is applicable in the case of simple linear regression if the single predictor has been chosen in advance of seeing the data subject to the analysis, but in the more general case of multiple regression (more than one x variable) problems of test multiplicity, to be discussed later, usually invalidate it. However, even if the F ratio cannot be used for quantitative statistical inference, it is still a valid qualitative index of the strength of a regression. See, for example, Draper and Smith (1998) or Neter et al. (1996) for discussions of the F test for overall significance of the regression.

7.2.5. Sampling Distributions of the Regression Coefficients

Another important use of the estimated residual variance is to obtain estimates of the sampling distributions of the regression coefficients. As statistics computed from a finite set of data subject to sampling variations, the computed regression intercept and slope, a and b, also exhibit sampling variability. That is, different batches of size n from the same data-generating process will yield different pairs of regression slopes and intercepts, and their sampling distributions characterize this batch-to-batch variability. Estimation of these sampling distributions allows construction of confidence intervals for the true population counterparts around the sample intercept and slope values, and provides a basis for hypothesis tests about the corresponding population values.

Under the assumptions listed previously, the sampling distributions for both intercept and slope are Gaussian. On the strength of the Central Limit Theorem, this result also holds at least approximately for any regression when n is large enough, because the estimated regression parameters (Equation 7.7) are obtained as the sums of large numbers of random variables. For the intercept the sampling distribution has parameters

$$\mu_a = a \tag{7.18a}$$

and

$$\sigma_a = s_e \left[\frac{\sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2} \right]^{1/2}. \tag{7.18b}$$

For the slope the parameters of the sampling distribution are

$$\mu_b = b \tag{7.19a}$$

and

$$\sigma_b = \frac{s_e}{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2 \right]^{1/2}}. \tag{7.19b}$$

Equations 7.18a and 7.19a indicate that the least-squares regression parameter estimates are unbiased. Equations 7.18b and 7.19b show that the precision with which the intercept and slope can be estimated from the data depends directly on the estimated standard deviation of the residuals, s_e, which is the square root of the MSE from the ANOVA table (Table 7.1). Additionally, the estimated slope and intercept are not independent, having correlation

$$r_{a,b} = \frac{-\bar{x}}{\frac{1}{n}\left(\sum_{i=1}^{n} x_i^2\right)^{1/2}}. \tag{7.20}$$

Taken together with the (at least approximately) Gaussian sampling distributions for a and b, Equations 7.18–7.20 define their joint bivariate normal (Equation 4.31) distribution. Equations 7.18b, 7.19b, and 7.20 are valid only for simple linear regression. With more than one predictor variable, analogous (vector) equations (Equation 11.40) must be used.

The output from regression packages will almost always include the standard errors (Equations 7.18b and 7.19b) in addition to the parameter estimates themselves. Some packages also include the ratios of the estimated parameters to their standard errors in a column labeled "*t* ratio." When this is done, a one-sample *t*-test (Equation 5.3) is implied, with the null hypothesis being that the underlying (population) mean for the parameter is zero. Sometimes a *p* value associated with this test is also automatically included in the regression output.

For the regression slope, this implicit *t*-test bears directly on the meaningfulness of the fitted regression. If the estimated slope is small enough that its true value could plausibly (with respect to its sampling distribution) be zero, then the regression is not informative, or useful for forecasting. If the slope is actually zero, then the value of the predictand specified by the regression equation is always the same and equal to its sample mean (cf. Equations 7.1 and 7.7b). If the assumptions regarding the regression residuals are satisfied, we would reject this null hypothesis at the 5% level if the estimated slope is, roughly, at least twice as large (in absolute value) as its standard error.

The same hypothesis test for the regression intercept often is offered by computerized statistical packages as well. Depending on the problem at hand, however, this test for the intercept may or may not be meaningful. Again, the *t* ratio is just the parameter estimate divided by its standard error, so the implicit null hypothesis is that the true intercept is zero. Occasionally, this null hypothesis is physically meaningful, and if so the test statistic for the intercept is worth looking at. On the other hand, it often happens that there is no physical reason to expect that the intercept might be zero. It may even be that a zero intercept is physically impossible. In such cases this portion of the automatically generated computer output is meaningless.

Example 7.1. A Simple Linear Regression

To concretely illustrate simple linear regression, consider the January 1987 minimum temperatures at Ithaca and Canandaigua from Table A.1 in Appendix A. Let the predictor variable, x, be the Ithaca minimum temperature, and the predictand, y, be the Canandaigua minimum temperature. The scatterplot of this data is shown the middle panel of the bottom row of the scatterplot matrix in Figure 3.31, and as part of Figure 7.10. A fairly strong, positive, and reasonably linear relationship is indicated.

Table 7.2 presents what the output from a typical statistical computer package would look like for this regression. The data set is small enough that the computational formulas can be worked through to verify the results. (A little work with a hand calculator will verify that $\Sigma x = 403$, $\Sigma y = 627$, $\Sigma x^2 = 10{,}803$, $\Sigma y^2 = 15{,}009$, and $\Sigma xy = 11{,}475$.) The upper portion of Table 7.2 corresponds to the template in Table 7.1, with the relevant numbers filled in. Of particular importance is MSE = 11.780, yielding as its square root the estimated sample standard deviation for the residuals, $s_e = 3.43$°F. This standard deviation addresses directly the precision of specifying the Canandaigua temperatures on the basis of the concurrent Ithaca temperatures, since we expect about 95% of the actual predictand values to be within $\pm 2s_e = \pm 6.9$°F of the temperatures given by the regression. The coefficient of determination is easily

TABLE 7.2 Example Output Typical of that Produced by Computer Statistical Packages, for Prediction of Canandaigua Minimum Temperature (*y*) using Ithaca Minimum Temperature (*x*) as the Predictor, from the January 1987 Data Set in Table A.1

Source	df	SS	MS	F
Total	30	2327.419		
Regression	1	1985.798	1985.798	168.57
Residual	29	341.622	11.780	
Variable	Coefficient	s.e.	*t* ratio	
Constant	12.4595	0.8590	14.504	
Ithaca Min	0.5974	0.0460	12.987	

computed as $R^2 = 1985.798/2327.419 = 85.3\%$. The Pearson correlation is $\sqrt{0.853} = 0.924$, as was given in Table 3.5. The value of the *F* statistic is very high, considering that the 99th percentile of its distribution under the null hypothesis of no real relationship is about 7.5. We also could compute the sample variance of the predictand, which would be the total mean square cell of the table, as $2327.419/30 = 77.58°\text{F}^2$.

The lower portion of Table 7.2 gives the regression parameters, *a* and *b*, their standard errors, and the ratios of these parameter estimates to their standard errors. The specific regression equation for this data set, corresponding to Equation 7.1, would be

$$T_{\text{Can.}} = \underset{(0.859)}{12.46} + \underset{(0.046)}{0.597}\, T_{\text{Ith.}}. \tag{7.21}$$

Thus the Canandaigua temperature would be estimated by multiplying the Ithaca temperature by 0.597 and adding 12.46°F. The intercept $a = 12.46°\text{F}$ has no special physical significance except as the predicted Canandaigua temperature when the Ithaca temperature is 0°F. Notice that the standard errors of the two coefficients have been written parenthetically below the coefficients themselves. Although this is not a universal practice, it is very informative to someone reading Equation 7.21 without the benefit of the information in Table 7.2. In particular, it allows the reader to get a sense for the significance of the slope (i.e., the parameter *b*). Since the estimated slope is about 13 times larger than its standard error it is almost certainly not really zero. This conclusion speaks directly to the question of the meaningfulness of the fitted regression. On the other hand, the corresponding implied hypothesis test for the intercept is much less interesting, because the possibility of a zero intercept is of no physical interest. ◇

7.2.6. Examining Residuals

It is not sufficient to feed data to a computer regression package and uncritically accept the results. Some of the results can be misleading if the assumptions underlying the computations are not satisfied. Since these assumptions pertain to the residuals, it is important to examine the residuals for consistency with the assumptions made about their behavior.

One easy and fundamental check on the residuals can be made by examining a scatterplot of the residuals as a function of the predicted value $\hat{y}$. Many statistical computer packages provide this capability as a standard regression option. Figure 7.4a shows the scatterplot of a hypothetical data set, with the least-squares regression line, and Figure 7.4b shows a plot for the resulting residuals as a function of the predicted values. The residual plot presents the impression of "fanning," or exhibition of increasing spread as $\hat{y}$ increases. That is, the variance of the residuals appears to increase as the predicted value increases. This condition of nonconstant residual variance is called *heteroscedasticity*. Since the computer program that fit the regression has assumed constant residual variance, the MSE given in the ANOVA table is an overestimate for smaller values of x and y (where the points cluster closer to the regression line), and an underestimate of the residual variance for larger values of x and y (where the points tend to be further from the regression line). If the regression is used as a forecasting tool, we would be overconfident about forecasts for larger values of y, and underconfident about forecasts for smaller values of y. In addition, the sampling distributions of the regression parameters will be more variable than implied by Equations 7.18 and 7.19. That is, the parameters will not have been estimated as precisely as the standard regression output would lead us to believe.

Often nonconstancy of residual variance of the sort shown in Figure 7.4b can be remedied by transforming the predictand y, perhaps by using a power transformation (Equations 3.20 or 3.23). Figure 7.5 shows the regression and residual plots for the same data as in Figure 7.4 after logarithmically transforming the predictand. Recall that the logarithmic transformation reduces all the data values, but reduces the larger values more strongly than the smaller ones. Thus the long right tail of the predictand has been pulled in relative to the shorter left tail, as in Figure 3.13. As a result, the transformed data points

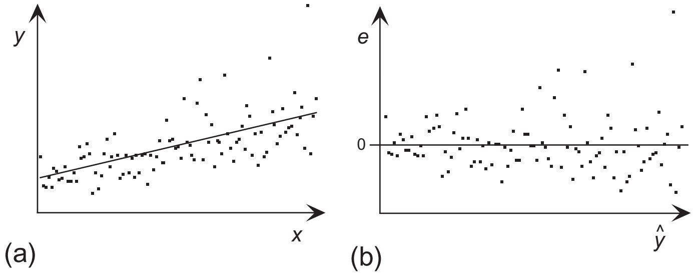


FIGURE 7.4 Hypothetical linear regression (a), and plot of the resulting residuals against the predicted values (b), for a case where the variance of the residuals is not constant. The scatter around the regression line in (a) increases for larger values of x and y, producing a visual impression of "fanning" in the residual plot (b). A transformation of the predictand is indicated.

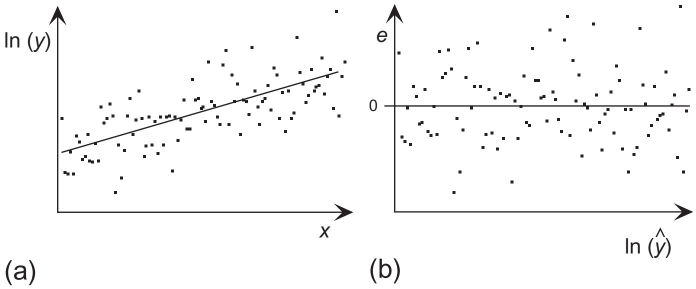


FIGURE 7.5 Scatterplot with regression (a), and resulting residual plot (b), for the same data in Figure 7.4, after logarithmically transforming the predictand. The visual impression of a horizontal band in the residual plot supports the assumption of constant variance of the residuals.

appear to cluster more evenly around the new regression line. Instead of fanning, the residual plot in Figure 7.5b gives the visual impression of a horizontal band, indicating appropriately constant variance of the residuals (*homoscedasticity*). Note that if the fanning in Figure 7.4b had been in the opposite sense, with greater residual variability for smaller values of $\hat{y}$ and lesser residual variability for larger values of $\hat{y}$, a transformation that stretches the right tail relative to the left tail (e.g., y^2) would have been appropriate.

It can also be informative to look at scatterplots of residuals as a function of the predictor variable. Figure 7.6 illustrates some of the forms such plots can take and their diagnostic interpretations. Figure 7.6a is similar to Figure 7.4b, in that the fanning of the residuals indicates nonconstancy of variance. Figure 7.6b illustrates a different form of heteroscedasticity that might be more challenging to remedy through a variable transformation. The type of residual plot in Figure 7.6c, with a linear dependence on the predictor of the linear regression, indicates that either the intercept a has been omitted or that the calculations have been done incorrectly. Deliberately omitting a regression intercept, called "forcing through the origin," is useful in some circumstances, but may not be appropriate even if it is known beforehand that the true relationship should pass through the origin. Particularly if data are available over only a restricted range, or if the actual relationship is nonlinear, a linear regression including an intercept term may yield better predictions. In this latter case a simple linear regression would be analogous to a first-order Taylor approximation about the mean of the training data.

Figure 7.6d shows a form for the residual plot that can occur when additional predictors would improve a regression relationship. Here the variance is reasonably constant in x, but the (conditional) average residual exhibits a dependence on x. Figure 7.6e illustrates the kind of behavior that can occur when a single outlier in the data has undue influence on the regression. Here the regression line has been pulled toward the outlying point in order to avoid the large squared error associated with it, leaving a trend in the other residuals. If the outlier were determined not to be a valid data point, it should either be corrected if possible or otherwise discarded. If it is a valid data point, a resistant approach such as LAD

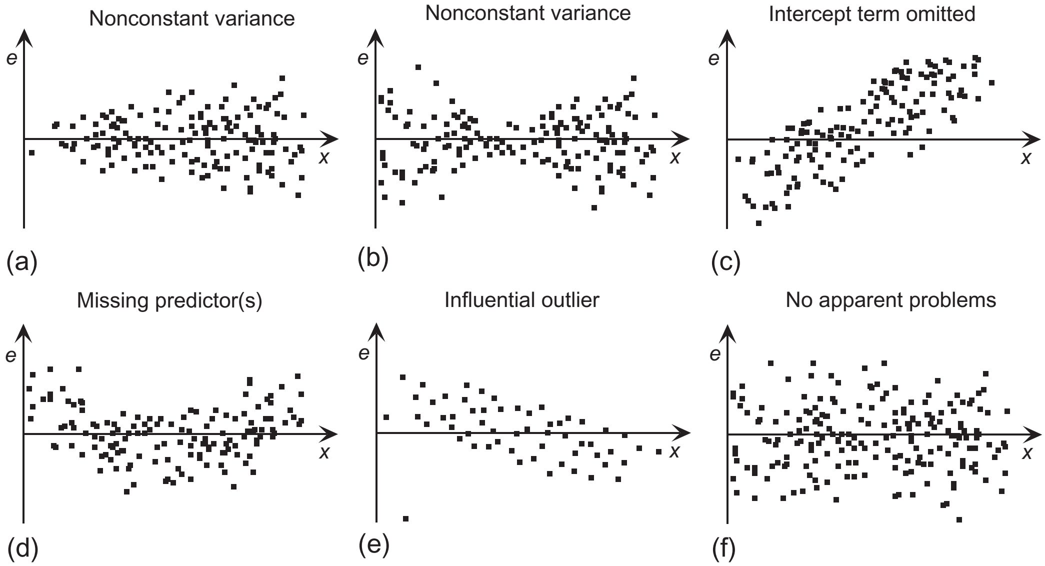


FIGURE 7.6 Idealized scatterplots of regression residuals vs. a predictor x, with corresponding diagnostic interpretations.

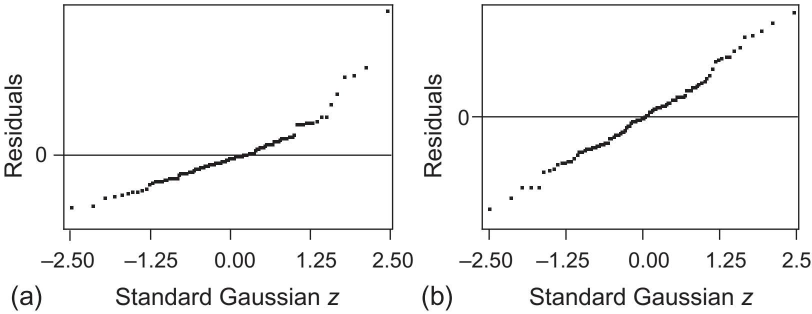


FIGURE 7.7 Gaussian quantile-quantile plots of the residuals for predictions of (a) the untransformed predictand in Figure 7.4a, and (b) the logarithmically transformed predictand in Figure 7.5. In addition to producing essentially constant residual variance, logarithmic transformation of the predictand has rendered the distribution of the residuals effectively Gaussian.

regression (Section 7.7.3) or median-slope regression (Section 7.7.2) might be more appropriate. Figure 7.6f again illustrates the desirable horizontally banded pattern of residuals, similar to Figure 7.5b.

A graphical impression of whether the residuals follow a Gaussian distribution can be obtained through a Q-Q plot. Such plots are often a standard option in statistical computer packages. Figure 7.7a and b show Q-Q plots for the residuals from Figures 7.4b and 7.5b, respectively. The residuals are plotted on the vertical, and the standard Gaussian variables corresponding to the empirical cumulative probability of each residual are plotted on the horizontal. The curvature apparent in Figure 7.7a indicates that the residuals from the regression involving the untransformed predictand are positively skewed relative to the (symmetric) Gaussian distribution. The Q-Q plot of residuals from the regression involving the logarithmically transformed predictand is very nearly linear. Evidently the logarithmic transformation has produced residuals that are close to Gaussian, in addition to stabilizing the residual variances. Similar conclusions could have been reached using a goodness-of-fit test (see Section 5.2.5).

It is also possible and desirable to investigate the degree to which the residuals are uncorrelated. This question is of particular interest when the underlying data are serially correlated, which is a common condition for atmospheric variables. A simple graphical evaluation can be obtained by plotting the regression residuals as a function of time. If groups of positive and negative residuals tend to cluster together (qualitatively resembling Figure 5.4b) rather than occurring more irregularly (as in Figure 5.4a), then time correlation can be suspected.

The *Durbin-Watson test* is a popular formal test for serial correlation of regression residuals that is included in many computer regression packages. This test examines the null hypothesis that the residuals are serially independent, against the alternative that they are consistent with a first-order autoregressive process (Equation 10.16). The Durbin-Watson test statistic,

$$d = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}, \tag{7.22}$$

computes the squared differences between pairs of consecutive residuals, divided by a scaling factor proportional to the residual variance. If the residuals are positively correlated, adjacent residuals will tend to be similar in magnitude, so the Durbin-Watson statistic will be relatively small. If the sequence

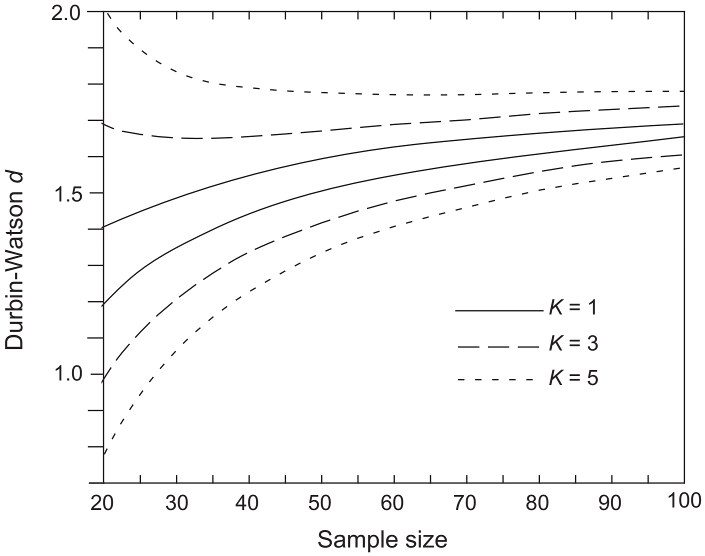


FIGURE 7.8 5%-level critical values for the Durbin-Watson statistic as a function of the sample size, for $K = 1, 3$, and 5 predictor variables. A test statistic d below the relevant lower curve results in a rejection of the null hypothesis of zero serial correlation. If the test statistic is above the relevant upper curve the null hypothesis is not rejected. If the test statistic is between the two curves the test is indeterminate without additional calculations.

of residuals is randomly distributed, the sum in the numerator will tend to be larger. Therefore the null hypothesis that the residuals are independent is rejected if the Durbin-Watson statistic is sufficiently small.

Figure 7.8 shows critical values for Durbin-Watson tests at the 5% level. These vary depending on the sample size, and the number of predictor (x) variables, K. For simple linear regression, $K = 1$. For each value of K, Figure 7.8 shows two curves. If the observed value of the test statistic falls below the lower curve, the null hypothesis is rejected and we conclude that the residuals exhibit significant serial correlation. If the test statistic falls above the upper curve, we do not reject the null hypothesis that the residuals are serially uncorrelated. If the test statistic falls between the two relevant curves, the test is indeterminate. The reason behind the existence of this unusual indeterminate condition is that the null distribution of the Durban-Watson statistic depends on the data set being considered. In cases where the test result is indeterminate according to Figure 7.8, some additional calculations (Durban and Watson, 1971) can be performed to resolve the indeterminacy, that is, to find the specific location of the critical value between the appropriate pair of curves, for the particular data at hand.

Example 7.2. Examination of the Residuals from Example 7.1

A regression equation constructed using autocorrelated variables as predictand and predictor(s) does not necessarily exhibit strongly autocorrelated residuals. Consider again the regression between Ithaca and Canandaigua minimum temperatures for January 1987 in Example 7.1. The lag-1 autocorrelations (Equation 3.36) for the Ithaca and Canandaigua minimum temperature data are 0.651 and 0.672, respectively. The residuals for this regression are plotted as a function of time in Figure 7.9. A strong serial correlation for these residuals is not apparent, and their lag-1 autocorrelation as computed using Equation 3.36 is only 0.191.

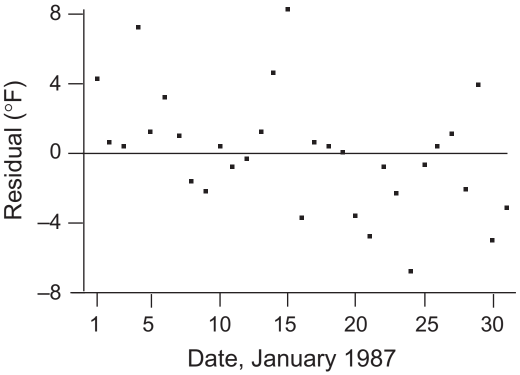


FIGURE 7.9 Residuals from the regression, Equation 7.21, plotted as a function of the date. A strong serial correlation is not apparent, but the tendency for a negative slope suggests that the relationship between Ithaca and Canandaigua temperatures may be changing through the month.

Having obtained the residuals for the Canandaigua vs. Ithaca minimum temperature regression, it is straightforward to compute the Durbin-Watson d (Equation 7.22). In fact, the denominator is simply the SSE from the ANOVA Table 7.2, which is 341.622. The numerator in Equation 7.22 must be computed from the residuals and is 531.36. These yield $d = 1.55$. Referring to Figure 7.8, the point at $n = 31$, $d = 1.55$ is well above the upper solid (for $K = 1$, since there is a single predictor variable) line, so the null hypothesis of uncorrelated residuals would not be rejected at the 5% level. Evidently the strong serial correlation exhibited by the predictand has been captured in the regression, through the strong serial correlation in the predictor. ◇

When regression residuals are autocorrelated, statistical inferences based upon their variance are degraded in the same way, and for the same reasons, that were discussed in Section 5.2.4 (e.g., Matalas and Sankarasubramanian, 2003; Santer et al., 2000; Zheng et al., 1997). In particular, positive serial correlation of the residuals leads to inflation of the variance of the sampling distribution of their sum or average, because these quantities are less consistent from batch to batch of size n. When a first-order autoregression (Equation 10.16) is a reasonable representation for these correlations (characterized by r_1) it is appropriate to apply the same variance inflation factor, $(1 + r_1)/(1–r_1)$ (the bracketed quantity in Equation 5.13), to the variance s_e^2 in, for example, Equations 7.18b and 7.19b (Matalas and Sankarasubramanian, 2003; Santer et al., 2000). The net effect is that the variance of the resulting sampling distribution is (appropriately) increased, relative to what would be calculated assuming independent regression residuals.

7.2.7. Prediction Intervals

Many times it is of interest to calculate *prediction intervals* around forecast values of the predictand (i.e., around the regression function), which are meant to bound a future value of the predictand with specified probability. When it can be assumed that the residuals follow a Gaussian distribution, it is natural to approach this problem using the unbiasedness property of the residuals (Equation 7.9), together with their estimated variance $\text{MSE} = s_e^2$. In terms of Figure 7.2, suppose the regression is being evaluated at a predictor value that is equal to x_3. The uncertainty associated with a corresponding future unknown value of the predictand would be represented approximately by the rightmost gray conditional residual distribution in Figure 7.2. Using Gaussian probabilities (Table B.1), we expect a 95% prediction interval for a future residual, or specific future forecast, to be approximately bounded by $\hat{y} \pm 2s_e$.

The $\pm 2s_e$ rule of thumb is often a quite good approximation to the width of a true 95% prediction interval, especially when the sample size is large. However, because both the sample mean of the predictand and the slope of the regression are subject to sampling variations, the prediction variance for future data (i.e., for data not used in the fitting of the regression) is somewhat larger than the regression MSE. For a forecast of y using the predictor value x_0, this prediction variance is given by

$$s_{\hat{y}}^2 = s_e^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]. \tag{7.23}$$

That is, the prediction variance is proportional to the regression MSE, but is larger to the extent that the second and third terms inside the square brackets are appreciably larger than zero. The second term derives from the uncertainty in estimating the true mean of the predictand from a finite sample of size n (compare Equation 5.4), and becomes very small for large sample sizes. The third term derives from the uncertainty in estimation of the slope (it is similar in form to Equation 7.19b), and indicates that predictions far removed from the center of the data used to fit the regression will be more uncertain than predictions made near the sample mean. However, even if the numerator in this third term is fairly large, the term itself will tend to be small if a large data sample was used to construct the regression equation, since there are n nonnegative terms of generally comparable magnitude in the denominator.

It is sometimes also of interest to compute *confidence intervals* for the regression function itself. These will be narrower than the prediction intervals for future individual data values, reflecting a smaller variance in a way that is analogous to the variance of a sample mean being smaller than the variance of its underlying data values. The variance for the sampling distribution of the regression function, or equivalently the variance of the conditional mean of the predictand given a particular predictor value x_0, is

$$s_{\bar{y}|x_0}^2 = s_e^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]. \tag{7.24}$$

This expression is similar to Equation 7.23, but is smaller by the amount s_e^2. That is, there are contributions to this variance due to uncertainty in the mean of the predictand (or, equivalently the vertical position of the regression line, or the intercept), corresponding to the first of the two terms in the square brackets; and to uncertainty in the slope, corresponding to the second term. There is no contribution to Equation 7.24 reflecting scatter of data around the regression line, which is the difference between Equations 7.23 and 7.24.

Figure 7.10 compares prediction and confidence intervals computed using Equations 7.23 and 7.24, in the context of the regression from Example 7.1. Here the regression (Equation 7.21) fit to the 31 data points (dots) is shown by the heavy solid line. The 95% prediction interval around the regression computed as $\pm 1.96\, s_{\hat{y}}$, using the square root of Equation 7.23, is indicated by the pair of slightly curved solid black lines. As noted earlier, these bounds are only slightly wider than those given by the simpler approximation $\hat{y} \pm 1.96\, s_e$ (dashed lines), because the second and third terms in the square brackets of

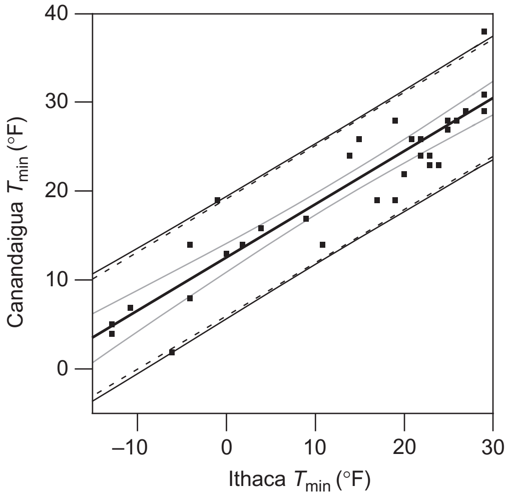


FIGURE 7.10 Prediction and confidence intervals around the regression derived in Example 7.1 (thick black line). Light solid lines indicate 95% prediction intervals for future data, computed using Equation 7.23, and the corresponding dashed lines simply locate the predictions ±1.96 s_e. Light gray lines locate 95% confidence intervals for the regression function (Equation 7.24). Data to which the regression was fit are also shown.

Equation 7.23 are relatively small, even for moderate n. The pair of gray curved lines locates the 95% confidence interval for the conditional mean of the predictand. These are much narrower than the prediction interval because they account only for sampling variations in the regression parameters, without direct contributions from the prediction variance s_e^2.

Equations 7.18 through 7.20 define the parameters of a bivariate normal distribution for the two regression parameters. Imagine using the methods outlined in Section 4.7 to generate pairs of intercepts and slopes according to that distribution, and therefore to generate realizations of plausible regression lines. One interpretation of the gray curves in Figure 7.10 is that they would contain 95% of those regression lines (or, equivalently, 95% of the regression lines computed from different samples of data of this kind, each with size $n = 31$). The minimum separation between the gray curves (at the average Ithaca $T_{min} = 13$°F) reflects the uncertainty in the intercept. Their spreading at more extreme temperatures reflects the fact that uncertainty in the slope (i.e., uncertainty in the angle of the regression line) will produce more uncertainty in the conditional expected value of the predictand at the extremes than near the mean, because Equations 7.6a and 7.7b show that any regression line must pass through the point located by the two sample means.

The result in Example 7.2 indicates that the residuals for this regression can reasonably be regarded as independent. Also, some of the sample lag-1 autocorrelation of $r_1 = 0.191$ can be attributable to the time trend evident in Figure 7.9. However, if the residuals are significantly correlated, and the nature of that correlation is plausibly represented by a first-order autoregression (Equation 10.16), it would be appropriate to increase the residual variances s_e^2 in Equations 7.23 and 7.24 by multiplying them by the variance inflation factor $(1 + r_1)/(1 – r_1)$.

Special care is required when computing prediction and confidence intervals for regressions involving transformed predictands. For example, if the relationship shown in Figure 7.5a (involving a log-transformed predictand) were to be used in forecasting, dimensional values of the predictand would need to be recovered in order to make the forecasts interpretable. That is, the predictand $\ln(\hat{y})$ would need to be back-transformed, yielding the forecast $\hat{y} = \exp.[\ln(\hat{y})] = \exp.[a + bx]$. Similarly, the limits of the prediction intervals would also need to be back-transformed. For example, the 95% prediction interval would be approximately $\ln(\hat{y}) \pm 1.96\ s_e$, because the regression residuals and their assumed Gaussian distribution pertain to the transformed predictand values. The lower and upper limits of this interval, when expressed on the original untransformed scale of the predictand, would be approximately $\exp[a + bx - 1.96\ s_e]$ and $\exp[a + bx + 1.96 s_e]$. These limits would not be symmetrical around $\hat{y}$, and would extend further for the larger values, consistent with the longer right tail of the predictand distribution.

Equations 7.23 and 7.24 are valid for simple linear regression. The corresponding equations for multiple regression are similar, but are more conveniently expressed in matrix algebra notation (see Example 11.2). However, as is the case for simple linear regression, the prediction variance is quite close to the MSE for moderately large samples.

7.3. MULTIPLE LINEAR REGRESSION

7.3.1. Extending Simple Linear Regression

Multiple linear regression is the more general (and more common) implementation of linear regression. As in the case of simple linear regression, there is still a single predictand, y, but in distinction there is more than one predictor (x) variable. The preceding treatment of simple linear regression was relatively lengthy, in part because most of what was presented generalizes readily to the case of multiple linear regression.

Let K denote the number of predictor variables. Simple linear regression is then the special case of $K = 1$. The prediction equation (corresponding to Equation 7.1) becomes

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_K x_K. \tag{7.25}$$

Each of the K predictor variables has its own coefficient, analogous to the slope, b, in Equation 7.1. For notational convenience, the intercept (or *regression constant*) is denoted as b_0 rather than a, as in Equation 7.1. These $K + 1$ regression coefficients often are called the *regression parameters*.

Equation 7.2 for the residuals is still valid, if it is understood that the predicted value $\hat{y}$ is a function of a vector of predictors, x_k, $k = 1, \ldots, K$. If there are $K = 2$ predictor variables, the residual can still be visualized as a vertical distance. In that case the regression function (Equation 7.25) is a surface rather than a line, and the residual corresponds geometrically to the distance above or below this surface along a line perpendicular to the (x_1, x_2) plane. The geometric situation is analogous for $K \geq 3$, but is not easily visualized. Also in common with simple linear regression, the average residual is guaranteed to be zero, so that the residual distributions are centered on the predicted values $\hat{y}_i$. Accordingly, these predicted values can be regarded as conditional means given particular values for a set of K predictors.

The $K + 1$ parameters in Equation 7.25 are found, as before, by minimizing the sum of squared residuals. This is achieved by simultaneously solving $K + 1$ equations analogous to Equation 7.5. This minimization is most conveniently done using matrix algebra, the details of which can be found in standard regression texts (e.g., Draper and Smith, 1998; Neter et al., 1996). The basics of the process are outlined in Example 11.2. In practice, the calculations usually are done using statistical software. They are again summarized in an ANOVA table, of the form shown in Table 7.3. As before, SST is

TABLE 7.3 Generic Analysis of Variance (ANOVA) Table for Multiple Linear Regression

Source	df	SS	MS	F
Total	$n-1$	SST		
Regression	K	SSR	$\text{MSR}=\text{SSR}/K$	$F=\text{MSR}/\text{MSE}$
Residual	$n-K-1$	SSE	$\text{MSE}=\text{SSE}/(n-K-1)=s_e^2$	

Table 7.1 for simple linear regression can be viewed as a special case, with $K=1$.

computed using Equation 7.13, SSR is computed using Equation 7.14a, and SSE is computed using the difference SST – SSR. The sample variance of the residuals is MSE = SSE/(n–K–1).

The coefficient of determination is computed according to Equation 7.17, although in the more general case of multiple linear regression it is the square of the Pearson correlation coefficient between the observed and predicted values of the predictand y. This generalizes the meaning of R^2 as the square of the Pearson correlation between x and y for simple linear regression, because in that setting the predicted value is a linear function of the single predictor x, and correlation is insensitive to linear transformations.

Expressions for the variances characterizing prediction uncertainty and conditional-mean predictand uncertainty in multiple regression, generalizing Equations 7.23 and 7.24, are given by Equations 11.43 and 11.42, respectively. The procedures presented previously for examination of residuals are applicable to multiple regression as well.

7.3.2. Derived Predictor Variables in Multiple Regression

Multiple regression opens up the possibility of an essentially unlimited number of potential predictor variables. An initial list of potential predictor variables can be expanded manyfold by also considering nonlinear mathematical transformations of these variables as potential predictors. The derived predictors must be nonlinear functions of the primary predictors in order for the computations (in particular, for the matrix inversion indicated in Equation 11.39) to be possible. Such *derived predictors* can be very useful in producing a good regression equation.

In some instances the most appropriate forms for predictor transformations may be suggested by a physical understanding of the data-generating process. In the absence of a strong physical rationale for particular predictor transformations, the choice of a transformation or set of transformations may be made purely empirically, perhaps by subjectively evaluating the general shape of the point cloud in a scatterplot, or the nature of the deviation of a residual plot from its ideal form. For example, the curvature in the residual plot in Figure 7.6d suggests that addition of the derived predictor $x_2 = x_1^2$ might improve the regression relationship. It may happen that the empirical choice of a transformation for a predictor variable in regression leads to a greater physical understanding, which is a highly desirable outcome in a research setting. This outcome would be less important in a purely forecasting setting, where the emphasis is on producing good forecasts rather than knowing precisely why the forecasts are good.

Transformations such as $x_2 = x_1^2$, $x_2 = \sqrt{x_1}$, $x_2 = 1/x_1$, or any other power transformation of an available predictor, can be adopted as potential predictors. Similarly, trigonometric (sine, cosine, etc.), exponential or logarithmic functions, or combinations of these are useful in some situations. Another commonly used transformation is to a *binary variable*, also known as a *dummy variable* or *indicator variable*. Binary variables take on one of two values (usually 0 and 1, although the particular choices do not affect subsequent use of the regression equation), depending on whether the variable being transformed is above or below a threshold or cutoff, c. That is, a binary variable x_2 could be constructed from another predictor x_1 according to the transformation

$$x_2 = \begin{cases} 1, & \text{if } x_1 > c \\ 0, & \text{if } x_1 \le c \end{cases}. \tag{7.26}$$

More than one binary predictor can be constructed from a single x_1 by choosing different values of the cutoff, c, for x_2, x_3, x_4, and so on.

Even though transformed variables may be nonlinear functions of other variables, the overall framework is still known as multiple linear regression. Once a derived variable has been defined it is just another variable, regardless of how the transformation was made. More formally, the "linear" in multiple linear regression refers to the regression equation being linear in the parameters, b_k.

Example 7.3. A Multiple Regression with Derived Predictor Variables

Figure 7.11 shows a scatterplot of the famous Keeling monthly averaged carbon dioxide (CO_2) concentration data from Mauna Loa in Hawaii, for the period March 1958 through December 2017. Representing the obvious time trend as a straight line yields the regression results presented in Table 7.4a, and the regression line is also plotted (dashed) in Figure 7.11. The results indicate a strong time trend, with the calculated standard error for the slope being much smaller than the estimated slope. The intercept merely estimates the CO_2 concentration at $t = 0$, or February 1958, so the implied test for

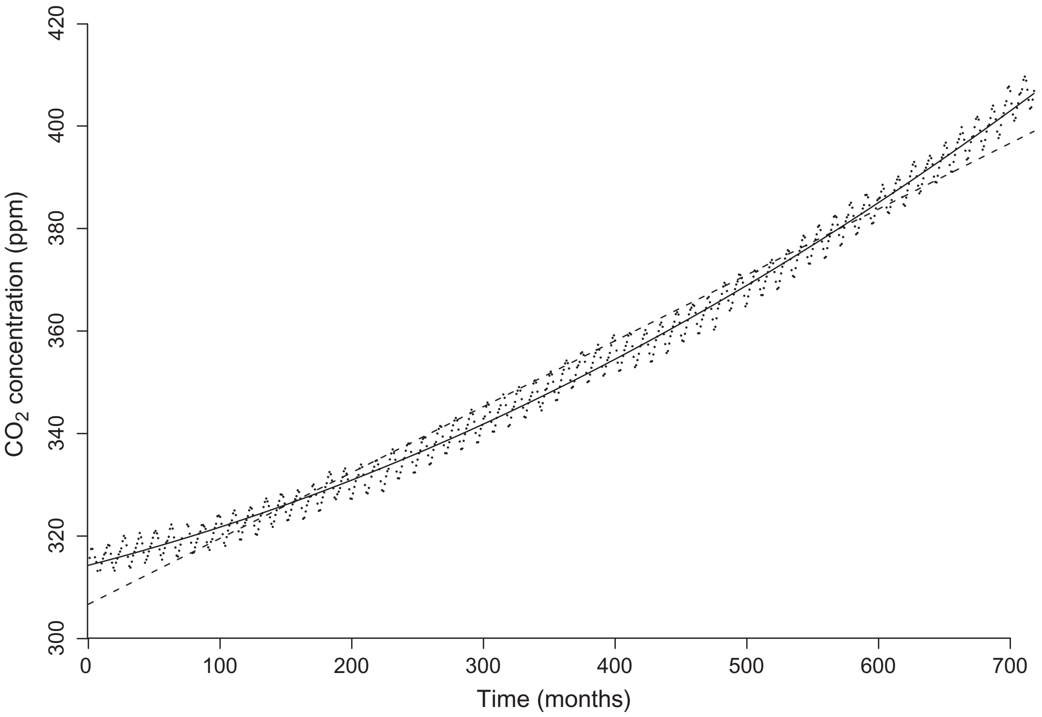


FIGURE 7.11 The Keeling Mauna Loa monthly CO_2 concentration data (March 1958–December 2017), with linear (dashed) and quadratic (solid) least-squares fits.

its difference from zero is of no interest. A literal interpretation of the MSE would suggest that a 95% prediction interval for measured CO_2 concentrations around the regression line would be about $\pm 2\sqrt{\text{MSE}} = \pm 8$ ppm.

However, examination of a plot of the residuals versus time for this linear regression would reveal a bowing pattern similar to that in Figure 7.6d, with a tendency for positive residuals at the beginning and end of the record, and with negative residuals being more common in the central part of the record. This can be discerned from Figure 7.11 by noticing that most of the points fall above the dashed line early and late in the record, and fall below the line toward the middle.

This problem with the residuals can be alleviated (and the regression consequently improved) by fitting a quadratic curve to the time trend. To do this, a second predictor is added to the regression, and that predictor is simply the square of the time variable. That is, a multiple regression with $K = 2$ is fit using the predictors $x_1 = t$ and $x_2 = t^2$. Once defined, x_2 is just another predictor variable, taking on values between 1^2 and $718^2 = 515{,}524$. The resulting least-squares quadratic regression is shown by the solid curve in Figure 7.11, and the corresponding regression statistics are summarized in Table 7.4b.

Of course the SST in Tables 7.4a and 7.4b are the same since both pertain to the same predictand, the CO_2 concentrations. For the quadratic regression, both the coefficients $b_1 = 0.0657$ and $b_2 = 0.0000870$ are substantially larger than their respective standard errors. The value of $b_0 = 314.3$ is again just the estimate of the CO_2 concentration at $t = 0$, and judging from the scatterplot this intercept is a better estimate of its true value than was obtained from the simple linear regression. The data points are fairly evenly scattered around the quadratic trend line throughout the time period, so the residual plot would exhibit the desired horizontal banding. Using this analysis, an approximate 95% prediction interval of $\pm 2\sqrt{\text{MSE}} = \pm 4.4$ ppm for CO_2 concentrations around the quadratic regression would be inferred throughout the range of these data.

The quadratic function of time provides a reasonable approximation of the annual-average CO_2 concentration for the 60 years represented by the regression, although we can find periods of time where the center of the point cloud wanders away from the curve. More importantly, however, a close inspection of the data points in Figure 7.11 reveals that they are not scattered randomly around the quadratic time trend. Rather, they execute a regular, nearly sinusoidal variation around the quadratic curve that is evidently an annual cycle. The resulting serial correlation in the residuals can easily be detected using the Durbin-Watson statistic, $d = 0.313$ (compare Figure 7.8). The CO_2 concentrations are lower in late summer and higher in late winter as a consequence of the annual cycle of photosynthetic carbon uptake by northern hemisphere land plants and carbon release from the decomposing dead plant parts. As will be shown in Section 10.4.2, this regular 12-month variation can be represented by introducing two more derived predictor variables into the equation, $x_3 = \cos(2\pi t/12)$ and $x_4 = \sin(2\pi t/12)$. Notice that both of these derived variables are functions only of the time variable t.

Table 7.4c indicates that, together with the linear and quadratic predictors included previously, these two harmonic predictors produce a very close fit to the data. The resulting prediction equation is

$$[CO_2] = \underset{(.1085)}{314.3} + \underset{(.0007)}{0.0657}\,t + \underset{(.0000)}{.0000872}\,t^2 + \underset{(.0503)}{1.149}\cos\left(\frac{2\pi t}{12}\right) + \underset{(.0502)}{2.582}\sin\left(\frac{2\pi t}{12}\right), \tag{7.27}$$

with all regression coefficients being much larger than their respective standard errors. The near equality of SST and SSR in Table 7.4c indicates that the predicted values are nearly coincident with the observed CO_2 concentrations (compare Equations 7.13 and 7.14a). The resulting coefficient of determination is $R^2 = 511{,}723/512356 = 99.88\%$, and the approximate 95% prediction interval implied by $\pm 2\sqrt{\text{MSE}}$ is only ± 1.9 ppm. A graph of Equation 7.27 would wiggle up and down around the solid curve in Figure 7.11, passing rather close to each of the data points. ◇

TABLE 7.4 ANOVA Tables and Regression Summaries for Three Regressions Fit to the 1958–2017 Keeling CO_2 Data in Figure 7.11

Source	df	SS	MS	F
(a) *Linear fit*				
Total	710	512,356		
Regression	1	501,014	501,014	31,317
Residual	709	11,342	15.998	
Variable	Coefficient	s.e.	t-ratio	
Constant	306.7	0.3027	1013	
t	0.1285	0.0007	177.0	
(b) *Quadratic fit*				
Total	710	512,356		
Regression	2	508,885	254,443	51,895
Residual	708	3471	4.903	
Variable	Coefficient	s.e.	t-ratio	
Constant	314.3	0.2536	1240	
t	0.0657	0.0016	40.6	
t^2	0.0000870	0.0000	40.1	
(c) *Including quadratic trend, and harmonic terms to represent the annual cycle*				
Total	710	512,356		
Regression	4	511,723	127,931	142,631
Residual	706	633.2	0.8969	
Variable	Coefficient	s.e.	t-ratio	
Constant	314.3	0.1085	2898	
t	0.0657	0.0007	94.9	
t^2	0.0000872	0.0000	93.8	
$\cos(2\pi t/12)$	1.149	0.0503	22.9	
$\sin(2\pi t/12)$	2.582	0.0502	51.4	

The variable t (time) is a consecutive numbering of the months, with March 1958 = 1 and December 2017 = 718. There are $n = 711$ data points and 7 missing months.

7.4. PREDICTOR SELECTION IN MULTIPLE REGRESSION

7.4.1. Why is Careful Predictor Selection Important?

There are almost always more potential predictors available than can be used in a statistical prediction procedure, and finding good subsets of these in particular cases is more difficult than might at first be imagined. The process is definitely not as simple as adding members of a list of potential predictors until an apparently good relationship is achieved. Perhaps surprisingly, there are dangers associated with including too many predictor variables in a forecast equation.

Example 7.4. An Overfit Regression

To illustrate the dangers of too many predictors, Table 7.5 presents total winter snowfall at Ithaca (inches) for the seven winters beginning in 1980 through 1986 and four arbitrary potential predictors: the U.S. federal deficit (in billions of dollars), the number of personnel in the U.S. Air Force, the sheep population of the United States (in thousands), and the average Scholastic Aptitude Test (SAT) scores of college-bound high-school students. Obviously these are nonsense predictors, which bear no real relationship to the snowfall at Ithaca.

Regardless of their lack of relevance, we can blindly offer these predictors to a computer regression package, and it will produce a regression equation. For reasons that will be made clear shortly, assume that the regression will be fit using only the six winters beginning in 1980 through 1985. That portion of available data used to produce the forecast equation is known as the *developmental sample*, *dependent sample*, or *training sample*. For the developmental sample of 1980–1985, the resulting equation is

$$Snow = 1161771 - 601.7\,yr - 1.733\,deficit + 0.0567\,AF\,pers. - 0.3799\,sheep + 2.882\,SAT$$

The ANOVA table accompanying this equation (not reproduced here) indicated MSE = 0.0000, $R^2 = 100.00\%$, and $F = \infty$, that is, a perfect fit!

Figure 7.12 shows a plot of the regression-specified snowfall totals (line segments) and the observed data (circles). For the developmental portion of the record, the regression does indeed represent the data exactly, as indicated by the ANOVA statistics, even though it is obvious from the nature of the predictor variables that the specified relationship is not physically meaningful. In fact, essentially any five predictors would have produced exactly the same perfect fit (although with different regression coefficients, b_k) to the six developmental data points. More generally, any $K = n - 1$ predictors will produce a perfect regression fit to any predictand for which there are n observations. This concept is easiest to see for the case of $n = 2$, where a straight line can be fit using any $K = 1$ predictor (simple linear regression), since a line can be found that will pass through any two points in the plane, and only an intercept and a slope are necessary to define a line. The problem, however, generalizes to any sample size.

TABLE 7.5 A small Data Set Illustrating the Dangers of Overfitting

Winter Beginning	Ithaca Snowfall (in.)	U.S. Federal Deficit ($\$ \times 10^9$)	U.S. Air Force Personnel	U.S. Sheep ($\times 10^3$)	Average SAT Scores
1980	52.3	59.6	557,969	12,699	992
1981	64.9	57.9	570,302	12,947	994
1982	50.2	110.6	582,845	12,997	989
1983	74.2	196.4	592,044	12,140	963
1984	49.5	175.3	597,125	11,487	965
1985	64.7	211.9	601,515	10,443	977
1986	65.6	220.7	606,500	9932	1001

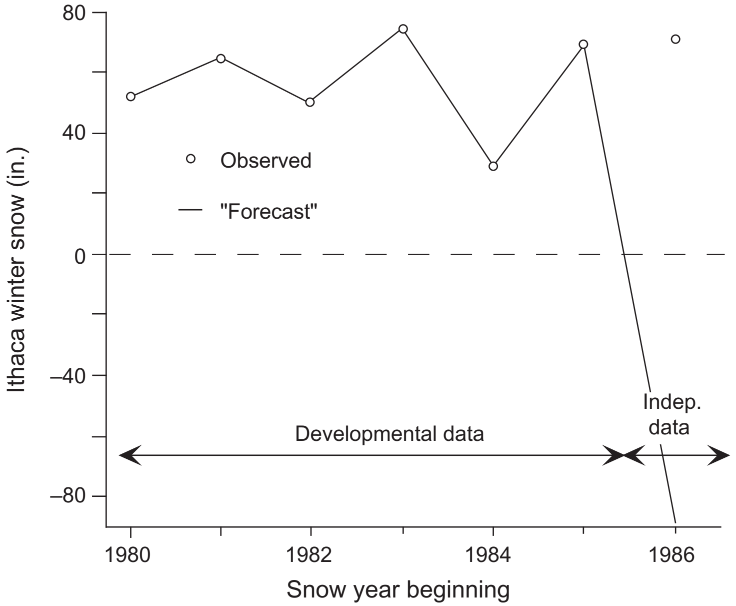


FIGURE 7.12 Forecasting Ithaca winter snowfall using the data in Table 7.5. The number of predictors is one fewer than the number of observations of the predictand in the developmental data, yielding perfect correspondence between the values specified by the regression and the predictand data for this portion of the record. The relationship falls apart completely when used with the 1986 data, which was not used in equation development. The regression equation has been grossly overfit.

Figure 7.12 indicates that the equation performs very poorly outside of the training sample, producing a meaningless forecast for negative snowfall during 1986–1987. Clearly, issuing forecasts equal to the climatological average total snowfall, or the snowfall for the previous winter, would yield better results than this overfit regression equation. ◇

Example 7.4 illustrates an extreme case of *overfitting* the data. Silver (2012) characterizes overfitting as "providing an overly specific solution to a general problem," and "the name given to the act of mistaking noise for a signal." That is, so many predictors have been used that an excellent fit has been achieved on the dependent data, but the fitted relationship falls apart when used with independent, or *verification data*—data not used in the development of the equation. In Example 7.4 the data for 1986 were reserved as a verification sample. Note that the problem of overfitting is *not* limited to cases where nonsense predictors are used in a forecast equation and will be a problem when too many meaningful predictors are included as well.

As ridiculous as it may seem, several important lessons can be drawn from Example 7.4:

- Begin development of a regression equation by choosing only physically reasonable or meaningful potential predictors. If the predictand of interest is surface temperature, for example, then temperature-related predictors such as the 1000–700 mb thickness (reflecting the mean virtual temperature in the layer), the 700 mb relative humidity (perhaps as a proxy for clouds), or the climatological average temperature for the forecast date (as a representation of the annual cycle of temperature) could be sensible candidate predictors. Understanding that clouds will form only in saturated air, a binary variable based on the 700 mb relative humidity also might be expected to contribute meaningfully to the regression. One consequence of this lesson is that a statistically literate person with insight into the physical problem ("domain expertise") may be more successful than a statistician at devising a forecast equation.

- A tentative regression equation needs to be tested on a sample of data not involved in its development. One way to approach this important step is simply to reserve a portion (perhaps a quarter, a third, or half) of the available data as the independent verification set and fit the regression using the remainder as the training set. The performance of the resulting equation will nearly always be better for the dependent than the independent data, since (in the case of least-squares regression) the coefficients have been chosen specifically to minimize the squared residuals in the developmental sample. A very large difference in performance between the dependent and independent samples would lead to the suspicion that the equation had been overfit. Often sufficient data are not available to reserve a separate verification set, in which case methods based on cross-validation (Section 7.4.4) are typically employed.
- A reasonably large developmental sample is needed if the resulting equation is to be "stable." Stability is usually understood to mean that the fitted coefficients are also applicable to independent (in particular, future) data, so that the resulting regression would be substantially unchanged if based on a different sample of the same kind of data. Stability thus relates to the precision of estimation of the regression coefficients, with smaller standard errors (Equations 7.18b and 7.19b, or square roots of the diagonal elements of Equation 11.40) being preferred. The number of coefficients that can be estimated with reasonable accuracy increases as the sample size increases, although in weather forecasting practice it has been found that there is little to be gained from including more than about a dozen predictor variables in a final regression equation (Glahn, 1985). In that kind of forecasting application there may be thousands of observations of the predictand in the developmental sample. Unfortunately, there is not a firm rule specifying a minimum ratio of sample size (number of observations of the predictand) to the number of predictor variables that can be supported in a final equation. Rather, testing on an independent data set is relied upon in practice to ensure stability of the regression.

7.4.2. Screening Predictors

Suppose the set of potential predictor variables for a particular problem could be assembled in a way that all physically relevant predictors were included, with exclusion of all irrelevant ones. This ideal can rarely, if ever, be achieved. Even if it could be, however, it generally would not be useful to include all the potential predictors in a final equation. This is because the predictor variables are almost always mutually correlated, so that the full set of potential predictors contains redundant information. Table 3.5, for example, shows substantial correlations among the six variables in Table A.1. Inclusion of predictors with strong mutual correlation is worse than superfluous, because this condition leads to poor estimates (high-variance sampling distributions) for the regression parameters. As a practical matter, then, we need a method to choose among potential predictors, and of deciding how many and which of them are sufficient to produce a good prediction equation.

In the jargon of statistical weather forecasting, the problem of selecting a good set of predictors from a pool of potential predictors is called *screening regression*, since the potential predictors must be subjected to some kind of screening, or filtering procedure. The most commonly used screening procedure is known as *forward selection* or *stepwise regression* in the broader statistical literature.

Given some number, M, candidate potential predictors for a least-squares linear regression, we begin the process of forward selection with the uninformative prediction equation $\hat{y} = b_0$. That is, only the intercept term is “in the equation,” and this intercept is necessarily equal to the sample mean of the predictand. On the first forward selection step, all M potential predictors are examined for the strength of their linear relationship to the predictand. In effect, all the possible M simple linear regressions between the available predictors and the predictand are computed, and that predictor whose linear regression is best among all candidate predictors is chosen as x_1. At this stage of the screening procedure, then, the prediction equation is $\hat{y} = b_0 + b_1 x_1$. Note that in general the intercept b_0 no longer will be the average of the y values.

At the next stage of the forward selection, trial regressions are again constructed using all remaining $M - 1$ predictors. However, all these trial regressions also contain the variable selected on the previous step as x_1. That is, given the particular x_1 chosen on the previous step, that predictor variable yielding the best regression $\hat{y} = b_0 + b_1 x_1 + b_2 x_2$ is chosen as x_2. This new x_2 will be recognized as best because it produces that regression equation with $K = 2$ predictors that also includes the previously chosen x_1, having the highest R^2, the smallest MSE, and the largest F ratio.

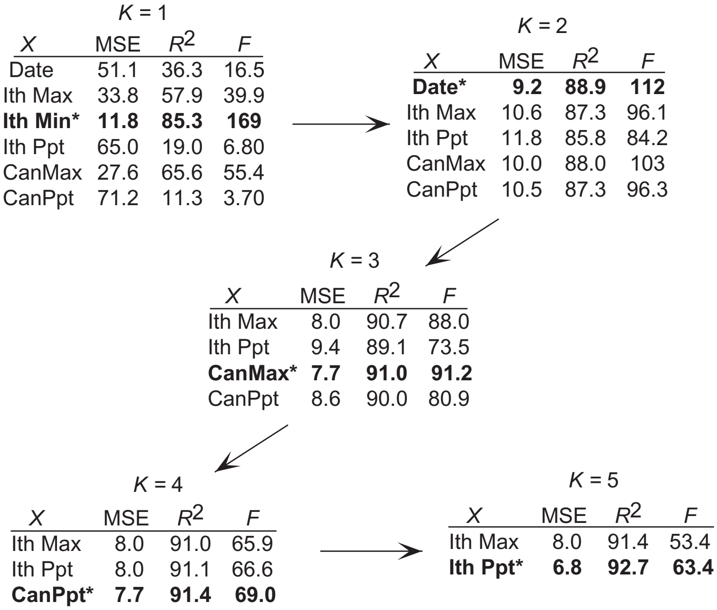
Subsequent steps in the forward selection procedure follow this pattern exactly: at each step, that member of the potential predictor pool not yet in the regression is chosen that produces the best regression in conjunction with the K–1 predictors chosen on previous steps. In general, when these regression equations are recomputed the regression coefficients for the intercept and for the previously chosen predictors will change. These changes will occur because the predictors usually are correlated to a greater or lesser degree, so that information about the predictand is spread among the predictors differently as more predictors are added to the equation.

Example 7.5. Equation Development Using Forward Selection

The concept of predictor selection can be illustrated with the January 1987 temperature and precipitation data in Table A.1. As in Example 7.1 for simple linear regression, the predictand is Canandaigua minimum temperature. The potential predictor pool consists of maximum and minimum temperatures at Ithaca, maximum temperature at Canandaigua, the logarithms of the precipitation amounts plus 0.01 in. (in order for the logarithm to be defined for zero precipitation) for both locations, and the day of the month. The date predictor is included on the basis of the trend in the residuals apparent in Figure 7.9. Note that this example is somewhat artificial with respect to statistical weather forecasting, since in general the predictors (other than the date) will not be known in advance of the time that the predictand (minimum temperature at Canandaigua) will be observed. However, this small data set serves to illustrate the principles.

Figure 7.13 diagrams the process of choosing predictors using forward selection. The numbers in each table summarize the comparisons being made at each step. For the first ($K = 1$) step, no predictors are yet in the equation, and all six potential predictors are under consideration. At this stage the predictor producing the best simple linear regression is chosen, as indicated by the smallest MSE, and the largest R^2 and F ratio among the six. This best predictor is the Ithaca minimum temperature, so the tentative regression equation is exactly Equation 7.21.

Having chosen the Ithaca minimum temperature in the first stage there are five potential predictors remaining, and these are listed in the $K = 2$ table. Of these five, the one producing the best predictions in an equation that also includes the Ithaca minimum temperature is chosen. Summary statistics for these five possible two-predictor regressions are also shown in the $K = 2$ table. Of these, the equation including Ithaca minimum temperature and the date as the two predictors is clearly best, producing MSE $= 9.2°\mathrm{F}^2$ for the dependent data.



K = 1

X	MSE	R^2	*F*
Date	51.1	36.3	16.5
Ith Max	33.8	57.9	39.9
Ith Min*	**11.8**	**85.3**	**169**
Ith Ppt	65.0	19.0	6.80
CanMax	27.6	65.6	55.4
CanPpt	71.2	11.3	3.70

K = 2

X	MSE	R^2	*F*
Date*	**9.2**	**88.9**	**112**
Ith Max	10.6	87.3	96.1
Ith Ppt	11.8	85.8	84.2
CanMax	10.0	88.0	103
CanPpt	10.5	87.3	96.3

K = 3

X	MSE	R^2	*F*
Ith Max	8.0	90.7	88.0
Ith Ppt	9.4	89.1	73.5
CanMax*	**7.7**	**91.0**	**91.2**
CanPpt	8.6	90.0	80.9

K = 4

X	MSE	R^2	*F*
Ith Max	8.0	91.0	65.9
Ith Ppt	8.0	91.1	66.6
CanPpt*	**7.7**	**91.4**	**69.0**

K = 5

X	MSE	R^2	*F*
Ith Max	8.0	91.4	53.4
Ith Ppt*	**6.8**	**92.7**	**63.4**

FIGURE 7.13 Diagram of the forward selection procedure for development of a regression equation for Canandaigua minimum temperature using as potential predictors the remaining variables in data set A.1, plus the date. At each step the variable is chosen (bold, starred) whose addition would produce the largest decrease in MSE or, equivalently, the largest increase in R^2 or F. At the final ($K = 6$) stage, only Ith. Max remains to be chosen, and its inclusion would produce MSE = 6.8, $R^2 = 93.0\%$, and $F = 52.8$.

With these two predictors now in the equation, there are only four potential predictors left at the $K = 3$ stage. Of these, the Canandaigua maximum temperature produces the best predictions in conjunction with the two predictors already in the equation, yielding MSE = 7.7°F^2 for the dependent data. Similarly, the best predictor at the $K = 4$ stage is Canandaigua precipitation, and the better predictor at the $K = 5$ stage is Ithaca precipitation. For $K = 6$ (all predictors in the equation) the MSE for the dependent data is 6.8°F^2, with $R^2 = 93.0\%$. ◇

An approach called *backward elimination* is an alternative to forward selection. The process of backward elimination is analogous but opposite to that of forward selection. Here the initial stage is a regression containing all M potential predictors, $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_M x_M$, so backward elimination will not be computationally feasible if $M \geq n$. Usually this initial equation will be grossly overfit, containing many redundant and some possibly useless predictors. At each step of the backward elimination procedure, the least important predictor variable is removed from the regression equation. That variable will be the one whose coefficient is smallest in absolute value, relative to its estimated standard error, so that in terms of the sample regression output tables presented earlier, the removed variable will exhibit the smallest (absolute) t ratio. As in forward selection, the regression coefficients for the remaining variables require recomputation if (as is usually the case) the predictors are mutually correlated.

The processes of both forward selection and backward elimination must be stopped at some intermediate stage, as discussed in the subsequent two sections. However, there is no guarantee that forward selection and backward elimination will choose the same subset of the potential predictor pool for the final regression equation. Other predictor selection procedures for multiple regression also exist, and these might select still different subsets. The possibility that a chosen selection procedure might not select the “right” set of predictor variables might be unsettling at first, but as a practical matter this is not usually an important problem in the context of producing an equation for use as a forecast tool. Correlations among the predictor variables result in the situation that nearly the same information about the predictand can be extracted from different subsets of the potential predictors. Therefore if the aim of the regression analysis is only to produce reasonably accurate forecasts of the predictand, the black box approach of empirically choosing a workable set of predictors is quite adequate. However, we should not be so complacent in a research setting, where one aim of a regression analysis could be to find specific predictor variables most directly responsible for the physical phenomena associated with the predictand.

7.4.3. Stopping Rules

Both forward selection and backward elimination require a stopping criterion or stopping rule. Without such a rule, forward selection would continue until all M candidate predictor variables were included in the regression equation, and backward elimination would continue until all predictors had been eliminated. It might seem that finding the stopping point would be a simple matter of evaluating the test statistics for the regression parameters and their nominal p values as supplied by the computer regression package. Unfortunately, because of the way the predictors are selected, these implied hypothesis tests are not quantitatively applicable, for two reasons. First, a sequence of tests is being computed, and the nominal p values do not account for this multiple testing. In addition, at each step (either in selection or elimination) predictor variables are not chosen randomly for entry or removal. Rather, the best or worst, respectively, among the available choices is selected.

The problem is illustrated in Figure 7.14, taken from the study of Neumann et al. (1977). The specific problem represented in this figure is the selection of exactly $K = 12$ predictor variables from pools of

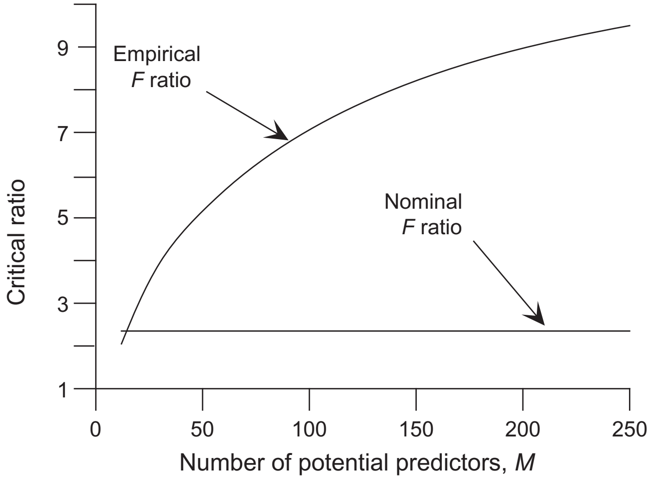


FIGURE 7.14 Comparison of the nominal and empirically (resampling-) estimated critical ($p = 0.01$) F ratios for overall significance in a particular regression problem, as a function of the number of potential predictor variables, M. The sample size is $n = 127$, with the best $K = 12$ predictor variables to be included in each final regression equation. The nominal F ratio of 2.35 is applicable only for the case of $M = K$. When the forward selection procedure can choose from among more than K potential predictors the true critical F ratio is substantially higher. The difference between the nominal and actual values widens as M increases. *From Neumann et al. (1977). © American Meteorological Society. Used with permission.*

potential predictors of varying sizes, M, when there are $n = 127$ observations of the predictand. Ignoring the problems of multiple testing and nonrandom predictor selection would lead us to declare as significant any regression for which the F ratio in the ANOVA table is larger than the nominal critical value of 2.35. Naïvely, this value would correspond to the minimum F ratio necessary to reject the null hypothesis of no real relationship between the predictand and the twelve predictors, at the 1% level. The curve labeled empirical F ratio was arrived at using a resampling test, in which the same meteorological predictor variables were used in a forward selection procedure to predict 100 artificial data sets of $n = 127$ independent Gaussian random numbers each. This procedure simulates a situation consistent with the null hypothesis that the predictors bear no real relationship to the predictand, while automatically preserving the correlations among this particular set of predictors.

Figure 7.14 indicates that the nominal regression diagnostics give the correct answer only in the case of $K = M$, for which there is no ambiguity in the predictor selection since all the $M = 12$ potential predictors must be used to construct the $K = 12$ predictor equation, and only a single test is being computed. When the forward selection procedure has available some larger number $M > K$ potential predictor variables to choose from, the true critical F ratio is higher, and sometimes by a substantial amount. Even though none of the potential predictors in the resampling procedure bears any real relationship to the artificial (random) predictand, the forward selection procedure chooses those predictors exhibiting the highest chance correlations with the predictand, and these relationships result in apparently large F ratio statistics. Put another way, the p value associated with the nominal critical $F = 2.35$ is too large (less significant), by an amount that increases as more potential predictors are offered to the forward selection procedure. To emphasize the seriousness of the problem, the nominal F ratio in the situation of Figure 7.14 for the very stringent 0.01% level test is only about 3.7. The practical result of relying literally on the nominal critical F ratio is to allow more predictors into the final equation than are meaningful, with the danger that the regression will be overfit. The F ratio in Figure 7.14 is a single-number regression diagnostic convenient for illustrating the effects of overfitting, but these effects would be reflected in other aspects of the ANOVA table also. For example, most if not all of the nominal t ratios for the individual cherry-picked predictors when $M >> K$ would be larger than 2 in absolute value, incorrectly suggesting meaningful relationships with the (random) predictand.

Unfortunately, the results in Figure 7.14 apply only to the specific data set from which they were derived. In order to employ this approach to estimate the true critical F-ratio using resampling methods it would need to be repeated for each regression to be fit, since the statistical relationships among the potential predictor variables will be different in different data sets. In practice, other less rigorous stopping criteria usually are employed. For example, we might stop adding predictors in a forward selection when none of the remaining predictors would reduce the R^2 by a specified amount, perhaps 0.05%.

The stopping criterion can also be based on the MSE. This choice is intuitively appealing because, as the standard deviation of the residuals around the regression function, $\sqrt{\text{MSE}}$ directly reflects the anticipated precision of a regression. For example, if a regression equation were being developed to forecast surface temperature, little would be gained by adding more predictors if the MSE were already $0.01°\text{F}^2$, since this would indicate a $\pm 2s_e$ (i.e., approximately 95%) prediction interval around the forecast value of about $\pm 2\sqrt{0.01°\text{F}^2} = 0.2°\text{F}$. So long as the number of predictors K is substantially less than the sample size n, adding more predictor variables (even meaningless ones) will decrease the MSE for the developmental sample. This concept is illustrated schematically in Figure 7.15. Ideally the stopping criterion would be activated at the point where the MSE does not decline appreciably with the addition of more predictors, at perhaps $K = 12$ predictors in the hypothetical case shown in Figure 7.15.

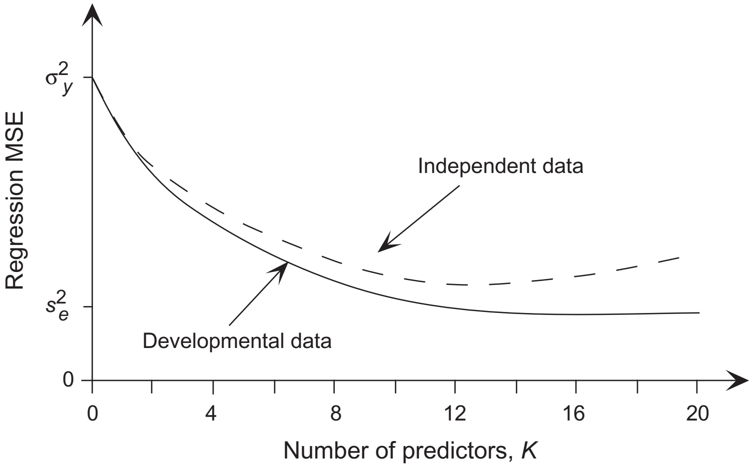


FIGURE 7.15 Schematic illustration of the regression MSE as a function of the number of predictor variables in the equation, K, for developmental data (solid), and for an independent verification set (dashed). *After Glahn (1985).*

Figure 7.15 indicates that the MSE for an independent data set will be larger than that achieved for the developmental data. This result should not be surprising, since the least-squares fitting procedure operates by optimizing the parameter values to minimize MSE for the developmental data. This underestimation of the independent-data MSE provided by the MSE for a forecast equation on developmental data is an expression of what is sometimes called *artificial skill* (Davis, 1976; Michaelson, 1987). The precise magnitudes of the differences in MSE between developmental and independent data sets are not determinable solely from the regression output using the developmental data. That is, having seen only the regressions fit to the developmental data, we cannot know the value of the minimum MSE for independent data. Neither can we know if it will occur at a similar point (at around $K = 12$ in Figure 7.15), or whether the equation has been overfit and the minimum MSE for the independent data will be for a substantially smaller K. This situation is unfortunate, because the purpose of developing a forecast equation is to specify future, unknown values of the predictand using observations of the predictors that have yet to occur.

Figure 7.15 also indicates that, for forecasting purposes, the exact stopping point is usually not critical as long as it is approximately right. Again, this is because the MSE tends to change relatively little through a range of K near the optimum, and for purposes of forecasting it is the minimization of the MSE rather than the specific identities of the predictors that is important. By contrast, if the purpose of the regression analysis is scientific understanding, the specific identities of chosen predictor variables can be of primary interest, and the magnitudes of the resulting regression coefficients may lead to significant physical insight. In this case it is not reduction of prediction MSE, per se, that is desired, but rather that causal relationships between particular variables be suggested by the analysis.

7.4.4. Cross-Validation

Sometimes regression equations to be used for weather forecasting are tested on a sample of independent data that has been held back during the development of the forecast equation. In this way, once the number K and specific identities of the predictors have been fixed, an estimate of the distances between the solid and dashed MSE lines in Figure 7.15 can be estimated directly from the reserved data. If the deterioration in forecast precision (i.e., the unavoidable increase in MSE) is judged to be acceptable, the equation can be used operationally.

This procedure of reserving an independent verification data set is actually a restricted case of a technique known as *cross-validation* (Efron and Gong, 1983; Efron and Tibshirani, 1993; Elsner and Schmertmann, 1994; Michaelson, 1987). Cross-validation simulates prediction for future, unknown data by repeating the entire fitting procedure on data subsets, and then examining the predictions made for the data portions left out of each of these subsets. The most frequently used procedure is known as *leave-one-out cross-validation*, in which the fitting procedure is repeated n times, each time with a sample of size n–1, because one of the predictand observations and its corresponding predictor set are left out in each replication of the fitting process. The result is n (often only slightly) different prediction equations.

In leave-one-out cross-validation, the estimate of the prediction MSE is computed by forecasting each of the omitted observation using the equation developed from the remaining n–1 data values, computing the squared difference between the prediction and predictand for each of these equations, and averaging the n squared differences. Thus leave-one-out cross-validation uses all n observations of the predictand to estimate the prediction MSE in a way that allows each observation to be treated, one at a time, as independent data.

More generally, J-fold cross-validation leaves out groups of size J sequentially, so that the fitting process is repeated n/J (approximately, unless this ratio is an integer) times, each with reduced sample size (approximately) $n-J$. Leave-one-out cross-validation thus corresponds to $J=1$. Hastie et al. (2009) suggest that choosing $J=5$ or $J=10$ is often a good rule of thumb. These choices produce data subsets that differ from each other more than in leave-one-out cross-validation, leading to a less variable, more stable, estimate of the MSE, and which also require substantially less computation. When the sample size n is small and the predictions will be evaluated using a correlation measure, leaving out $J>1$ values at a time can be especially advantageous (Barnston and van den Dool, 1993).

It should be emphasized that each repetition of the cross-validation exercise is a repetition of the entire fitting algorithm, not a refitting of the specific statistical model derived from the full data set using n–J data values. In particular, different prediction variables must be allowed to enter for different cross-validation subsets. DelSole and Shukla (2009) provide a cautionary analysis showing that failure to respect this precept can lead to random-number predictors exhibiting apparently real, cross-validated predictive ability. Any data transformations (e.g., standardizations with respect to climatological values) also need to be defined (and therefore possibly recomputed) without any reference to the withheld data in order for them to have no influence on the equation that will be used to predict them in the cross-validation exercise. However, the ultimate product equation, to be used for operational forecasts, would be fit using all the data after we are satisfied with the cross-validation results.

Cross-validation requires some special care when the data are serially correlated. In particular, data records adjacent to or near the omitted observation(s) will tend to be more similar to them than randomly selected ones, so the omitted observation(s) will be more easily predicted than the future observations they are meant to simulate. A solution to this problem is to leave out blocks of an odd number of consecutive observations, L, so the fitting procedure is repeated n–$L+1$ times on samples of size n–L (Burman et al., 1994; Elsner and Schmertmann, 1994). The blocklength L is chosen to be large enough for the correlation between its middle value and the nearest data used in the cross-validation fitting to be small, and the cross-validation prediction is made only for that middle value. For $L=1$ this moving-blocks cross-validation reduces to leave-one-out cross-validation.

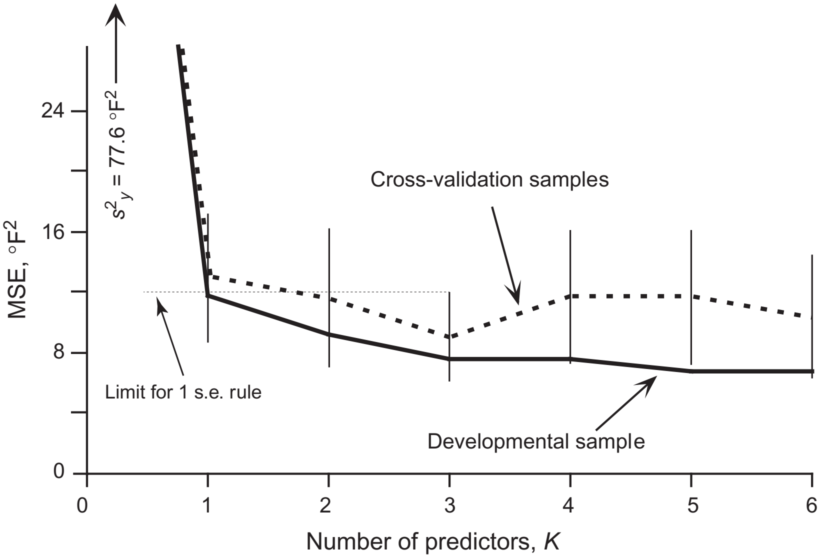


FIGURE 7.16 Plot of residual mean-squared error as a function of the number of regression predictors specifying Canandaigua minimum temperature, using the January 1987 data in Table A.1. Solid line shows MSE for developmental data (starred predictors in Figure 7.13). Dashed line shows estimated MSE achievable on independent data, with the same numbers of (possibly different) predictors, as computed through cross-validation, leaving out blocks of seven consecutive days. Whiskers show ±1 standard errors around the cross-validated MSE estimates, and the horizontal dotted line locates the one-s.e. rule threshold. This plot is a real-data example corresponding to the idealization in Figure 7.15.

Example 7.6. Protecting against Overfitting Using Cross-Validation

Having used all the available developmental data to fit the regressions in Example 7.5, what can be done to ensure that these prediction equations have not been overfit? Fundamentally, what is desired is a measure of how the regressions will perform when used on data not involved in the fitting. Cross-validation is an especially appropriate tool for this purpose in the present example, because the small ($n = 31$) sample would be inadequate if a substantial portion of it had to be reserved for a validation sample.

Figure 7.16 evaluates MSEs for six regression equations obtained with forward selection. This figure shows real results in the same form as the idealization of Figure 7.15. The solid line indicates the MSE achieved on the developmental sample, obtained by adding the predictors in the order shown in Figure 7.13. Because a regression chooses precisely those coefficients minimizing MSE for the developmental data, this quantity is expected to be higher when the equations are applied to independent data. An estimate of how much higher is given by the average MSE from the cross-validation samples (dashed line). Because these data are autocorrelated, a simple leave-one-out cross-validation is expected to underestimate the prediction MSE. Here the cross-validation has been carried out omitting blocks of length $L = 7$ consecutive days, and repeating the entire forward selection procedure $n–L + 1 = 25$ times. Since the lag-1 autocorrelation for the predictand is approximately $r_1 = 0.6$ and the autocorrelation function exhibits approximately exponential decay (similar to that in Figure 3.22), the correlation between the predictand in the centers of the seven-day moving blocks and the nearest data used for equation fitting is $0.6^4 = 0.13$, corresponding to $R^2 = 1.7\%$, indicating near-independence.

Each cross-validation point in Figure 7.16 represents the average of 25 squared differences between an observed value of the predictand at the center of a block, and the forecast of that value produced by a regression equation fit to all the data except those in that block. Predictors are added to each of these equations according to the usual forward selection algorithm. The order in which the predictors are added in one of these 25 regressions is often the same as that indicated in Figure 7.13 for the full data set, but this order is not forced onto the cross-validation samples, and indeed is different for some of the data partitions.

The differences between the dashed and solid lines in Figure 7.16 are indicative of the expected prediction errors for future independent data (dashed), and those that would be inferred from the MSE on the dependent data as provided by the ANOVA table (solid). The minimum cross-validation MSE at $K=3$ suggests that the best regression for these data may be the one with three predictors, and that it should produce prediction MSE on independent data of around 9.1°F^2, yielding $\pm 2s_e$ confidence limits of ±6.0°F. ◇

Recognizing that the cross-validated estimates of prediction MSE are subject to sampling variations, and that the cross-validated MSE estimates may be very similar near the minimum, Hastie et al. (2009) note that the possible resulting overfitting can be addressed using the so-called *one-s.e. rule*. The idea is to choose the simplest model (here, the regression with smallest K) whose accuracy is comparable to that of the best model, in the sense that its accuracy is no more than one standard error worse than that of the best model. The vertical whiskers in Figure 7.16 show ±1 s.e. around each of the average cross-validated MSE estimates, where the standard errors are computed as the standard deviation of the 25 cross-validated MSE estimates divided by $\sqrt{25}$ (i.e., the square root of Equation 5.4). The light dotted horizontal line in Figure 7.16 locates this threshold level. The cross-validated MSE for $K=1$ is above the one s.e. limit, but the value for $K=2$ is below it, so the two-predictor regression would be chosen by the one-s.e. rule in this case.

A potentially interesting but as yet little-used twist on cross-validation has been proposed by Hothorn et al. (2005). The idea is to repeatedly bootstrap the available data sample, and then use the data values not included in the current bootstrap sample (which will number approximately 0.368 n, on average) as the independent test data. Results over all bootstrap data splits are then aggregated to yield the final cross-validation estimate. Hastie et al. (2009) suggest a similar idea, where predictions for each data value are averaged over the bootstrap iterations that did not draw that sample, and these average results are then averaged over all underlying n data values.

Before leaving the topic of cross-validation it is worthwhile to note that the procedure is sometimes mistakenly referred to as the *jackknife*, a relatively simple resampling procedure that was introduced in Section 5.3.5. The confusion is understandable because the jackknife is computationally analogous to leave-one-out cross-validation. Its purpose, however, is to estimate the bias and/or standard deviation of a sampling distribution nonparametrically, and using only the data in a single sample. Given a sample of n independent observations, the idea in jackknifing is to recompute a statistic of interest n times, omitting a different one of the data values each time. Attributes of the sampling distribution for the statistic can then be inferred from the resulting n-member jackknife distribution (Efron, 1982; Efron and Tibshirani, 1993). The jackknife and leave-one-out cross-validation share the mechanics of repeated recomputation on reduced samples of size $n-1$, but cross-validation seeks to infer future forecasting performance, whereas the jackknife seeks to nonparametrically characterize the sampling distribution of a sample statistic.

7.5. REGULARIZATION/SHRINKAGE METHODS FOR MULTIPLE REGRESSION

One reason for the widespread use of the regression methods described in Sections 7.2 and 7.3 flows from the *Gauss-Markov theorem*, which shows that the least-squares parameter estimates b_k have the smallest variance (i.e., narrowest sampling distributions, and so are most consistent from batch to batch of training data) of any linear and unbiased estimates. (Unbiasedness implies that the mean of the sampling distribution locates the true value, on average.) However, in many instances the predictor variables in a multiple regression may exhibit some strong correlations among themselves, which leads to these sampling-distribution variances, although minimum conditional on the unbiasedness, being comparatively large and indeed large enough to negatively impact the accuracy of the predictions made with the resulting equations.

Although lack of bias sounds like a virtuous attribute because of the nonstatistical connotations of the word, better predictions may be possible when regression predictors are strongly correlated if the unbiasedness restriction is relaxed. As indicated by Equation 7.23, prediction accuracy is directly related to the regression MSE, which in turn depends on both the sampling variance and the bias:

$$\text{MSE} = \text{variance} + (\text{bias})^2. \tag{7.28}$$

For an unbiased procedure, MSE is minimized by minimizing variance. But more generally MSE results from a trade-off between variance and bias, so that nonzero bias can be associated with smaller MSE if a reduction in variance can be achieved that is greater than the squared bias. Figure 7.17 illustrates the trade-off in terms of the sampling distributions for a hypothetical regression parameter. The dashed PDF represents the sampling distribution for the unbiased estimate, which is centered on the true value, E[b]. A possible sampling distribution for a biased estimator of the same parameter is shown as the solid density. Even though the biased estimates here will be too small on average, they will typically be closer to the true value because of the reduced sampling variance, and so usually would exhibit the smaller MSE. Bias and sampling error in parameter estimation propagates to bias and random errors in the resulting forecasts.

Regularization is the general term for allowing bias, with the aim of reducing MSE. It may be useful in regression when the correlations among the predictors are sufficiently strong by yielding regression equations that produce more accurate predictions, and parameter estimates that can be more meaningfully interpreted. When regression predictors are strongly correlated among themselves, a large positive estimated coefficient for one predictor can be balanced by a large negative coefficient for another. The magnitude of

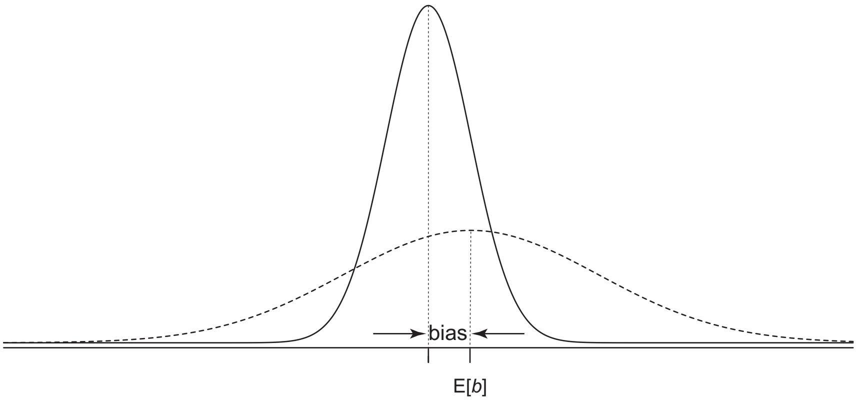


FIGURE 7.17 Illustration of the potential benefit of biased estimation, by reducing sampling variability sufficiently to yield a smaller MSE. Dashed density represents the sampling distribution of the unbiased estimator, and solid density represents sampling distribution of the biased estimator. Even though the biased estimator is too small on average, the estimation error is typically reduced.

such a trade-off between estimated coefficients for strongly correlated predictors can depend on the random sampling variations of the training data at hand, and so will not be reproducible in future independent data, leading to a lack of stability in the sense used in Section 7.4.1, and therefore to poor predictions. That is, small changes in the training data may lead to substantial changes in the estimated regression coefficients, which is symptomatic of large estimation variance. Another consequence can be that the estimated coefficients may not make sense physically, exhibiting the wrong sign relative to physical considerations, for example. In addition, the estimated regression coefficients tend to be too large in absolute value (Hoerl and Kennard, 1970), as is suggested by the breadth of the dashed probability density in Figure 7.17.

Regularization methods suppress these problems by allowing some bias in estimation of the regression parameters. In the process they also tend to reduce, or "shrink," the absolute magnitudes of the estimated coefficients. The interpretability of the resulting regression parameters may be improved in the sense that a relatively small subset of the predictors exhibiting the largest effects might be identified. Another advantage of regularization methods is that, unlike conventional least-squares regression, they can be applied when there are more potential predictors K than data values n.

The possible need for regularization methods in multiple regression can be diagnosed by looking at the squared multiple correlation coefficients for each of the predictor variables, in relation to all of the others. These will correspond to the R^2 value in a multiple regression in which the x variable of interest is predicted by the remaining K–1 predictors. Alternatively, these squared multiple correlations can be computed from the elements that appear on the diagonal of the inverse (Section 11.3.2) of the correlation matrix (Section 3.6.4) for the K predictor variables, according to $R_k^2 = 1 – 1/r_{kk}^{-1}$ for the kth element of the inverse correlation matrix. Marquardt (1970) suggests the rule of thumb that use of regularization in regression is appropriate if the largest of these squared multiple correlations is larger than 0.9 in absolute value, corresponding to a diagonal element in the inverse-correlation matrix larger than 10. Equivalently, the smallest eigenvalue (Section 11.3.3) of the correlation matrix for the x variables will be nearly zero.

7.5.1. Ridge Regression

Ridge regression, also known as *Tikhonov regularization*, is an early regularization procedure, dating from Hoerl and Kennard (1970). The basic idea is to control the potentially excessive magnitudes of estimated regression coefficients by imposing a limit, or budget, jointly on their magnitudes, according to

$$\sum_{k=1}^{K} b_k^{*2} \le c, \tag{7.29}$$

where the b_k* denote the regularized parameter estimates. In order to avoid undue emphasis on predictors that happen to exhibit large variance, perhaps arbitrarily because of the measurement scales of the corresponding predictors, the predictor variables in ridge regression are first standardized to have zero (sample) mean and unit variance according to Equation 3.27. That is, they are expressed as standardized anomalies. Often the predictand y is either centered (expressed as anomalies, with zero mean) or fully standardized, in which cases there is no intercept parameter b_0*, but in any case the magnitude of the intercept is not penalized in Equation 7.29.

The ridge regression parameter estimates are obtained by minimizing a penalized residual sum of squares,

$$SSE_p = \sum_{i=1}^{n} e_i^2 + \lambda \sum_{k=1}^{K} b_k^{*2}, \tag{7.30}$$

where λ is the nonnegative regularization (or biasing) parameter. The regression parameters can be obtained analytically through the modification to Equation 11.39

$$\boldsymbol{b}^* = \left([X]^T[X] + \lambda[I]\right)^{-1}[X]^T\boldsymbol{y}. \tag{7.31}$$

The parameter c in Equation 7.29 is a decreasing function of λ, but their particular functional relationship depends on the training data being used. Estimation bias increases with increasing λ but estimation variance decreases, so that an optimum value yields minimum MSE. Unless all the predictor variables are uncorrelated there is a nonzero value of λ for which the biased MSE* is smaller than its conventional unconstrained counterpart (Hoerl and Kennard, 1970).

It is conventional to display the results of ridge regression by plotting the estimated regression parameters b_k* as functions of the regularization parameter λ, which is known as the *ridge trace*. For sufficiently large λ the regression parameters will all have been forced to shrink to zero, which is generally a noninformative result. Although early practitioners would typically choose λ according to subjective rules of thumb (Hoerl and Kennard, 1970), the more modern approach is to optimize the regularization parameter using cross-validation (Section 7.4.4) to minimize the estimated prediction MSE, often in conjunction with the one-s.e. rule.

Example 7.7. Equation Development Using Ridge Regression

In Example 7.5, the use of forward selection was illustrated using the data in Appendix A.1, with the Canandaigua minimum temperature as the predictand. The remaining meteorological variables and the date were used as potential predictors, with the precipitation values transformed logarithmically. The largest of the six squared multiple correlation coefficients for these predictors is 0.94, exhibited by the Canandaigua maximum temperature, mainly because of its large correlation with the Ithaca maxima (Table 3.5). The squared multiple correlation for the Ithaca maxima is 0.93. Regularization may therefore be advantageous.

Figure 7.18 shows the ridge trace for the resulting regressions. For small values of the regularization parameter the estimated regression coefficients are nearly the same as for the unconstrained full model in Example 7.5. Although these coefficients have been computed using standardized variables, the scale on the vertical axis pertains to the variables on their original scales. As the regularization parameter is increased the regression coefficients are progressively shrunk toward zero. This effect is most pronounced for the two precipitation variables, which are strongly correlated with each other, and only modestly correlated with the predictand. In addition, the coefficients for the Canandaigua precipitation predictor changes sign to more physically plausible positive values as the regularization parameter increases.

A subjective interpretation of the ridge trace in Figure 7.18 would likely lead to removing the two precipitation predictors from the regression, both because they exhibit the greatest shrinkage, and because the coefficient for the Canandaigua precipitation changes sign, whereas the traces for the other predictors are nearly flat (e.g., Hoerl and Kennard, 1970; Marquardt and Snee, 1975). However, the procedure will not force estimated coefficients to zero unless the regularization parameter is very large. The dashed vertical line locates the regularized regression, with $\lambda = 3.96$, chosen by 10-fold cross-validation, using the one-s.e. rule. ◇

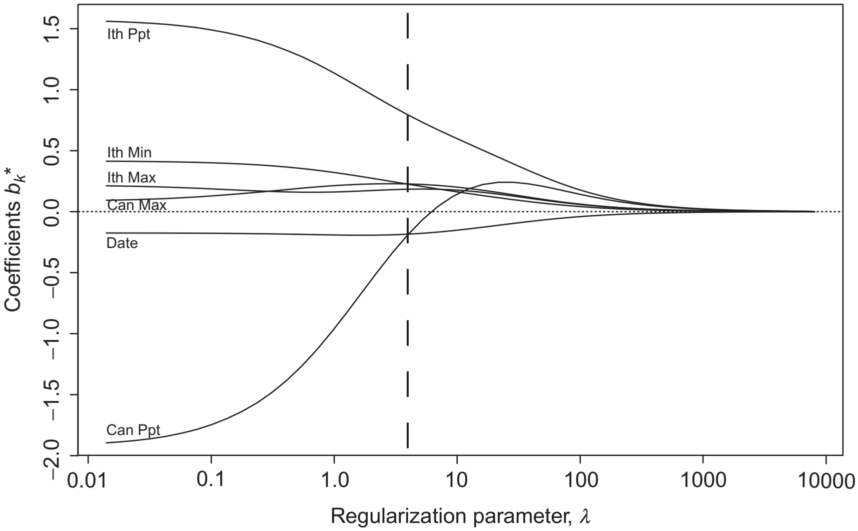


FIGURE 7.18 Ridge trace for regressions predicting Canandaigua minimum temperatures, using the other elements in Appendix A.1 as predictors. Regressions have been computed using standardized values, but the scale for the parameter estimates on the vertical axis pertains to the original dimensional values. Vertical dashed line locates the value of the regularization parameter chosen using 10-fold cross-validation with the one-s.e. rule.

7.5.2. The Lasso

The *Lasso* (least absolute shrinkage and selection operator, Tibshirani, 1996), or *L1 regularization*, is an alternative regularization procedure that constrains the magnitudes of the estimated regression coefficients according to

$$\sum_{k=1}^{K} \left|b_k^*\right| \le c. \tag{7.32}$$

Comparing to Equation 7.29 for ridge regression, the Lasso parameter magnitude budget is formulated in terms of the absolute values rather than the squares of the estimated coefficients. In common with ridge regression, standardized values of the predictors, and often the predictand as well, are used. The penalty in Equation 7.32 does not involve an intercept, if any.

Computation of Lasso regularization proceeds by minimizing

$$SSE_p = \sum_{i=1}^{n} e_i^2 + \lambda \sum_{k=1}^{K} \left|b_k^*\right|, \tag{7.33}$$

which is the counterpart of Equation 7.30 for ridge regression. As was also the case for ridge regression, the limit c of the parameter budget in Equation 7.32 is a decreasing function of λ, with the details of the relationship depending on the training data used. However, there is no closed-form expression for the parameter estimates as functions of the regularization parameter λ, corresponding to Equation 7.31 for ridge regression, and numerical methods must be used to evaluate them.

Lasso regularization is computed for a range of the regularization parameter λ, with the results displayed in graphical form using a *coefficient profile graph*, which is the counterpart of the ridge trace in ridge regression. The estimated regression parameters shrink toward zero as λ increases so that the ceiling c progressively constrains their magnitudes. Also in common with ridge regression, a best value for λ is usually chosen to minimize estimated prediction MSE using cross-validation, often in conjunction with the one-s.e. rule. However, unlike ridge regression, Lasso regularization will progressively force the regularized coefficient to exactly zero as the parameter λ is increased, so that it may lead automatically to a subset of "best" predictors.

Example 7.8. Equation Development using the Lasso

The regression situation described in Examples 7.5, 7.6, and 7.7, where the Canandaigua minimum temperatures were predicted using the other quantities in Appendix A.1, can also be addressed with Lasso regularization. As before, the precipitation predictors are log-transformed, and all quantities are standardized.

Figure 7.19 shows the coefficient profile graph for this analysis, where the vertical axis pertains to regularized regression coefficients relating to the original untransformed variables. Qualitatively, the coefficient profile graph is similar to the ridge trace in Figure 7.18 for the ridge regressions based on the same data. For very small values of λ the parameter estimates approach their conventional least-squares counterparts. They are shrunk toward zero with increasing λ, and similarities in shape in the two figures for the various predictors can be discerned. Most notably, the coefficients for the two precipitation predictors begin with opposite signs and large absolute values, but decrease quickly to zero.

A notable difference in Figure 7.19 is that the predictors are progressively shrunk to exactly zero as λ is increased. The dashed vertical line locates the value of $\lambda = 1.37$ that is chosen using 10-fold cross-validation with the one-s.e. rule. At this point the nonzero predictors are Ithaca minima, Canandaigua maxima, and the date. Because the Lasso forces predictors to zero, moving from right to left in Figure 7.19 in effect traces out the results of a forward selection algorithm. Unlike in the forward

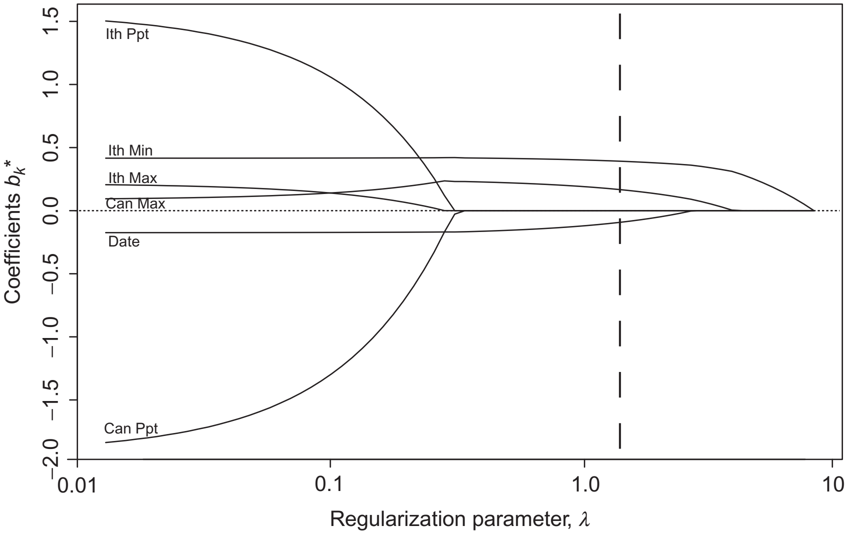


FIGURE 7.19 Coefficient profile graph for lasso regularization of regressions predicting Canandaigua minimum temperatures using the other elements of Appendix A.1 as predictors. Regressions have been computed using standardized values, but the scale for the parameter estimates on the vertical axis pertains to the original dimensional values. Vertical dashed line locates the value of the regularization parameter chosen using 10-fold cross-validation with the one-s.e. rule.

selection exercise in Example 7.5, here the Canandaigua maxima appear to be more important than the date predictor. ◇

7.6. NONLINEAR REGRESSION

7.6.1. Generalized Linear Models

Although linear least-squares regression accounts for the overwhelming majority of regression applications, it is also possible to fit regression functions that are nonlinear in the regression parameters. Nonlinear regression can be appropriate when a nonlinear relationship is dictated by nature of the physical problem at hand, and/or the usual assumptions of Gaussian residuals with constant variance are untenable. In these cases the fitting procedure is usually iterative and based on maximum likelihood methods (see Section 4.6).

This section introduces two such regression structures, both of which are important examples of a class of nonlinear statistical models known as *generalized linear models* (GLM) (McCullagh and Nelder, 1989). Generalized linear models extend linear statistical models, such as multiple linear regression, by representing the predictand as a nonlinear transformation of a linear regression function. The nonlinearity is represented by a 1-to-1 (and therefore invertible) function known as the *link function*, $g(\hat{y})$. Accordingly, the GLM extension of the ordinary linear multiple regression (Equation 7.25) is

$$g(\hat{y}) = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_K x_K, \tag{7.34}$$

where the specific form of the link function is chosen according to the nature of the predictand data. Comparing Equation 7.34 and 7.25 shows that ordinary linear regression is a special case of a GLM model, with the identity link, i.e., $g(\hat{y}) = \hat{y}$. Because the link function will be invertible, GLM equations are often written equivalently as

$$\hat{y} = g^{-1}(b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_K x_K). \tag{7.35}$$

7.6.2. Logistic Regression

One important advantage of statistical over (deterministic) dynamical forecasting methods is the capacity to produce probability forecasts. Inclusion of probability elements into the forecast format is advantageous because it provides an explicit expression of the inherent uncertainty or state of knowledge about the future weather, and because probabilistic forecasts allow users to extract more value from them when making decisions (e.g., Katz and Murphy, 1997a; Krzysztofowicz, 1983; Murphy, 1977; Thompson, 1962). In a sense, ordinary linear regression produces probability information about a predictand, for example, through the 95% prediction interval around the regression function approximated by the $\pm 2\sqrt{\text{MSE}}$ rule. More narrowly, however, probability forecasts are forecasts for which the predictand is expressed explicitly in terms of probability, rather than as the value of a physical variable.

The simplest setting for probability forecasting produces single probabilities for binary outcomes, for example, rain tomorrow or not. Systems for producing this kind of probability forecast are developed in a regression setting by first transforming the predictand to a binary (or dummy) variable, taking on the values zero (e.g., for no rain) and one (for any nonzero rain amount). That is, regression procedures are implemented after applying Equation 7.26 to the predictand, y, rather than to a predictor. In a sense, zero and one can be viewed as probabilities of the dichotomous event not occurring or occurring, respectively, after it has been observed.

The easiest approach to regression when the predictand is binary is to use the machinery of ordinary multiple regression as described previously. In the meteorological literature this is called Regression Estimation of Event Probabilities (REEP) (Glahn, 1985; Miller, 1964). The main justification for the use of REEP is that it is no more computationally demanding than the fitting of any other linear regression, and so historically has been used when computational resources have been limiting. The resulting predicted values are usually between zero and one, and it has been found through operational experience that these predicted values can usually be treated as specifications of probabilities for the event $\{Y=1\}$. However, one obvious problem with REEP is that some of the resulting forecasts may not lie on the unit interval, particularly when the regression relationship is relatively strong (so that the slope parameter is relatively large in absolute value, Brelsford and Jones, 1967), there are relatively few predictors (so there is limited scope for compensating influences among them), or the predictands are near the limits or outside of their ranges in the training data. This logical inconsistency usually causes little difficulty in an operational setting because multiple-regression forecast equations with many predictors rarely produce such nonsense probability estimates. When the problem does occur the forecast probability is usually near zero or one, and the operational forecast can be issued as such.

Two other difficulties associated with forcing a linear regression onto a problem with a binary predictand are that the residuals are clearly not Gaussian, and their variances are not constant. Because the predictand can take on only one of two values, a given regression residual can also take on only one of two values, and so the residual distributions are Bernoulli (i.e., binomial, Equation 4.1, with $N=1$). Furthermore, the variance of the residuals is not constant, but depends on the ith predicted probability p_i according to $(p_i)(1-p_i)$.

It is possible to simultaneously bound the regression estimates for binary predictands on the interval (0, 1), and to accommodate the Bernoulli distributions for the regression residuals, using a more theoretically satisfying technique known as log*istic regression*. Some examples of logistic regression in the atmospheric science literature are Applequist et al. (2002), Bröcker (2010), Hamill et al., 2004, Hilliker and Fritsch (1999), Lehmiller et al. (1997), and Lemcke and Kruizinga (1988).

Logistic regressions are fit to binary predictands using the log-odds, or *logit*, link function $g(p)=\ln[p/(1–p)]$ (Equation 3.24), yielding the generalized linear model

$$\ln\left(\frac{p}{1-p}\right)=b_0+b_1x_1+\cdots+b_Kx_K, \tag{7.36a}$$

which can be expressed also in the form of Equation 7.35 as

$$p=\frac{\exp(b_0+b_1x_1+\cdots+b_Kx_K)}{1+\exp(b_0+b_1x_1+\cdots+b_Kx_K)}=\frac{1}{1+\exp(-b_0-b_1x_1-\cdots-b_Kx_K)}. \tag{7.36b}$$

Here the predicted value p results from the ith set of predictors (x_1, x_2, …, x_K) of n such sets. Geometrically, logistic regression is most easily visualized for the single-predictor case ($K=1$), for which Equation 7.36b is an S-shaped curve that is a function of x_1. In the limit of $b_0+b_1x_1 \rightarrow +\infty$ the exponential function in the first equality of Equation 7.36b becomes arbitrarily large so that the predicted value p_i approaches one. As $b_0+b_1x_1 \rightarrow -\infty$, the exponential function approaches zero and thus so does the predicted value. Depending on the parameters b_0 and b_1, the function rises gradually or abruptly from zero to one (or falls, for $b_1 < 0$, from one to zero) at intermediate values of x_1. Thus it is guaranteed that logistic regression will produce properly bounded probability estimates. The logistic function is convenient mathematically, but it is not the only function that could be used in this context. Another alternative yielding a very similar shape involves using the inverse Gaussian CDF for the link function, yielding $p = \Phi(b_0 + b_1x_1 + \ldots + b_Kx_K)$, which is known as *probit regression*.

Equation 7.36a shows that logistic regression can be viewed as linear in terms of the logarithm of the odds ratio, $p/(1–p)$. Superficially it appears that Equation 7.36a could be fit using ordinary linear regression, except that the predictand is binary, so the left-hand side will be either ln(0) or ln(∞). However, fitting the regression parameters can be accomplished using the method of maximum likelihood, recognizing that the residuals are Bernoulli variables. Assuming that Equation 7.36 is a reasonable model for the smooth changes in the probability of the binary outcome as a function of the predictors, the probability distribution function for the *i*th residual is Equation 4.1, with $N = 1$, and p_i as specified by Equation 7.36b. The corresponding likelihood is of the same functional form, except that the values of the predictand *y* and the predictors *x* are fixed, and the probability p_i is the variable. If the *i*th residual corresponds to a "success" (i.e., the event occurs, so $y_i = 1$), the likelihood is $\Lambda = p_i$ (as specified in Equation 7.36b), and otherwise $\Lambda = 1–p_i = 1/(1+\exp[b_0 + b_1x_1 + \ldots + b_Kx_K])$. If the *n* sets of observations (predictand and predictor(s)) are independent, the joint likelihood for the $K+1$ regression parameters is simply the product of the *n* individual likelihoods, or

$$\Lambda(\boldsymbol{b}) = \prod_{i=1}^{n} \frac{y_i \exp(b_0 + b_1x_1 + \cdots + b_Kx_K) + (1 - y_i)}{1 + \exp(b_0 + b_1x_1 + \cdots + b_Kx_K)}. \tag{7.37}$$

Since the *y*'s are binary [0, 1] variables, each factor in Equation 7.37 for which $y_i = 1$ is equal to p_i (Equation 7.36b), and the factors for which $y_i = 0$ are equal to $1 - p_i$. As usual, it is more convenient to estimate the regression parameters by maximizing the log-likelihood

$$L(\boldsymbol{b}) = \ln[\Lambda(\boldsymbol{b})] = \sum_{i=1}^{n} \{y_i(b_0 + b_1x_1 + \cdots b_Kx_K) - \ln[1 + \exp(b_0 + b_1x_1 + \cdots b_Kx_K)]\}. \tag{7.38}$$

The combinatorial factor in Equation 4.1 has been omitted because it does not involve the unknown regression parameters, and so will not influence the process of locating the maximum of the function. Usually statistical software will be used to find the values of the *b*'s maximizing this function, using iterative methods such as those outlined in Sections 4.6.2 or 4.6.3.

Some software will display information relevant to the strength of the maximum likelihood fit using what is called the *analysis of deviance* table, which is analogous to the ANOVA table for linear regression (Table 7.3). The idea underlying an analysis of deviance table is the likelihood ratio test (Equation 5.20). As more predictors and thus more regression parameters are added to Equation 7.36, the log-likelihood will progressively increase as more latitude is provided to accommodate the data. Whether that increase is sufficiently large to reject the null hypothesis that a particular, smaller, regression equation is adequate, is judged in terms of twice the difference of the log-likelihoods relative to the χ^2 distribution, with degrees-of-freedom ν equal to the difference in numbers of parameters between the null-hypothesis regression and the more elaborate regression being considered. More about analysis of deviance can be learned from sources such as Healy (1988) or McCullagh and Nelder (1989).

The likelihood ratio test is appropriate when a single candidate logistic regression is being compared to a null model. Often H_0 will specify that all the regression parameters except b_0 are zero, in which case the question being addressed is whether the predictors $\boldsymbol{x}$ being considered are justified in favor of the constant (no-predictor) model with $b_0 = \ln\left[\Sigma y_i / n/(1 - \Sigma y_i/n)\right]$. However, if multiple alternative logistic regressions are being entertained, computing the likelihood ratio test for each alternative raises the problem of test multiplicity (see Section 5.4.1). In such cases it is better to compute either the *Bayesian Information Criterion* (BIC) statistic (Schwarz, 1978)

$$BIC = -2L(\boldsymbol{b}) + (K+1)\ \ln(n) \tag{7.39}$$

or the *Akaike Information Criterion* (AIC) (Akaike, 1974)

$$AIC = -2L(\boldsymbol{b}) + 2(K+1), \tag{7.40}$$

for each candidate model. Both the AIC and BIC statistics consist of twice the negative of the log-likelihood plus a penalty for the number of parameters fit, and the preferred regression will be the one minimizing the chosen criterion. The BIC statistic will generally be better for large-n problems since its probability of selecting the proper member of the class of models considered approaches 1 as n—>∞. For smaller sample sizes BIC often chooses models that are simpler than justified by the data, in which cases AIC may be preferred. The AIC is biased for small sample size relative to the number of parameters estimated, and so will tend to yield overfit models. In such cases the corrected AIC (Hurvich and Tsai, 1989)

$$AIC_C = -2L(\boldsymbol{b}) + 2(K+1) + \frac{2(K+1)(K+2)}{n-K-2}, \tag{7.41}$$

will generally be preferred, where again $K+1$ is the number of parameters (including intercept) estimated for a given model. Equation 7.41 converges to Equation 7.40 for $n >> K$.

The Lasso (Section 7.5.2) is another alternative for selecting among potential predictors in a logistic regression (Bröcker, 2010). In that case predictors whose regression coefficients have not been shrunk to zero when the regularization parameter has been optimized through cross-validation, possibly using the one-s.e. rule, would be retained.

Example 7.9. Comparison of REEP and Logistic Regression

Figure 7.20 compares the results of REEP (dashed) and logistic regression (solid) for some of the January 1987 data from Table A.1. The predictand is daily Ithaca precipitation, transformed to a binary variable using Equation 7.26 with $c = 0$. That is, $y = 0$ if the precipitation is zero, and $y = 1$ otherwise. The predictor is the Ithaca minimum temperature for the same day. The REEP (linear regression) equation has been fit using ordinary least squares, yielding $b_0 = 0.208$ and $b_1 = 0.0212$. This equation specifies negative probability of precipitation if the temperature predictor is less than about −9.8°F, and specifies probability of precipitation greater than one if the minimum temperature is greater than about 37.4°F. The parameters for the logistic regression, fit using maximum likelihood, are $b_0 = -1.76$ and $b_1 = 0.117$. The logistic regression curve produces probabilities that are similar to the REEP specifications through most of the temperature range, but are constrained by the functional form of Equation 7.36 to lie between zero and one, even for extreme values of the predictor.

Maximizing Equation 7.38 for logistic regression with a single ($K = 1$) predictor is simple enough that the Newton-Raphson method (see Section 4.6.2) can be implemented easily and is reasonably robust to poor initial guesses for the parameters. The counterpart to Equation 4.87 for this problem is

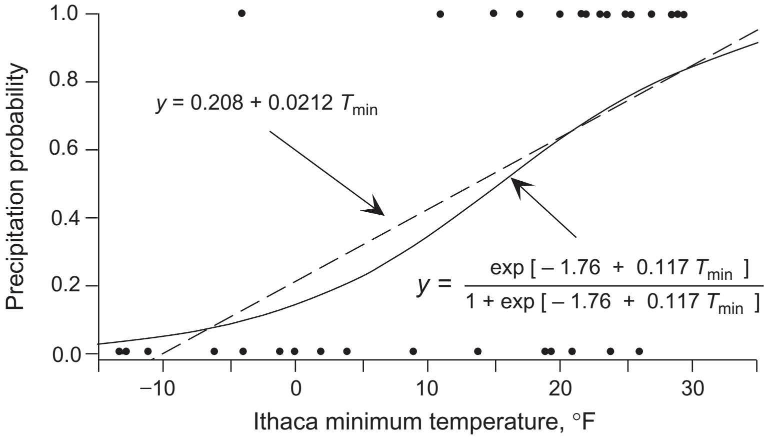


FIGURE 7.20 Comparison of regression probability forecasting using REEP (dashed) and logistic regression (solid) using the January 1987 data set in Table A.1. The linear function was fit using least squares, and the logistic curve was fit using maximum likelihood, to the data shown by the dots. The binary predictand $y = 1$ if Ithaca precipitation is greater than zero, and $y = 0$ otherwise.

$$\begin{bmatrix} b_0^* \\ b_1^* \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} - \begin{bmatrix} \sum_{i=1}^{n} (p_i^2 - p_i) & \sum_{i=1}^{n} x_i (p_i^2 - p_i) \\ \sum_{i=1}^{n} x_i (p_i^2 - p_i) & \sum_{i=1}^{n} x_i^2 (p_i^2 - p_i) \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{n} (y_i - p_i) \\ \sum_{i=1}^{n} x_i (y_i - p_i) \end{bmatrix}, \tag{7.42}$$

where p_i is a function of the regression parameters b_0 and b_1, and depends also on the predictor data x_i, as shown in Equation 7.36b. The first derivatives of the log-likelihood (Equation 7.38) with respect to b_0 and b_1 are in the vector enclosed by the rightmost square brackets, and the second derivatives are contained in the matrix to be inverted. Beginning with an initial guess for the parameters (b_0, b_1), updated parameters ($b^*{}_0$, $b^*{}_1$) are computed and then resubstituted into the right-hand side of Equation 7.42 for the next iteration. For example, assuming initially that the Ithaca minimum temperature is unrelated to the binary precipitation outcome, so $b_0 = -0.0645$ (the log of the observed odds ratio, for constant $p = 15/31$) and $b_1 = 0$; the updated parameters for the first iteration are $b^*{}_0 = -0.0645–(-0.251)(-0.000297)–(0.00936)(118.0) = -1.17$, and $b^*{}_1 = 0–(0.00936)(-0.000297)–(-0.000720)(118.0) = 0.085$. These updated parameters increase the log-likelihood from -21.47 for the constant model (calculated using Equation 7.38, imposing $b_0 = -0.0645$ and $b_1 = 0$), to -16.00. After four iterations the algorithm has converged, with a final (maximized) log-likelihood of -15.67.

Is the logistic relationship between Ithaca minimum temperature and the probability of precipitation statistically significant? This question can be addressed using the likelihood ratio test (Equation 5.20). The appropriate null hypothesis is that $b_1 = 0$, so $L(H_0) = -21.47$, and $L(H_A) = -15.67$ for the fitted regression. If H_0 is true then the observed test statistic $\Lambda^* = 2\,[L(H_A) - L(H_0)] = 11.6$ is a realization from the χ^2 distribution with $\nu = 1$ (the difference in the number of parameters between the two regressions), and the test is 1-tailed because small values of the test statistic are favorable to H_0. Referring to the first row of Table B.3, it is clear that the regression is significant at the 0.1% level.

Another approach to assessing the significance of this regression is through use of the estimated sampling distributions of the fitted coefficients, as outlined in Section 4.6.4. For sufficiently large sample sizes, these sampling distributions are Gaussian, with variances estimated by the diagonal elements of the information matrix, which in the present example is the negative of the inverted matrix in Equation 7.42. The estimated sampling variance for the b_1 coefficient will be in the lower-right corner of that matrix, which is 0.00190. Accordingly the t statistic addressing the null hypothesis that $b_1 = 0$ would be $0.117/\sqrt{0.00190} = 2.68$, yielding a one-tailed p value of 0.0037. ◇

7.6.3. Poisson Regression

Another regression setting where the residual distribution may be poorly represented by the Gaussian is the case where the predictand consists of counts, that is, each of the y's is a nonnegative integer. Particularly if these counts tend to be small, the residual distribution is likely to be asymmetric, and we would like a regression predicting these data to be incapable of implying nonzero probability for negative counts.

A natural probability model for count data is the Poisson distribution (Equation 4.12). Recall that one interpretation of a regression function is as the conditional mean of the predictand, given specific value (s) of the predictor(s). If the outcomes to be predicted by a regression are Poisson-distributed counts, but the Poisson parameter μ may depend on one or more predictor variables, we can structure a regression to specify the Poisson mean as a nonlinear function of those predictors using the link function $g(\mu) = \ln(\mu)$. The resulting GLM can then be written as

$$\ln(\mu) = b_0 + b_1 x_1 + \cdots + b_K x_K, \tag{7.43a}$$

or

$$\mu = \exp\left[b_0 + b_1 x_1 + \cdots + b_K x_K\right]. \tag{7.43b}$$

Equation 7.43 is not the only function that could be used for this purpose, but framing the problem in this way makes the subsequent mathematics quite tractable, and the logarithmic link function ensures that any predicted Poisson mean is nonnegative. Some applications of Poisson regression are described in Elsner and Schmertmann (1993), Paciorek et al. (2002), Parisi and Lund (2008), Solow and Moore (2000), and Tippet et al. (2011).

Having framed the regression in terms of Poisson distributions for the y_i conditional on the corresponding set of predictor variables $\boldsymbol{x}_i = \{x_1, x_2, \ldots, x_K\}$, the natural approach to parameter fitting is to maximize the Poisson log-likelihood, written in terms of the regression parameters. Again assuming independence among the n data values, the log-likelihood is

$$L(\boldsymbol{b}) = \sum_{i=1}^{n} \{y_i\,(b_0 + b_1 x_1 + \cdots + b_K x_K) - \exp\,(b_0 + b_1 x_1 + \cdots + b_K x_K)\}, \tag{7.44}$$

where the term involving $y!$ from the denominator of Equation 4.12 has been omitted because it does not involve the unknown regression parameters and so will not influence the process of locating the maximum of the function. An analytic maximization of Equation 7.44 in general is not possible, so that statistical software will approximate the maximum iteratively, typically using one of the methods outlined in Sections 4.6.2 or 4.6.3. For example, if there is a single ($K = 1$) predictor, the Newton-Raphson method (see Section 4.6.2) iterates the solution according to

$$\begin{bmatrix} b_0^* \\ b_1^* \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} - \begin{bmatrix} -\sum_{i=1}^{n} \mu_i & -\sum_{i=1}^{n} x_i \mu_i \\ -\sum_{i=1}^{n} x_i \mu_i & -\sum_{i=1}^{n} x_i^2 \mu_i \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{n} (y_i - \mu_i) \\ \sum_{i=1}^{n} x_i (y_i - \mu_i) \end{bmatrix}, \tag{7.45}$$

where μ_i is the conditional mean as a function of the ith set of regression parameters as defined in Equation 7.43b. Equation 7.45 is the counterpart of Equation 4.87 for fitting the gamma distribution, and Equation 7.42 for logistic regression.

Example 7.10. A Poisson Regression

Consider the annual counts of tornados reported in New York state for 1959–1988, in Table 7.6. Figure 7.21 shows a scatterplot of these as a function of average July temperatures at Ithaca in the corresponding years. The solid curve is a Poisson regression function, and the dashed line shows the ordinary linear least-squares linear fit. The nonlinearity of the Poisson regression is quite modest over the range of the training data, although the regression function would remain strictly positive regardless of the magnitude of the predictor variable.

The relationship is weak, but slightly negative. The significance of the Poisson regression usually would be judged using the likelihood ratio test (Equation 5.20). The maximized log-likelihood (Equation 7.44) is 74.26 for $K = 1$, whereas the log-likelihood with only the intercept $b_0 = \ln(\Sigma y/n) = 1.526$ is 72.60. Comparing $\Lambda^* = 2\,(74.26–72.60) = 3.32$ to the χ^2 distribution quantiles in Table B.3 with $\nu = 1$ (the difference in the number of fitted parameters) indicates that b_1 would be judged significantly different from zero at the 10% level, but not at the 5% level. Alternatively the estimated sampling variance, given by the lower-right element of the negative of the inverse matrix in Equation 7.45 (Section 4.6.4) is 0.00325, so that the t statistic for the null hypothesis that $b_1 = 0$ is $-0.104/\sqrt{0.00325} = -1.82$, corresponding to the 2-tailed p value 0.069. For the linear regression, the t ratio

TABLE 7.6 Numbers of Tornados Reported Annually in New York State, 1959–1988

1959	3	1969	7	1979	3
1960	4	1970	4	1980	4
1961	5	1971	5	1981	3
1962	1	1972	6	1982	3
1963	3	1973	6	1983	8
1964	1	1974	6	1984	6
1965	5	1975	3	1985	7
1966	1	1976	7	1986	9
1967	2	1977	5	1987	6
1968	2	1978	8	1988	5

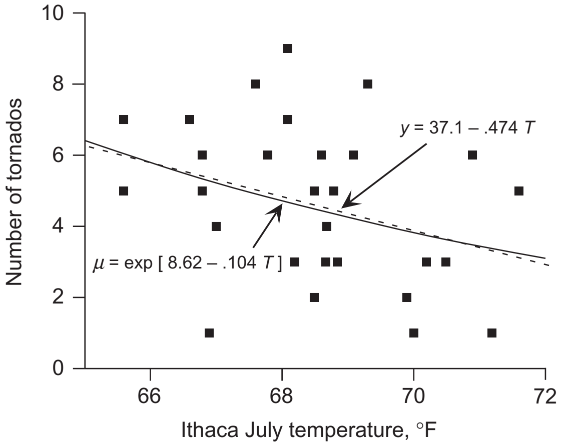


FIGURE 7.21 Annual New York tornado counts, 1959–1988 (Table 7.6), as a function of average Ithaca July temperature in the same year. Solid curve shows the Poisson regression fit using maximum likelihood (Equation 7.45), and dashed line shows ordinary least-squares linear regression.

for the slope parameter b_1 is -1.86, implying a two-tailed p value of 0.068, which is an essentially equivalent result.

The primary difference between the Poisson and linear regressions in Figure 7.21 is in the residual distributions, and therefore in the probability statements about the specified predicted values. Consider, for example, the number of tornados specified when $T = 70$°F. For the linear regression, $\hat{y} = 3.92$ tornados, with a Gaussian $s_e = 2.1$. Rounding to the nearest integer (i.e., using a continuity correction), the linear regression assuming Gaussian residuals implies that the probability for a negative number of tornados is $\Phi[(-0.5–3.92)/2.1] = \Phi[-2.10] = 0.018$, rather than the true value of zero. On the other hand, conditional on a temperature of 70°F, the Poisson regression specifies that the number of tornados will be distributed as a Poisson variable with mean $\mu = 3.82$. Using this mean, Equation 4.12 yields $\Pr\{Y < 0\} = 0$, $\Pr\{Y = 0\} = 0.022$, $\Pr\{Y = 1\} = 0.084$, $\Pr\{Y = 2\} = 0.160$, and so on. ◇

7.7. NONPARAMETRIC REGRESSION

7.7.1. Local Polynomial Regression and Smoothing

Regression settings where we may be unwilling to assume a specific mathematical form, such as Equation 7.25, can be approached nonparametrically using *local polynomial regression*. This method operates by fitting a separate low-order polynomial regression model in a limited neighborhood around each of many target points, building up the regression function as the collection of predictions at these target points. These local regressions are usually linear or quadratic in the predictor variable. Most often the approach is used as a smoothing technique, for example, as an aid to guiding the eye through a scatterplot, in which case it is often referred to as *loess smoothing*. It is also used to interpolate the predictand variable between observed data points, as well as providing the basis for predictions of future data.

Given n developmental data pairs x_i and y_i, the goal is to estimate or predict $\hat{y}(x_0)$ at a target point x_0. The target point need not be one of the values x_i in the training data. For each weighted local quadratic regression around a target point x_0, the regression parameters b_0, b_1, and b_2 are estimated by minimizing

$$\sum_{i=1}^{n} w_i(x_0)\left[y_i - b_0 - b_1(x_0 - x_i) - b_2(x_0 - x_i)^2\right]^2. \tag{7.46}$$

Weighted local linear regressions are obtained by minimizing Equation 7.46 while constraining $b_2 = 0$. Having fit these parameters for a given target point x_0, the estimate or prediction there is $\hat{y}(x_0) = b_0$.

Training-data values further away from the target point are deemphasized or neglected entirely using the weights

$$w_i(x_0) = \begin{cases} h\left(\dfrac{x_i - x_0}{\eta(x_0)}\right), & x_0 - \eta(x_0) < x_i < x_0 + \eta(x_0) \\ 0, & \text{otherwise} \end{cases}, \tag{7.47}$$

where $\eta(x_0)$ is the bandwidth around the target point. Most frequently the tricube kernel,

$$h(u) = \left(1 - |u|^3\right)^3, \tag{7.48}$$

is used as the weighting function, although others such as those in Table 3.1 could be chosen instead.

The results of local regressions depend strongly on the bandwidth, which may be defined using the nearest-neighbor fraction λ. For each target-point regression the nearest λn training-data points receive nonzero weights in Equation 7.47. Specifically,

$$\eta(x_0) = \left|x_0 - d_{(\lambda n)}\right|, \tag{7.49}$$

where

$$d_i = |x_0 - x_i|, \quad i = 1, \ldots, n \tag{7.50}$$

are the distances of the training predictors to the target point, and the parenthetical subscript in Equation 7.49 (possibly rounded to a nearest integer) denotes the order statistic (Section 3.1.2), or the λnth smallest of the n distances. Chosen in this way, the bandwidth will tend to be narrower in regions of higher data density or away from the boundaries and tend to be wider in regions of lower data density or nearer the boundaries.

Larger values of the tuning parameter λ produce smoother results whereas smaller values allow the method to adapt more easily to local variations in the data. When the purpose of the local regression is data smoothing, the nearest-neighbor fraction could be chosen subjectively after viewing results of a range of possibilities. On the other hand, an objective method such as cross validation is more appropriate when the purpose is to obtain a predictive model.

Example 7.11. Smoothing the Mauna Loa CO_2 data

Figure 7.22 repeats the time series of monthly CO_2 measurements from Mauna Loa for 1958–2017, and the least-squares quadratic fit to those data from Figure 7.11 (light line). The heavy line shows the loess smooth of these data using weighted local linear regressions with nearest-neighbor fraction $\lambda = 0.10$, so that 10% of the training data are used in the local regressions for each of the many target points x_0. The two curves are very similar, although the local regression approach captures the dips below

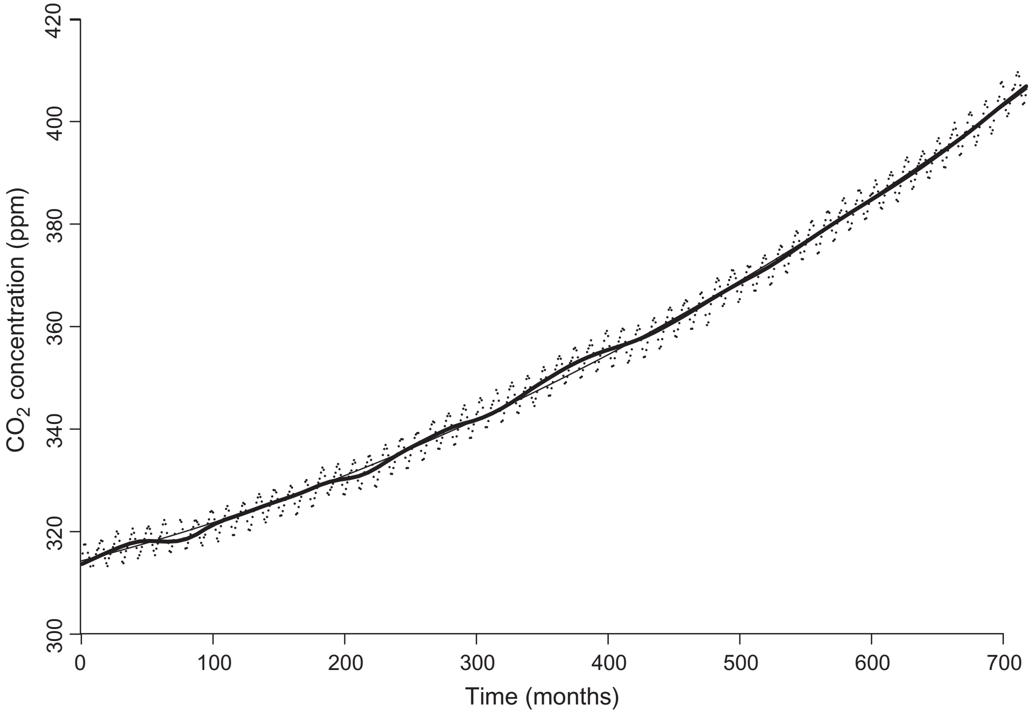


FIGURE 7.22 Loess smooth of the Mauna Loa CO$_2$ data using weighted local linear regressions with nearest-neighbor fraction $\lambda = 0.10$ (heavy curve). The light curve is the conventional quadratic least-squares fit reproduced from Figure 7.11.

the least-squares function around months 75 and 225, and the increase above the smoother function around month 375. Repeating the calculations with wider bandwidths defined by $\lambda = 0.30$ yields results nearly identical to the conventional regression. ◇

Local linear regressions are effective in defining the long-term trend in Figure 7.22 because its curvature is modest. Local linear regressions may be less effective for data exhibiting stronger curvature, as they tend to undershoot peaks and overshoot valleys (Hastie et al., 2009; Loader, 1999), and in such situations local quadratic regressions usually provide better results.

The development in this section has been in terms of a single predictor variable, but the method extends naturally to settings with more than one predictor (Loader, 1999). For example, with 2 predictors the result is a surface above the plane defined by those two predictors. However, the method is generally not useful when there are more than three predictors, because there will be few training-data values in high-dimensional local neighborhoods (Hastie et al., 2009).

7.7.2. Theil-Sen (Median-Slope) Regression

Section 5.3.2 described the nonparametric Mann-Kendall trend test, which involves differencing all distinct pairs of the data values, and so relates to Kendall's τ (Equation 3.35) through Equation 5.32. Although the Mann-Kendall test can address the null hypothesis of no monotonic association, it cannot provide an estimate of the magnitude of any association or address the sampling distribution of such an estimate. Theil-Sen median-slope regression (Sen, 1968; Theil, 1950) provides these nonparametrically, assuming that the form of the association is linear.

As the name suggests, the method operates by computing the median of a collection of slope estimates. These estimates are all the pairwise combinations of

$$b_{i,j} = \frac{y_i - y_j}{x_i - x_j}, \tag{7.51}$$

for pairs in which $x_i \neq x_j$. Here y is the response variable and x is the independent variable (time index if the setting is a time series trend). If there are no repeated x values there are $n(n–1)/2$ such pairs. The median-slope regression is then defined by the slope

$$b = \text{ median}\left(b_{i,j}\right) \tag{7.52a}$$

over the number of pairs for which $x_i \neq x_j$. The estimate for the intercept is

$$a = \text{median}\left(y_i - bx_i\right), \tag{7.52b}$$

over the n pairs in the training data, which result in a rank correlation between the x's and residuals from the regression line being approximately zero.

Median-slope regression is robust to non-Gaussian distribution and resistant to high fractions of outlying data, and yet performs reasonably well when the conventional assumptions in least-squares regression are met. Pairwise slopes for points involving outlying data will typically be extreme in absolute value, but because the slope estimate in Equation 7.51 is the median of all pairwise slopes it will be little affected by a modest number of erroneously large values. The slope estimate in Equation 7.52a is unbiased for an underlying true linear slope. Inferences regarding the estimated slope can be approached through bootstrapping.

Example 7.12. Comparison of Theil-Sen and Ordinary Linear Regression

Example 7.1 considered conventional least-squares regression between the Ithaca minimum temperature in Appendix A.1 as the predictor and the Canandaigua minimum temperature as predictand. There are six Ithaca temperatures appearing twice, and one Ithaca temperature appearing three times, yielding nine combinations of the 31 days with equal predictor values, so that there are (31)(30)/2–9 = 456 date pairs for which $x_i \neq x_j$ in Equation 7.51.

The median of these pairwise slopes is 0.600, which compares well with the least-squares slope estimate of 0.597 in Table 7.2 and Equation 7.21. The median of the 31 intercept estimates (Equation 7.52b with $b = 0.600$) is 12.80, which also compares closely with the least-squares value of 12.46. It is unsurprising that these pairs of estimated regression parameters correspond so closely, because the relationship between the two maximum temperatures is rather well behaved, as can be seen in the appropriate panel of Figure 3.31.

Bootstrapping these data, and applying Equations 7.51, 7.52a, and 7.52b to each bootstrap sample, yields estimated (simple percentile-method) confidence intervals of (9.36, 14.12) for the intercept a, and (0.500, 0.724) for the slope b. Corresponding results using ±2 standard error intervals for these parameters based on the conventional least-squares fitting in Table 7.2 and Equation 7.21 yields the intervals (10.74, 14.17) for the intercept and (0.505, 0.689) for the slope. It is unsurprising that the conventionally

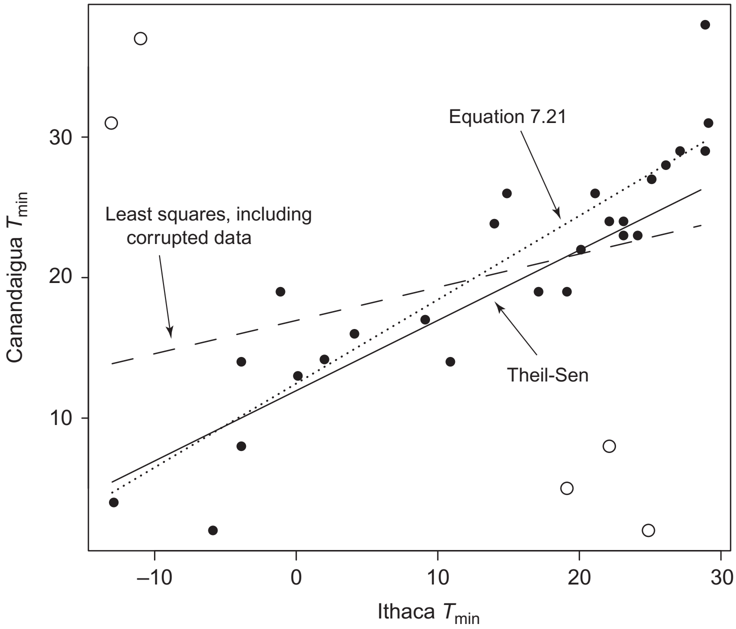


FIGURE 7.23 Scatterplot of January 1987 Canandaigua and Ithaca minimum temperatures, with five of the Canandaigua temperatures corrupted to outlying values (open circles). Solid line shows Theil-Sen median-slope regression, and dashed line shows least-squares fit to the data including the outliers. Dotted line shows least-squares fit to the uncorrupted data (Equation 7.21).

computed intervals are narrower in this instance because the data are plausibly Gaussian, contain no outliers, and appear to be homoscedastic. However, the degradation in estimation precision for the Theil-Sen estimators is quite modest.

The differences between the Theil-Sen and least-squares regressions are more substantial if the data set is contaminated with erroneous outlying values. Figure 7.23 shows a scatterplot in which five of the Canandaigua temperatures (open circles) have been artificially corrupted to be outliers. The least-squares fit to these data ($a = 16.8$, $b = 0.236$, dashed line) is adversely affected by the outlying data, whereas the Theil-Sen equation ($a = 11.8$, $b = 0.50$, solid line) differs only modestly from the least-squares fit (Equation 7.21) to the full correct data set ($a = 12.5$, $b = 0.60$, solid line). ◇

Examples of use of Theil-Sen regression in climatology and hydrology can be found in Hirsch et al. (1982), Huth and Pokorná (2004), Lettenmaier et al. (1994), and Romanic et al. (2015). Some alternative approaches to robust regression are considered by Muhlbauer et al. (2009).

7.7.3. Quantile Regression

Ordinary least-squares regression aims to find a function specifying the conditional mean of the predictand, given particular value(s) of the predictor(s). Alternatively, one can imagine seeking a function to specify the conditional median, or indeed functions specifying other conditional quantiles. This is the setting of *quantile regression* (Koenker and Bassett, 1978), which was introduced into the meteorological literature by Bremnes (2004).

Recall that the quantile q_p is the magnitude of the random variable exceeding p x 100% of the probability in its distribution. For example, $q_{.50}$ denotes the median or 50th percentile. In quantile regression a relatively small set of quantiles is selected; for example, Bremnes (2004) considered the five predictand quantiles $q_{.05}$, $q_{.25}$, $q_{.50}$, $q_{.75}$, and $q_{.95}$. Each selected quantile is represented by a distinct regression equation,

$$q_p(\boldsymbol{x}_i) = b_{q,0} + \sum_{k=1}^{K} b_{q,k} x_{i,k} \tag{7.53}$$

where $\boldsymbol{x}_i$ denotes the vector of K predictor variables $x_{i,1}$, $x_{i,2}$, …, $x_{i,K}$. The predictors $x_{i,k}$ can be different for regression equations pertaining to different quantiles, and the considerations regarding predictor selection that were outlined in Section 7.4 apply also to quantile regression. The coefficients $b_{q,0}$ and $b_{q,k}$ are estimated separately for each quantile by numerically minimizing

$$\sum_{i=1}^{n} \rho_p \left[y_i - q_p(\boldsymbol{x}_i) \right], \tag{7.54}$$

where n is the training-data size, and

$$\rho_p(e_i) = \begin{cases} e_i p & , \mathrm{e}_i \geq 0 \\ e_i(p-1) & , \mathrm{e}_i < 0 \end{cases} \tag{7.55}$$

is called the check function. Because $0 < p < 1$ in Equation 7.55 the check function is nonnegative, so that Equation 7.54 is effectively a weighted sum of absolute values of the errors in the square brackets. Unlike least-squares regression, an analytic minimization of Equation 7.54 does not exist, so that parameter estimates must be computed numerically, usually using a linear programming algorithm.

When the target quantile is the median $q_{0.5}$, quantile regression is equivalent to *least absolute deviation* (LAD) *regression* (Bloomfield and Steiger, 1980; Mielke Jr. et al., 1996; Narula and Wellington, 1982). That is, computing a quantile regression for the median is equivalent to minimizing the sum over the n training samples of the absolute value of the square-bracketed quantity in Equation 7.54, because in that case $\rho_{0.5}[\varepsilon_i] = 0.5\,|\varepsilon_i|$. With this perspective it can be easily seen that LAD regression, and quantile regression more broadly, is resistant to outliers because the residuals e_i are not squared. Interestingly, the idea behind LAD regression predates least-squares regression, and apparently Gauss developed least-squares regression as an analytically tractable alternative (Mielke, 1991).

When the assumptions of least-squares regression are met, namely, homoscedastic Gaussian residual distributions, conditional predictand quantile functions can instead be obtained easily and analytically using Equation 7.23. In addition to being resistant to outliers, quantile regression can also be useful when the residual distributions are asymmetric, or exhibit dispersion that depends on the predictor variable(s), or both. Figure 7.24, showing maximum winter Northern Hemisphere sea-ice extent for 1979–2010, illustrates the point by comparing linear quantile regression functions to the ordinary least-squares fit. The quantile-regression median function is very near the line for $q_{0.8}$ and much further from the $q_{0.2}$ line, and the least-squares mean function is substantially below the quantile-regression median, all of which indicate very substantial negative skewness for the residuals. In addition, the dispersion indicated by the distances between the $q_{0.8}$ and $q_{0.2}$ lines shows increasing variability of winter sea-ice extent later in the record (i.e., nonconstant residual variance), illustrating that quantile regression can easily represent heteroscedasticity.

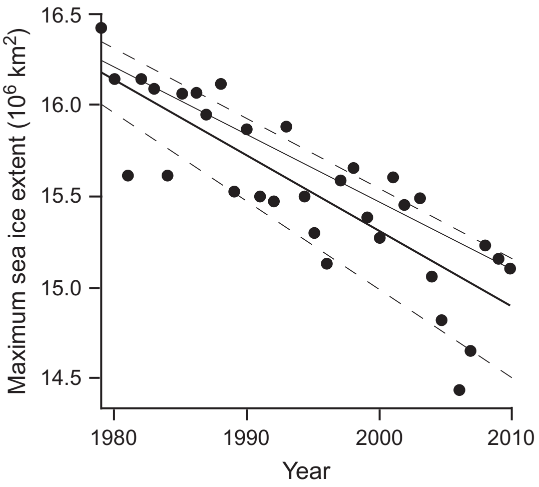


FIGURE 7.24 Linear quantile regressions representing the 20th and 80th percentiles (dashed), and median (light solid) of Northern Hemisphere maximum ice cover extent, 1979–2010. Heavy solid line shows the least-squares linear fit. *After Tareghian and Rasmussen (2013).*

One potential problem with quantile regressions arises as a consequence of the prediction equations for each target quantile being derived independently of one another. Especially in settings where training data are limited, or in instances where equations for a large number of densely spaced quantiles are estimated, the procedure may yield probabilistically incoherent results, meaning that the forecast cumulative probability for a higher quantile may be smaller than the probability for a lower quantile. Such cases would be reflected in plots like those in Figure 7.24 exhibiting functions that cross. For example, extrapolating much beyond 2010 in Figure 7.24 would imply the impossible $q_{0.5} > q_{0.8}$.

Some additional recent examples of the use of quantile regression include the studies of Ben Bouallègue (2016), Bentzien and Friederichs (2012), Jagger and Elsner (2009), Nielsen et al. (2006), and Wasko and Sharma (2014).

7.8. "MACHINE-LEARNING" METHODS

The names "machine-learning" (e.g., Efron and Hastie, 2016; Hsieh, 2009), or "statistical learning" (e.g., Hastie et al., 2009), or "artificial intelligence" methods (e.g., Haupt et al. 2009) refer to computationally intensive algorithms that have been developed relatively recently, enabled by the ongoing exponential increases in computing capacity. These nonparametric and nonlinear methods are extremely flexible and data adaptive. Since they typically involve large numbers of parameters, their estimation usually must involve large training data sets, and their effective use generally requires vigilant attention to potential overfitting. Unlike many traditional statistical methods, they can be deployed in settings where the number of fitted parameters is larger than the training-sample size. On the other hand, their structures are often less easily interpretable than are more conventional statistical methods, so that they are sometimes disparaged as "black boxes." These nonlinear methods may or may not outperform traditional linear statistical methods in atmospheric prediction, even as they are substantially more expensive (Mao and Monahan, 2018).

Machine learning methods are only beginning to be applied to meteorological and climatological forecasting problems, and this section outlines a few of the more common approaches.

7.8.1. Binary Regression Trees and Random Forests

Regression Trees

As the name suggests, *binary regression trees* (Breiman et al., 1984) are regression models, in the sense that each specifies a conditional mean for the predictand variable y conditional on a particular set of predictor variables x_k, $k = 1, \ldots, K$. An individual tree (i.e., a particular regression model) is built through a sequence of binary splittings of the training data made on the basis of the predictor variables, where each split progressively decreases the sum of squares of the training-sample predictand across the groups defined by the binary splits.

At the first step of the algorithm (the "base of the tree trunk"), all possible definitions for two groups g_1 and g_2 of the predictand, defined by binary splits of each of the predictors at each of its possible values, are examined to find the minimum of the combined sum of squares

$$SS = \sum_{y \in g_1} \left(y - \bar{y}_{g_1} \right)^2 + \sum_{y \in g_2} \left(y - \bar{y}_{g_2} \right)^2. \tag{7.56}$$

Thus $K(n–1)$ potential binary splits are evaluated at the initial step in order to find the one minimizing Equation 7.56. This number increases as the algorithm progresses, because the number of groups that can be potentially split increases. The minimization of Equation 7.56 defines the binary split that makes the resulting two (sub)groups of predictand values as different from each other as possible. The search for minimum sum-of-squares binary splits of one of the previously defined groups based on the predictors continues for the second and subsequent steps of the algorithm, yielding a sequence of branches for the tree that ends with a stopping criterion being satisfied. Definition of this endpoint (the "pruning") typically involves the cross-validated minimization of some function of the prediction error, often the mean-squared error (Hastie et al., 2009).

At the final stage, $s+1$ groups have been defined on the basis of s algorithm steps, and each of the $s+1$ terminal branches will be associated with some number of the training-data predictand values (the "leaves"). Using the tree for forecasting involves simply following the branching pattern according to the particular set of K predictor variables at hand, and then estimating the conditional predictand mean, given the predictor set, as the sample mean of the leaves at the end of the terminal branch. Unless the tree is rather large, the identities and specific values of the predictor variables that define each binary split can be easily interpreted.

When the predictand variable is not a continuous quantity, but rather consists of a finite number of discrete classes or groups, the same tree-building approach can be taken to build classification trees (see Section 15.6.2).

Example 7.13. A Simple Binary Regression Tree

Examples 7.5–7.8 illustrated specifications of the January 1987 Canandaigua minimum temperatures as the predictand, on the basis of the other variables listed in Table A.1 as predictors. Figure 7.25 shows the binary regression tree for the same problem.

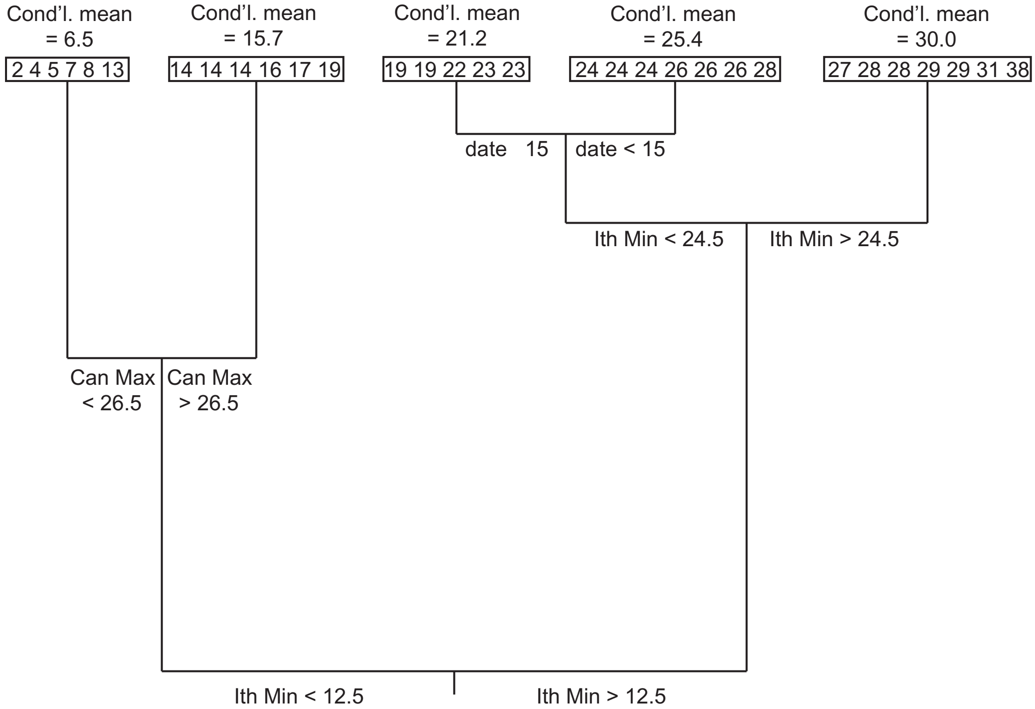


FIGURE 7.25 Binary regression tree for prediction of Canandaigua minimum temperatures, using the other variables in Table A.1 as potential predictors. Boxed values are training-data predictands at each of the five terminal branches, and the five corresponding conditional means are the possible predicted values.

The first binary split, at the base of the tree, indicates that the groups of predictand values most different from each other are defined by whether the Ithaca minimum temperature is colder or warmer than 12.5°F, with colder Ithaca temperatures predicting colder Canandaigua temperatures. The next most effective split, indicated by the next-highest horizontal bar, is based on whether the Canandaigua maximum temperature is below 26.5°F or not, again with colder predictor temperatures indicating colder predictand values. If the Ithaca minimum temperature is warmer than 12.5°F, the most effective binary split is defined by whether the Ithaca minimum temperature is colder or warmer than 24.5°F, and if colder the final split is defined by the date.

In common with the regression models developed in Examples 7.5 and 7.8, the Ithaca minima, Canandaigua maxima, and the date have been chosen as predictors here, but their use in prediction of the Canandaigua minima is nonlinear and discontinuous. The boxed values above the terminal branches show the training-data predictand values defined by the tree structure below (although these are not usually included in the portrayal of a regression tree). Evidently the binary tree structure has been very effective in segregating relatively homogeneous groups of training-data predictand temperatures, with only the 28° value at the second-warmest terminal branch being apparently out of place. The five corresponding conditional means are the possible values that the regression could return as predictions. For example, on any day for which the Ithaca minimum temperature is at least 25°, the corresponding conditional mean value for Canandaigua minimum temperature would be 30°. If the Ithaca minimum is at most 12° and the Canandaigua maximum temperature is at least 27°, the regression tree predicts 15.7° for the Canandaigua minimum temperature. ◇

Random Forests

Binary regression trees are simple to fit, easy to interpret, provide automatic selection of predictor variables, and can be used even when $n < K$. On the other hand, individual trees provide only discontinuous forecasts (e.g., five possible values in Figure 7.25). In addition, they can be unstable, in the sense of their structure varying strongly for different training data sets, and so may perform poorly on independent data. However, these problems can be ameliorated by averaging the predictions of many different regression trees. Such a collection of trees is known as a *random forest* (Breiman, 2001; Hastie et al., 2009).

The collection of regression trees in a random forest differ from each other in that each is computed based on a different bootstrap sample of the training data, which is called bootstrap aggregation, or *bagging* (Breiman, 1996; Hastie et al., 2009). Differences among trees in the random forest are further increased (and the required computations also reduced) by allowing only a random subset of the predictors to be considered as candidates for defining the binary splitting at each new branch. Typical choices for the number of predictors selected for consideration at each step are $K/3$ or $\sqrt{K}$. Forecasts are derived from the random forest by following the branches for each tree as directed by the predictor variables pertaining to the tree in question to their terminal branches, and then averaging the conditional means for the terminal branches of each tree in the forest.

A random forest delivers an estimate of the conditional mean of the predictand, given a particular set of predictor variables. The idea can be extended to estimation of the full predictive distribution using *quantile regression forests* (Meinshausen, 2006). Despite the similarity of the names, these are quite different from the quantile regressions described in Section 7.7.3. Forecast distributions from quantile random forests are based on the subsets of training-sample predictands at each of the terminal branches in the random forest (i.e., the leaves), which define an empirical predictive distribution for that terminal branch. The quantile-regression-forest predictive distribution is then formed as the average of the empirical distributions at the terminal leaves, over all trees in the random forest. Herman and Schumacher (2018), Taillardat et al. (2017), and Whan and Schmeits (2018) provide operationally oriented MOS (Section 7.9.2) examples. Because the resulting predictive distribution is the average of the conditional training-data empirical distributions for the terminal leaves, incoherent probability forecasts (such as negative event probabilities or nonnegative probabilities for impossible outcomes) cannot be produced. On the other hand, neither can nonzero probability be assigned to predictand values outside their range in the training data, so that extreme-value forecasts may be problematic.

7.8.2. Artificial Neural Networks

Artificial Neural Networks (ANNs) provide another nonlinear and nonparametric regression approach. The structure of ANNs is meant to emulate a particular conceptual model of brain functioning (McCulloch and Pitts, 1943). The idea is that an individual neuron sums the signals it receives from other neurons, and if that sum is sufficiently large it "fires" a signal of its own to other neurons. If the received signals are not strong enough the neuron is silent and does not signal to other neurons in the network.

Structures of particular ANNs, representing highly simplified abstractions of this underlying biological neural model, are typically represented by diagrams such as those in Figure 7.26, called feed-forward *perceptrons*. The circles in this figure represent individual neurons, often called nodes, which are arranged in layers. Each node is connected to all other nodes in adjacent layers, as indicated by the lines, and the information flow is exclusively from left to right (i.e., feeding forward). The predictor data, consisting of $K = 4$ predictor variables in both panels of Figure 7.26, provide the initial signal to the first

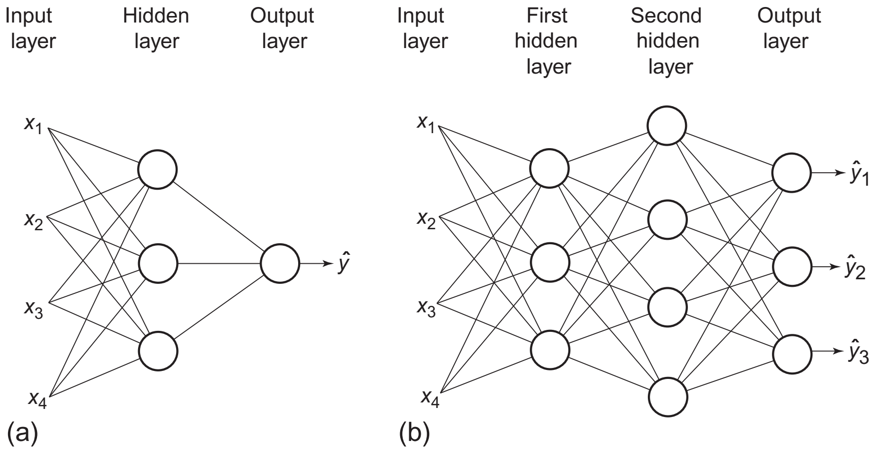


FIGURE 7.26 Two example ANN diagrams, both of which operate on $K = 4$ predictor variables. Panel (a) shows a network with a single hidden layer and one predictand variable. The network in panel (b) has two hidden layers and three related predictand outputs.

of possibly multiple "hidden" layers. The ANN represented in Figure 7.26a has a single hidden layer, and the ANN in Figure 7.26b has two hidden layers. The final, or output, layer transforms the outputs from the last hidden layer to the predicted value or values. For the ANN in Figure 7.26a this predictand is a single scalar value, and in Figure 7.26b the outputs are three related predictands.

Diagrams such as those in Figure 7.26 actually represent a sequence of parallel computations on the predictor variables. In particular, each circle indicates a transformation of a linear combination of the values it receives. For example, consider the simpler ANN represented in Figure 7.26a, and define z_j, $j = 1, 2, 3$, to be the output of the jth neuron in the hidden layer,

$$z_j = h\left(b_{j,0} + \sum_{k=1}^{K} b_{j,k} x_k\right), \tag{7.57}$$

where $h(\bullet)$ is called an *activation function*, which in general will be nonlinear (with the possible exception of the operations in the output layer). The b parameters are regression coefficients that need to be estimated using the training data.

Adhering to the original on/off concept of neuron interaction would suggest choosing a step function for the activation function. More commonly, smooth sigmoid (S-shaped) and bounded functions are chosen for this purpose. The logistic function

$$h(u) = \frac{1}{1 + e^{-u}} \tag{7.58}$$

(see also Equation 7.36b and Figure 7.20) is a very common choice, which is bounded by $0 < h(u) < 1$, and so approximates a step function in a way that is smooth and differentiable. For large positive values of its argument it is nearly fully "on," and for large negative argument values it is nearly fully "off." Another common choice for the activation function is the hyperbolic tangent function,

$$h(u) = \frac{e^u - e^{-u}}{e^u + e^{-u}}, \tag{7.59}$$

which is also smooth, differentiable, sigmoid in shape, and bounded as $-1 < h(u) < 1$. For u in the neighborhood of 0, Equations 7.58 and 7.59 approximate linear responses.

The activation function for the output layer is often different from that chosen for the hidden layers. In a conventional regression setting where there is a single scalar predictand, such as the ANN diagrammed in Figure 7.26a, the activation function in the output layer is usually just the identity $h(u) = u$. If the ANN is meant to yield probabilities for multiple discrete outcomes, such as diagrammed in Figure 7.26b, a reasonable choice is the *softmax* function

$$h(u_j) = \frac{e^{u_j}}{\sum_{m=1}^{J} e^{u_m}}, \tag{7.60}$$

which yields a collection of J positive values that sum to one. If the predictand is the probability for a single binary outcome, using the logistic function (Equation 7.58) at the single node of the output layer is a natural choice. These latter two cases are essentially probabilistic discrimination and classification (Chapter 15) applications.

The foregoing discussion indicates that ANNs are elaborations on the structure of generalized linear models (Section 7.6.1). Indeed, abstracting Figure 7.26a to a "network" with no hidden layer, and a single output node employing the logistic activation function, yields an ordinary logistic regression (Section 7.6.2). On the other hand, the complexity, and thus the flexibility, of ANN models expands very rapidly as layers and nodes are added. For example, possibly the simplest ANN network structure consists of a single, two-node hidden layer, and a single output node containing the identity activation function, which would be represented mathematically as

$$\hat{y} = b_0 + b_1\, h\left(b_{1,0} + \sum_{k=1}^{K} b_{1,k} x_k\right) + b_2\, h\left(b_{2,0} + \sum_{k=1}^{K} b_{2,k} x_k\right). \tag{7.61}$$

This is a highly flexible equation, but even this extremely simple ANN would require estimation of $2(K+1)+3$ parameters whose interpretation in terms of the underlying physical variables would be very difficult.

Parameter estimation typically begins using random, near-zero initial guesses for the parameters (and so a near-linear initial model), and proceeds iteratively by minimizing a penalty function such as squared prediction error or the negative of maximized likelihood. Because the number of parameters to be estimated may be quite large, regularization methods (Section 7.5), which penalize a model for large or numerous nonzero values for the parameters multiplying the inputs (the "intercept" parameters are usually not restricted), are generally needed to prevent overfitting. Accordingly, the predictor variables should be standardized (converted to standardized anomalies) to have comparable magnitudes, so that the regularization constraints are equitably distributed among them. Efron and Hastie (2016) and Hastie et al. (2009) provide details on fitting algorithms and their regularization.

There appear to be no clear-cut rules or guidelines for choosing the number of hidden layers or the numbers of nodes in each hidden layer. Hsieh (2009) notes that accurate representation of more complicated and irregular functional relationships requires more hidden nodes, whereas smoother relationships can be captured using fewer parameters. Hastie et al. (2009) state that typical ANN models usually contain somewhere between 5 and 100 hidden nodes in total. Efron and Hastie (2016) suggest that it is better to have too many rather than too few hidden nodes, since overall model complexity can be controlled through regularized parameter estimation.

7.8.3. Support Vector Regression

Consider again the conventional multiple regression model in Equation 7.25. Some of the possible approaches to estimating the parameters b_0, b_1, …, b_K, can be generalized as finding the parameter set that minimizes the function

$$\sum_{i=1}^{n} g(y_i - \hat{y}) + \lambda \sum_{k=1}^{K} b_k^2 \tag{7.62}$$

over the n training-data values. For example, if the error penalty $g(u) = u^2$ then Equation 7.62 represents ordinary least-squares regression (Section 7.3) if $\lambda = 0$, and ridge regression (Section 7.5.1) if $\lambda \neq 0$. Similarly, if $g(u) = |u|$ then Equation 7.62 represents LAD regression (quantile regression for the median, Section 7.7.3), with ($\lambda \neq 0$) or without ($\lambda = 0$) regularization.

In *support vector regression*, the parameters are estimated by minimizing Equation 7.62 using the "δ-insensitive" error penalty

$$g(u) = \begin{cases} 0 & , |u| \leq \delta \\ |u| - \delta, & |u| > \delta \end{cases}, \tag{7.63}$$

that ignores regression residuals smaller than δ, and imposes an absolute-error penalty otherwise. Thus support vector regression seeks to orient the regression function in a way that maximizes the number of training-data points with residuals smaller than δ in absolute value and is fairly robust to outlying data because of the absolute error rather than squared-error loss specified by Equation 7.63. The parameter estimates depend only on the predictor values yielding residuals larger than δ in absolute value, which are the "support vectors." The two adjustable parameters δ and λ might be estimated through cross-validation.

The foregoing development described linear support vector regression. More typically nonlinear support vector regression is of interest, for which the regression equation can be written

$$\hat{y} = b_0 + \sum_{m=1}^{M} b_m h_m(x_1, x_2, \ldots, x_K). \tag{7.64}$$

Here each of the basis functions h_m depends in general nonlinearly on all K of the predictor variables. Linear support vector regression is then a special case, with $M = K$, and $h_m(\mathbf{x}) = x_m$. The computational details are somewhat elaborate and can be found in Hastie et al. (2009) and Hsieh (2009). When the predictand values indicate memberships in discrete classes or groups, the same general approach can be used for classification (Section 15.6.1).

7.9. OBJECTIVE FORECASTS USING TRADITIONAL STATISTICAL METHODS

7.9.1. Classical Statistical Forecasting

Construction of weather forecasts through purely statistical means—that is, without the benefit of information from fluid-dynamical weather prediction models—has come to be known as classical statistical forecasting. This name reflects the long history of the use of purely statistical forecasting methods, dating from the time before the availability of dynamical forecast information. The accuracy of dynamical forecasts has advanced sufficiently that pure statistical forecasting is currently used in practical settings only for very short lead times or for fairly long lead times.

Very often classical forecast products are based on multiple regression equations of the kinds described in Sections 7.2 and 7.3. These statistical forecasts are objective in the sense that a particular set of inputs or predictors will always produce the same forecast for the predictand, once the forecast equation has been developed. However, many subjective decisions necessarily go into the development of the forecast equations.

The construction of a classical statistical forecasting procedure follows from a straightforward implementation of the ideas presented in previous sections of this chapter. Required developmental data consist of past values of the quantity to be forecast, and a matching collection of potential predictors whose values will be known prior to the forecast time. A forecasting procedure is developed using this set of historical data, which can then be used to forecast future values of the predictand on the basis of future observations of the predictor variables. It is thus a characteristic of classical statistical weather forecasting that the time lag is built directly into the forecast equation through the time-lagged relationships between the predictors and the predictand.

For lead times up to a few hours, purely statistical forecasts still find productive use. This short-lead forecasting niche is known as *nowcasting*. Dynamically based forecasts are not practical for nowcasting because of the delays introduced by the processes of gathering weather observations, data assimilation (calculation of initial conditions for the dynamical model), the actual running of the forecast model, and the postprocessing and dissemination of the results.

One very simple statistical approach that can produce competitive nowcasts is use of *conditional climatology*, that is, historical statistics subsequent to (conditional on) analogous weather situations in the past. The result could be a conditional frequency distribution for the predictand, or a single-valued forecast corresponding to the expected value (mean) of that conditional distribution. A more sophisticated approach is to construct a regression equation to forecast a few hours ahead. For example, Vislocky and Fritsch (1997) compare these two approaches for forecasting airport ceiling and visibility at lead times of one, three, and six hours.

At lead times beyond perhaps two weeks, purely statistical forecasts are again competitive with dynamical forecasts. At these longer lead times the sensitivity of dynamical models to the unavoidable small errors in their initial conditions, described in Chapter 8, makes explicit forecasting of specific weather events problematic. This estimated limit of approximately two weeks for explicit dynamical predictability of the atmosphere has remained basically unchanged since the question began being investigated in the 1960s (e.g., Buizza and Leutbecher, 2015; Lorenz, 1982; Simmons and Hollingsworth, 2002).

Although long-lead forecasts for seasonally averaged quantities currently are made using dynamical models (e.g., Barnston et al., 2003; Kirtman et al., 2014; Stockdale et al., 2011), comparable or even better predictive accuracy at substantially lower cost is still obtained through statistical methods (e.g., Hastenrath et al., 2009; Moura and Hastenrath, 2004; Quan et al., 2006; Van den Dool, 2007; Wilks, 2008; Zheng et al., 2008). Possibly the failure of dynamical methods to consistently outperform relatively simple statistical methods for seasonally averaged quantities is due to the linearization of these prediction problems that is induced by the inherent long time averaging (Yuval and Hsieh, 2002; Hsieh, 2009), through effects described by the Central Limit Theorem (Sections 4.4.2 and 12.5.1).

Often the predictands in these seasonal forecasts are spatial patterns, and so the forecasts involve multivariate statistical methods that are more elaborate than those described in Sections 7.2 and 7.3 (e.g., Barnston, 1994; Mason and Mimmack, 2002; Ward and Folland, 1991; Wilks, 2008, 2014a, 2014b; see Sections 14.2.3, 14.3.2, and 15.4). However, univariate regression methods are still appropriate and useful for single-valued predictands. For example, Knaff and Landsea (1997) used ordinary least-squares regression for seasonal forecasts of tropical sea-surface temperatures with observed sea-surface temperatures as predictors, and Elsner and Schmertmann (1993) used Poisson regression for seasonal prediction of hurricane numbers.

Example 7.14. A Set of Classical Statistical Forecast Equations

The flavor of classical statistical forecast methods can be appreciated by looking at the NHC-67 procedure for forecasting hurricane movement (Miller et al., 1968). This relatively simple set of regression equations was used as part of the operational suite of forecast models at the U.S. National Hurricane Center until 1988 (Sheets, 1990). Since hurricane movement is a vector quantity, each forecast consists of two equations: one for northward movement and one for westward movement. The two-dimensional forecast displacement is then computed as the vector sum of the northward and westward forecasts.

The predictands were stratified according to two geographical regions: north and south of 27.5°N latitude. That is, separate forecast equations were developed to predict storms located on either side of this latitude at the time of forecast initialization. This division was based on the subjective experience of the developers regarding the responses of hurricane movement to the larger-scale flow, and in particular that storms moving in the trade winds in the lower latitudes tend to behave less erratically. Separate forecast equations were also developed for "slow" vs. "fast" storms. The choice of these two stratifications was also made subjectively, on the basis of the experience of the developers. Separate equations are also needed for each forecast lead time (0–12 h, 12–24 h, 24–36 h, and 36–48 h, yielding a total of 2 (displacement directions) × 2 (regions) × 2 (speeds) × 4 (lead times) = 32 separate regression equations in the NHC-67 package.

The available developmental data set consisted of 236 northern cases (initial position for hurricanes) and 224 southern cases. Candidate predictor variables were derived primarily from 1000-, 700-, and 500-mb heights at each of 120 gridpoints in a 5° x 5° coordinate system that follows the storm. Predictors derived from these $3 \times 120 = 360$ geopotential height predictors, including 24-h height changes at each level, geostrophic winds, thermal winds, and Laplacians of the heights, were also included as candidate predictors. Additionally, two persistence predictors, observed northward and westward storm displacements in the previous 12 h, were included.

With vastly more potential predictors than observations, some screening procedure is clearly required. Here forward selection was used, with the (subjectively determined) stopping rule that no more than 15 predictors would be in any equation, and new predictors would be only included to the extent that they increased the regression R^2 by at least 1%. This second criterion was apparently sometimes relaxed for regressions with few predictors.

TABLE 7.7 Regression Results for the NHC-67 Hurricane Forecast Procedure, for the 0–12 h Westward Displacement of Slow Southern-Zone Storms, Indicating the Order in Which the Predictors were Selected and the Resulting R^2 at Each Step

Predictor	Coefficient	Cumulative R^2
Intercept	−2709.5	–
P_X	0.8155	79.8%
Z_{37}	0.5766	83.7%
P_Y	−0.2439	84.8%
Z_3	−0.1082	85.6%
P_{51}	−0.3359	86.7%

The meanings of the symbols for the predictors are P_X = westward displacement in the previous 12 h, Z_{37} = 500 mb height at the point 10° north and 5° west of the storm, P_Y = northward displacement in the previous 12 h, Z_3 = 500 mb height at the point 20° north and 20° west of the storm, and P_{51} = 1000 mb height at the point 5° north and 5° west of the storm. Distances are in nautical miles, and heights are in meters.
From Miller et al. (1968). © American Meteorological Society. Used with permission.

Table 7.7 presents the results for the 0–12 h westward displacement of slow southern storms in NHC-67. The five predictors are shown in the order they were chosen by the forward selection procedure, together with the R^2 value achieved on the developmental data at each step. The coefficients are those for the final ($K = 5$) equation. The most important single predictor was the persistence variable (P_X), reflecting the tendency of hurricanes to change speed and direction fairly slowly. The 500 mb height at a point north and west of the storm (Z_{37}) corresponds physically to the steering effects of midtropospheric flow on hurricane movement. Its coefficient is positive, indicating a tendency for westward storm displacement given relatively high heights to the northwest, and slower or eastward (negative westward) displacement of storms located southeast of 500 mb troughs. The final two or three predictors appear to improve the regression only marginally—the predictor Z_3 increases the R^2 by $<1\%$—and it is quite possible that the $K = 2$ or $K = 3$ predictor models might have been chosen, and might have been equally accurate for independent data, if cross-validation had been known to and computationally feasible for the developers. Remarks in Neumann et al. (1977) in relation to Figure 7.14, concerning the fitting of the similar NHC-72 regressions, are also consistent with the idea that the equation represented in Table 7.7 may have been overfit. ◇

7.9.2. Model Output Statistics (MOS)

Pure classical statistical weather forecasts for lead times in the range of a few days to a week or two are generally no longer employed, since dynamical models now allow more accurate forecasts at these time scales. However, raw dynamical forecast outputs typically exhibit systematic errors which can be corrected through statistical postprocessing. Generally this process is carried out using large multiple regression equations in a way that is analogous to the classical approach, so that many of the same technical considerations pertaining to equation fitting apply. The difference has to do with the range of available predictor variables. In addition to conventional predictors such as current meteorological observations, the date, or climatological values of a particular meteorological element, predictor variables taken from the outputs of the dynamical models are also used.

There are three reasons why statistical reinterpretation of dynamical forecast output is useful for practical weather forecasting:

- There are important differences between the real world and its representation in the dynamical models, and these differences have important implications for the forecast enterprise (e.g., Gober et al., 2008). Dynamical models necessarily simplify and homogenize surface conditions, by representing the world as an array of gridpoints to which the forecast output pertains. Small-scale effects (e.g., of topography or small bodies of water) important to local weather may not be represented in a dynamical model. Also, locations and variables for which forecasts are needed may not be represented explicitly. However, statistical relationships can be developed between the information provided by the dynamical models and desired forecast quantities and locations to help alleviate these problems.
- Dynamical models are not complete and true representations of the workings of the atmosphere, particularly at the smaller time and space scales, and they are inevitably initialized at states that differ from the true initial state of the atmosphere. For both of these reasons, their forecasts are subject to errors. To the extent that these errors are systematic, statistical postprocessing can compensate and correct the resulting forecast biases.
- The dynamical models are deterministic. That is, even though the future state of the weather is inherently uncertain, a single integration is capable of producing only a single forecast for any meteorological element, given a particular set of initial model conditions. Using dynamical forecast information in conjunction with statistical methods allows quantification and expression of the uncertainty associated with different forecast situations. For example, it is possible to derive probability forecasts, using methods such as REEP or logistic regression, using predictors taken from even a single deterministic dynamical integration. Although ensemble forecasting (Chapter 8) is increasingly used to represent forecast uncertainty, these dynamical forecasts also benefit from statistical postprocessing (Sections 8.3 and 8.4).

The first statistical approach to be developed for taking advantage of deterministic dynamical forecasts was called *perfect prog* (Klein et al., 1959), which is short for perfect prognosis. As the name implies, the perfect prog technique made no attempt to correct for possible dynamical model errors or biases, but rather took their forecasts for future atmospheric variables at face value—assuming them to be perfect. The perfect prog method involved developing regression equations relating predictands of interest to simultaneously observed predictor variables. At first, it might seem that this would not be a productive approach to forecasting. For example, tomorrow's 1000–850 mb thickness may be an excellent predictor for tomorrow's maximum temperature, but tomorrow's thickness will not be known until tomorrow. However, in implementing the perfect-prog approach, it is the dynamical forecasts of the predictors (e.g., today's dynamical forecast for tomorrow's thickness) that are substituted into the regression equation as predictor values. Therefore the forecast time lag in the perfect-prog approach is contained entirely in the dynamical model. A key advantage of the perfect prog approach in the early days of dynamical weather forecasting was that it did not require an archive of past forecasts to fit and implement it.

The *Model Output Statistics* (MOS) approach (Carter et al., 1989; Glahn and Lowry, 1972) began to be used in preference to perfect prog after a sufficient historical archive of dynamical forecasts had been developed. The MOS approach extends classical statistical forecasting by including dynamical forecasts available at the time the forecast must be issued, but which pertain to the future time being forecast. Preference for the MOS approach over perfect prog derives from its capacity to include directly in the regression equations the influences of specific characteristics of particular dynamical models. Separate MOS forecast equations must be developed for different forecast lead times. This is because the error characteristics of the dynamical forecasts are different at different lead times, producing, for example, different statistical relationships between observed temperature and forecast thicknesses for 24 h versus 48 h in the future. In addition, since the MOS equations are tuned to the particular error characteristics of the model for which they were developed, different MOS equations will, in general, be required for use with different dynamical models.

The classical, perfect-prog, and MOS approaches are most commonly based on multiple linear regression, exploiting correlations between a predictand and available predictors (although nonlinear regressions can also be used: e.g., Lemcke and Kruizinga, 1988; Marzban et al., 2007; Vislocky and Fritsch, 1995b; Wilks, 2009, 2018b). In the classical approach it is the correlations between today's values of the predictors and tomorrow's predictand that forms the basis of the forecast. For the perfect-prog approach it is the simultaneous correlations between today's values of both predictand and predictors that are the statistical basis of the prediction equations. In the case of MOS forecasts, the prediction equations are constructed on the basis of correlations between dynamical forecasts as predictor variables, and the subsequently observed value of tomorrow's predictand.

Fitting MOS equations requires an archived record of forecasts from the same dynamical model that will ultimately be used to provide input to the MOS equations. Typically, several years of archived dynamical forecasts are required to develop a stable set of MOS forecast equations (e.g., Jacks et al., 1990). This requirement can be a substantial limitation, because the dynamical models are not static. Rather, these models regularly undergo changes aimed at improving their performance. Minor changes in a dynamical model leading to reductions in the magnitudes of its random errors will not substantially degrade the performance of a set of MOS equations (e.g., Erickson et al., 1991). However, modifications to the model that change—even substantially reducing—systematic errors will require redevelopment of accompanying MOS forecast equations. Since it is a change in the dynamical model that will have necessitated the redevelopment of a set of MOS forecast equations, it is often the case that a sufficiently long developmental sample of predictors from the improved dynamical model will not be immediately available. However, as the quality of dynamical forecast models continues to improve, fewer dynamical predictors are required to achieve good results using MOS regressions, so that shorter training periods can be used (Glahn, 2014).

The MOS approach to statistical forecasting has two advantages over the perfect prog approach that make MOS the method of choice when practical. The first of these is that model-calculated, but unobserved, quantities such as vertical velocity can be used as predictors. However, the dominating advantage of MOS over perfect prog is that systematic errors exhibited by the dynamical model are accounted for in the process of developing the MOS equations. Since the perfect-prog equations are developed without reference to the characteristics of any particular dynamical model, they cannot account for or correct their forecast errors. The MOS development procedure allows compensation for these systematic errors when the forecasts are computed. Systematic errors include such problems as progressive cooling or warming biases in the dynamical model with increasing forecast lead time, a tendency for synoptic features to move too slowly or too quickly in the dynamical model, and even the unavoidable decrease in forecast accuracy at increasing lead times.

The compensation for systematic errors in a dynamical model that is accomplished by MOS forecast equations is easiest to see in relation to a simple bias in an important predictor. Figure 7.27 illustrates

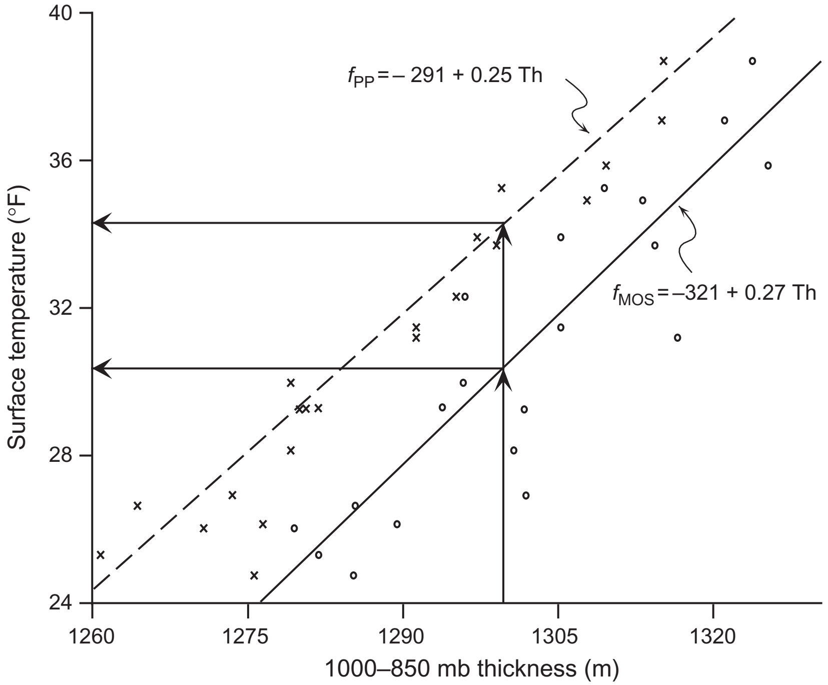


FIGURE 7.27 Illustration of the capacity of a MOS equation to correct for systematic bias in a hypothetical dynamical model. The x's represent observed, and the circles represent dynamically forecast 1000–850 mb thicknesses, in relation to hypothetical surface temperatures. The bias in the dynamical model is such that the forecast thicknesses are too large by about 15 m, on average. The MOS equation (solid line) is calibrated for this bias and produces a reasonable temperature forecast (lower horizontal arrow) when the forecast thickness is 1300 m. The perfect-prog equation (dashed line) incorporates no information regarding the attributes of the dynamical model and produces a surface temperature forecast (upper horizontal arrow) that is too warm as a consequence of the thickness bias.

a hypothetical case, where surface temperature is to be forecast using the 1000–850 mb thickness. The x's in the figure represent the (unlagged, or simultaneous) relationship of a set of observed thicknesses with observed temperatures, and the circles represent the relationship between thicknesses previously forecast by a dynamical model, with the same temperature data. As drawn, the hypothetical dynamical model tends to forecast thicknesses that are too large by about 15 m. The scatter around the perfect-prog regression line (dashed) derives from the fact that there are influences on surface temperature other than those captured by the concurrent 1000–850 mb thickness. The scatter around the MOS regression line (solid) is greater, because in addition it reflects errors in the dynamical model.

The observed thicknesses (x's) in Figure 7.27 appear to forecast the simultaneously observed surface temperatures reasonably well, yielding an apparently good perfect-prog regression equation (dashed line). The relationship between forecast thickness and observed temperature represented by the MOS equation (solid line) is substantially different, because it includes the tendency for this dynamical model to systematically overforecast thickness. If this model produces a thickness forecast of 1300 m (vertical arrows), the MOS equation corrects for the bias in the forecast thickness and produces a reasonable temperature forecast of about 30°F (lower horizontal arrow). Loosely speaking, the MOS knows that when

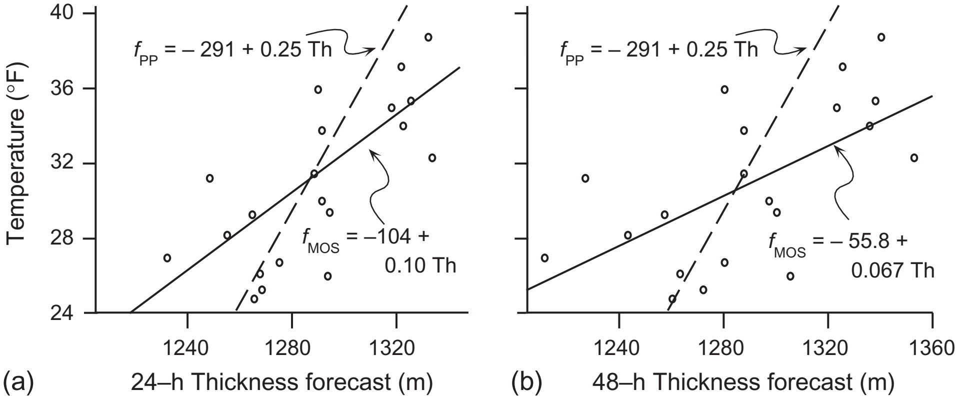


FIGURE 7.28 Illustration of the capacity of a MOS equation to account for the systematic tendency of dynamical forecasts to become less accurate at longer lead times. The points in these panels are simulated thickness forecasts, constructed from the x's in Figure 7.27 by adding random errors to the thickness values. As the forecast accuracy degrades at longer lead times, the perfect-prog equation (dashed line, reproduced from Figure 7.27) is increasingly overconfident and tends to forecast extreme temperatures too frequently. At longer lead times (b) the MOS equations increasingly provide forecasts near the climatological average temperature (30.8°F in this example).

this dynamical model forecasts 1300 m, a more reasonable expectation for the true future thickness is closer to 1285 m, which in the climatological data (x's) corresponds to a temperature of about 30°F. The perfect-prog equation, on the other hand, operates under the assumption that a dynamical model will forecast the future thickness perfectly. It therefore yields a temperature forecast that is too warm (upper horizontal arrow) when supplied with a thickness forecast that is too large.

A more subtle systematic error exhibited by all dynamical weather forecasting models is the degradation of forecast accuracy at increasing lead time. The MOS approach accounts for this type of systematic error as well. The situation is illustrated in Figure 7.28, which is based on the hypothetical observed data in Figure 7.27. The panels in Figure 7.28 simulate the relationships between forecast thicknesses from an unbiased dynamical model at 24- and 48-h lead time and the surface temperature, and have been constructed by adding random errors to the observed thickness values (x's) in Figure 7.27. These random errors exhibit $\sqrt{\text{MSE}} = 20$ m for the 24-h lead time and $\sqrt{\text{MSE}} = 30$ m at the 48-h lead time. The increased scatter of points for the simulated 48-h lead time illustrates that the regression relationship is weaker when the dynamical model is less accurate.

The MOS equations (solid lines) fit to the two sets of points in Figure 7.28 reflect the progressive loss of predictive accuracy of the dynamical model at longer lead times. As the scatter of points increases the slopes of the MOS forecast equations become more horizontal, leading to temperature forecasts that are more like the climatological mean temperature, on average. This characteristic is reasonable and desirable, since as the dynamical model provides less information about the future state of the atmosphere at longer lead times, temperature forecasts differing substantially from the climatological average temperature are progressively less well justified. In the limit of a few weeks lead time, a dynamical model will really provide no more information than will the climatological value of the predictand, so that the slope of the corresponding MOS equation would be zero, and the appropriate temperature forecast consistent with this (lack of) information would simply be the climatological average temperature. Thus it is sometimes said that MOS "converges to the climatology." By contrast, the perfect-prog equation (dashed lines, reproduced from Figure 7.27) take no account of the decreasing accuracy of the dynamical model at longer lead times and continue to produce temperature forecasts as if the thickness forecasts were perfect. Figure 7.28 emphasizes that the result is overconfident temperature forecasts, with both very warm and very cold temperatures forecast much too frequently.

Although MOS postprocessing of dynamical forecasts is strongly preferred to perfect prog and to the raw dynamical forecasts themselves, the pace of changes made to dynamical models continues to accelerate as computing capabilities progressively increase. Operationally it would not be practical to wait for two or three years of new dynamical forecasts to accumulate before deriving a new MOS system, even if the dynamical model were to remain static for that period of time. One option for maintaining MOS systems in the face of this reality is to retrospectively *reforecast* weather for previous years using the current updated dynamical model (Hagedorn, 2008; Hamill et al., 2004, 2013). Because daily weather data are typically strongly autocorrelated, the reforecasting process is more efficient if several days are omitted between the reforecast days (Hamill et al., 2004). Even if the computing capacity to reforecast is not available, a significant portion of the benefit of fully calibrated MOS equations can be achieved using a few months of training data (Mao et al., 1999; Neilley et al., 2002). Alternative common approaches include using longer developmental data records together with whichever version of the dynamical model was current at the time, and weighting the more recent forecasts more strongly. This can be done either by downweighting forecasts made with older model versions (Wilson and Vallée, 2002, 2003), or by gradually downweighting older data, usually using an algorithm called the *Kalman filter* (Cheng and Steenburgh, 2007, Crochet, 2004, Cui et al. 2012, Galanis and Anadranistakis, 2002, Homleid, 1995, Kalnay, 2003, Mylne et al., 2002b, Valée et al., 1996), although other approaches are also possible (Yuval and Hsieh, 2003).

7.9.3. Operational MOS Forecasts

Interpretation and extension of dynamical forecasts using MOS systems has a long-standing history at a number of national meteorological centers, including those in the Netherlands (Lemcke and Kruizinga, 1988), Britain (Francis et al., 1982), Italy (Conte et al., 1980), China (Lu, 1991), Spain (Azcarraga and Ballester, 1991), Canada (Brunet et al., 1988), and the United States (Carter et al., 1989; Glahn et al., 2009a), among others. Most of MOS applications have been oriented toward ordinary weather forecasting, but the method is equally well applicable in areas such as postprocessing of dynamical seasonal forecasts (e.g., Lepore et al., 2017; Shongwe et al., 2006; Vigaud et al., 2017).

MOS forecast products can be quite extensive, as exemplified by Table 7.8, which shows a collection of MOS forecasts for LaGuardia airport, New York City, for the 0600 UTC forecast cycle on 23 February 2018. This is one of hundreds of such panels for locations in the United States, for which these forecasts are issued four times daily and posted on the internet by the U.S. National Weather Service. Forecasts for a wide variety of weather elements are provided, at lead times up to 60 h and at intervals as close as 3 h. After the first few lines indicating the dates and times (UTC), are forecasts for daily maximum and minimum temperatures; temperatures, dew point temperatures, cloud coverage, wind speed, and wind direction at 3 h intervals; probabilities of measurable precipitation at 6- and 12-h intervals; forecasts for precipitation amount; thunderstorm probabilities; and forecast ceiling, visibility, and obstructions to visibility. Similar panels, based on several other dynamical models, are also produced and posted.

The MOS equations underlying forecasts such as those presented in Table 7.8 are seasonally stratified, usually with separate forecast equations for the "warm season" (April through September) and "cool season" (October through March). This two-season stratification allows the MOS forecasts to incorporate different relationships between predictors and predictands at different times of the year. A finer stratification (three-month seasons, or separate month-by-month equations) would probably be preferable if sufficient developmental data were available.

TABLE 7.8 Example MOS Forecasts Produced by the U.S. National Meteorological Center for LaGuardia Airport, New York City, Shortly after 0600 UTC on 23 February 2018

KLGA		GFS	MOS	GUIDANCE			2/23/2018				0600	UTC									
DT	/FEB	23			/FEB		24						/FEB		25					/	
HR	12	15	18	21	00	03	06	09	12	15	18	21	00	03	06	09	12	15	18	00	06
X/N					44				42				59				40			53	
TMP	37	38	39	40	41	42	44	45	46	51	56	56	52	46	43	43	43	44	48	52	50
DPT	28	28	32	37	40	40	41	41	40	37	36	36	36	35	35	36	37	38	39	43	37
CLD	OV	OV	OV	OV	OV	OV	OV	OV	OV	BK	OV	OV	OV	OV	OV	OV	OV	OV	OV	OV	BK
WDR	07	09	09	10	13	22	24	27	30	33	36	36	05	07	07	07	08	08	06	01	30
WSP	08	09	08	06	04	06	07	06	07	07	04	07	07	07	10	14	16	15	08	06	06
P06			41		70		60		9		3		33		70		97		91	66	8
P12					70				60				38				97			91	
Q06			0		1		1		0		0		0		2		3		4	2	0
Q12					1				1				0				3			4	
T06		1/	0	0/	0	0/	0	0/	0	0/	1	0/	0	1/	0	0/	0	1/	0	1/	3
T12						2/	0			1/	5			1/	0			3/	0	3/	8
POZ	6	11	7	6	3	1	2	2	3	1	0	0	1	0	2	3	2	1	0	0	2
POS	21	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
TYP	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R
SNW									0								0				
CIG	4	4	3	3	3	2	4	6	8	8	7	7	6	4	4	2	2	3	2	8	8
VIS	7	7	5	3	2	3	2	6	5	7	7	7	7	5	4	4	3	3	3	5	7
OBV	N	N	BR	BR	BR	BR	FG	BR	BR	N	N	N	N	BR	BR	BR	BR	BR	BR	BR	N

A variety of weather elements are forecast, at lead times up to 60 h and at intervals as close as 3 h.

The forecast equations for all elements except temperatures, dew points, and winds are "regionalized." That is, developmental data from groups of nearby and climatically similar stations were composited in order to increase the sample size when deriving the forecast equations. For each regional group, then, forecasts are made with the same equations and the same regression coefficients. This does not mean that the forecasts for all the stations in the group are the same, however, since interpolation of the dynamical output to the different forecast locations yields different predictor values. Some of the MOS equations also contain predictors representing local climatological values, which introduces further differences in the forecasts for the different stations. Regionalization is especially valuable for producing good forecasts for relatively rare events.

In order to enhance consistency among the forecasts for different but related weather elements, some of the MOS equations are developed "simultaneously." This means that the same predictor variables, although with different regression coefficients, are forced into prediction equations for related predictands in order to enhance the consistency of the forecasts. This is an instance of *multivariate multiple regression* (e.g., Johnson and Wichern, 2007). For example, it would be physically unreasonable and clearly undesirable for the forecast dew point to be higher than the forecast temperature. To help ensure that such inconsistencies appear in the forecasts as rarely as possible, the MOS equations for maximum temperature, minimum temperature, and the 3-h temperatures and dew points all contain the same

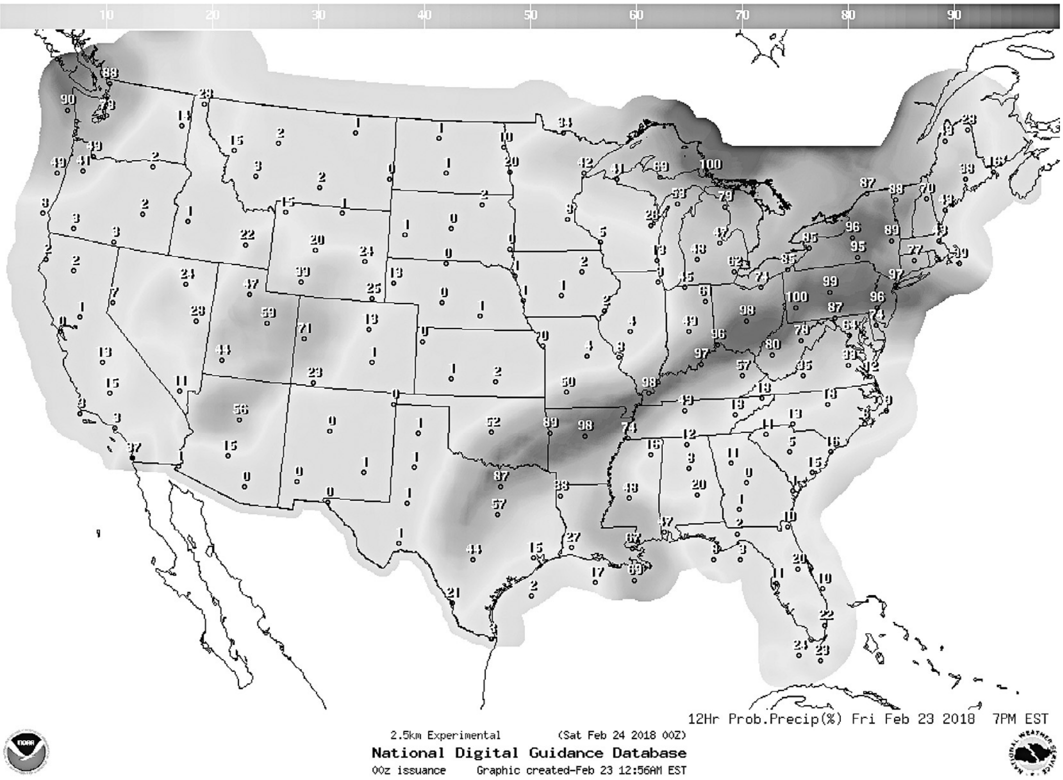

10
20
30
40
50
60
70
80
90
12Hr Prob.Precip(%) Fri Feb 23 2018 7PM EST
2.5km Experimental (Sat Feb 24 2018 00Z)
National Digital Guidance Database
00z issuance Graphic created-Feb 23 12:56AM EST

FIGURE 7.29 Example MOS forecasts in map form. The predictand is the probability (× 100) of at least 0.01 in. of precipitation during the 12-h period 7 AM-7 PM EST, 23 February 2018. *From sats.nws.noaa.gov/~mos/gmos/conus25/view_maps_js.php.*

predictor variables. Similarly, the four groups of forecast equations for wind speeds and directions, the 6- and 12-h precipitation probabilities, the 6- and 12-h thunderstorm probabilities, and the probabilities for precipitation types, were also developed simultaneously to enhance their consistency.

Because MOS forecasts are made for a large number of locations, it is possible to view them as maps, which are also posted on the internet. Some of these maps display selected quantities from the MOS panels such as the one presented in Table 7.8. Figure 7.29 shows a forecast map for the variable listed as "P12" in Table 7.8, meaning probabilities of at least 0.01 in. of (liquid-equivalent) precipitation, accumulated over a 12-h period.

7.10. SUBJECTIVE PROBABILITY FORECASTS

7.10.1. The Nature of Subjective Forecasts

Most of this chapter has dealt with objective forecasts, or forecasts produced by means that are automatic. Objective forecasts are determined unambiguously by the nature of the forecasting procedure and the values of the variables that are used to drive it. However, objective forecasting procedures necessarily rest on a number of subjective judgments made during their development. Nevertheless, some people feel more secure with the results of objective forecasting procedures, seemingly taking comfort from their lack of contamination by the vagaries of human judgment. Apparently, such individuals feel that objective forecasts are in some way less uncertain than human-mediated forecasts.

One very important—and perhaps irreplaceable—role of human forecasters in the forecasting process is in the subjective integration and interpretation of objective forecast information. These objective forecast products often are called forecast guidance, and include deterministic forecast information from dynamical integrations, and statistical guidance from MOS systems or other interpretive statistical products. Human forecasters also use, and incorporate into their judgments, available atmospheric observations (surface maps, radar images, etc.), and prior information ranging from persistence or simple climatological statistics, to their individual previous experiences with similar meteorological situations. The result is (or should be) a forecast reflecting, to the maximum practical extent, the forecaster's state of knowledge about the future evolution of the atmosphere.

Human forecasters can rarely, if ever, fully describe or quantify their personal forecasting processes (Stuart et al., 2007). Thus the distillation by a human forecaster of disparate and sometimes conflicting information is known as *subjective* forecasting. A subjective forecast is one formulated on the basis of the judgment of one or more individuals. Making a subjective weather forecast is a challenging process precisely because future states of the atmosphere are inherently uncertain. The uncertainty will be larger or smaller in different circumstances—some forecasting situations are more difficult than others—but it will never really be absent. Doswell (2004) provides some informed perspectives on the formation of subjective judgments in weather forecasting.

Since the future states of the atmosphere are inherently uncertain, a key element of a good and complete subjective weather forecast is the reporting of some measure of the forecaster's uncertainty. This point has been recognized since at least the 19th century (Murphy, 1998). It is the forecaster who is most familiar with the atmospheric situation, and it is therefore the forecaster who is in the best position to evaluate the uncertainty associated with a given forecasting situation. Although it is common for nonprobabilistic forecasts (i.e., forecasts containing no expression of uncertainty) to be issued, such as "tomorrow's maximum temperature will be 27°F," an individual issuing this forecast would not seriously expect the temperature to be exactly 27°F. Given a forecast of 27°F, temperatures of 26°F or 28°F would generally be regarded as nearly as likely, and in this situation the forecaster would usually not really be surprised to see tomorrow's maximum temperature anywhere between 25° and 30°F.

Although uncertainty about future weather can be reported verbally using phrases such as "chance" or "likely," such qualitative descriptions are open to different interpretations by different people (e.g., Murphy and Brown, 1983). Even worse, however, is the fact that such qualitative descriptions do not precisely reflect the forecaster's uncertainty about, or degree of belief in, the future weather. The forecaster's state of knowledge is most accurately reported, and the needs of the forecast user are best served, if the intrinsic uncertainty is quantified in probability terms. Thus the Bayesian view of probability as the degree of belief of an individual holds a central place in subjective forecasting. Note that since different forecasters have somewhat different information on which to base their judgments (e.g., different sets of experiences with similar past forecasting situations), it is perfectly reasonable to expect that their probability judgments may differ somewhat as well.

7.10.2. The Subjective Distribution

Before a forecaster reports a subjective degree of uncertainty as part of a forecast, he or she needs to have a mental image of that uncertainty. The information about an individual's uncertainty can be thought of as residing in their *subjective distribution* for the event in question. The subjective distribution is a probability distribution in the same sense as the parametric distributions described in Chapter 4. Sometimes, in fact, one of the distributions specifically described in Chapter 4 may provide a very good approximation to an individual's subjective distribution. Subjective distributions are interpreted from a Bayesian perspective as the quantification of an individual's degree of belief in each of the possible outcomes for the variable being forecast.

Each time a forecaster prepares to make a forecast, he or she internally develops a subjective distribution. The possible weather outcomes are subjectively weighed, and an internal judgment is formed as to their relative likelihoods. This process occurs whether or not the forecast is to be a probability forecast, or indeed whether or not the forecaster is even consciously aware of the process. However, unless we believe that uncertainty can somehow be expunged from the process of weather forecasting, it should be clear that better forecasts will result when forecasters think explicitly about their subjective distributions and the uncertainty that those distributions describe.

It is easiest to approach the concept of subjective probabilities with a familiar but simple example. Subjective probability-of-precipitation (PoP) forecasts have been routinely issued in the United States since 1965. These forecasts specify the probability that measurable precipitation (i.e., at least 0.01 in.) will occur at a particular location during a specified time period. The forecaster's subjective distribution for this event is so simple that we might not notice that it is a probability distribution. However, the events "precipitation" and "no precipitation" divide the sample space into two MECE events. The distribution of probability over these events is discrete and consists of two complementary elements: one probability for the event "precipitation" and another probability for the event "no precipitation," and either one of these two probabilities can be easily computed from the other. This distribution will be different for different forecasting situations, and perhaps for different forecasters assessing the same situation. However, the only thing about a forecaster's subjective distribution for the PoP that can change from one forecasting occasion to another is the probability, and this will be different to the extent that the forecaster's degree of belief regarding the future precipitation occurrence is different. The PoP ultimately issued by the forecaster should be the forecaster's subjective probability for the event "precipitation," or perhaps a suitably rounded version of that probability. That is, it is the forecaster's job to evaluate the uncertainty associated with the possibility of future precipitation occurrence and to report that uncertainty to the users of the forecasts.

7.10.3. Central Credible Interval Forecasts

It has been argued here that inclusion of some measure of the forecaster's uncertainty should be included in any weather forecast. Forecast users can use the added uncertainty information to make better, economically more favorable, decisions (e.g., Roulston et al., 2006). Historically, resistance to the idea of probability forecasting has been based in part on the practical consideration that the forecast format should be compact and easily understandable. In the case of PoP forecasts, the subjective distribution is sufficiently simple that it can be reported with a single number, and is no more cumbersome than issuing a nonprobabilistic forecast of "precipitation" or "no precipitation." When the subjective distribution is continuous, however, some approach to sketching its main features is a practical necessity if its probability information is to be conveyed succinctly in a publicly issued forecast. Discretizing a continuous subjective distribution is one approach to simplifying it in terms of one or a few easily expressible quantities. Alternatively, if the forecaster's subjective distribution on a given occasion can be reasonably well approximated by one of the parametric distributions described in Chapter 4, another approach to simplifying its communication could be to report the parameters of the approximating distribution. There is no guarantee, however, that subjective distributions will always (or even ever) correspond to a familiar parametric form.

One very attractive and workable alternative for introducing probability information into forecasts for continuous meteorological variables is the use of *credible interval forecasts*. This forecast format has been used operationally in Sweden (Ivarsson et al., 1986), but to date has been used only experimentally in the United States (Murphy and Winkler, 1974; Peterson et al., 1972; Winkler and Murphy, 1979). In unrestricted form, a credible interval forecast requires specification of three quantities: two points defining an interval for the continuous forecast variable, and a probability (according to the forecaster's subjective distribution) that the forecast quantity will fall in the designated interval. Usually the requirement is also made that the credible interval be located in the middle of the subjective distribution. In this case the specified probability is distributed equally on either side of the subjective median, and the forecast is called a *central credible interval* forecast.

There are two special cases of the central credible interval forecast format, each requiring that only two quantities be communicated. The first is the fixed-width central credible interval forecast. As the name implies, the width of the central credible interval is the same for all forecasting situations and is specified in advance for each predictand. Thus the forecast includes a location for the interval, generally specified as its midpoint, and a probability that the outcome will occur in the forecast interval. For example, the Swedish central credible interval forecasts for temperature are of the fixed-width type, with the interval size specified to be ±3 °C around the midpoint temperature. These forecasts thus include a forecast temperature, together with a probability that the subsequently observed temperature will be within 3 °C of the forecast temperature. The two forecasts 15 °C, 90% and 15 °C, 60% would both indicate that the forecaster expects the temperature to be about 15 °C, but the inclusion of

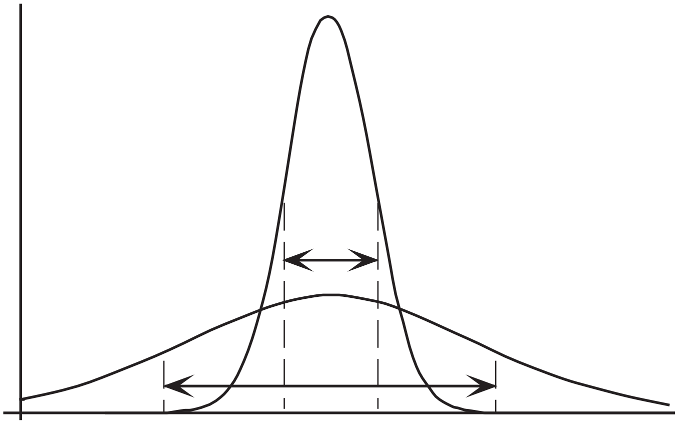


FIGURE 7.30 Two hypothetical subjective distributions shown as probability density functions. The two distributions have the same mean, but reflect different degrees of uncertainty. The tall, narrow distribution represents an easier (less uncertain) forecasting situation, and the broader distribution represents a more difficult forecast problem. Arrows delineate 75% central credible intervals in each case.

probabilities in the forecasts shows that much more confidence can be placed in the former as opposed to the latter of the two forecasts of 15 °C. Because the forecast interval is central, these two forecasts would also imply 5%, and 20% chances, respectively, for the temperature to be colder than 12° or warmer than 18°.

Some forecast users would find the unfamiliar juxtaposition of a temperature and a probability in a fixed-width central credible interval forecast to be somewhat jarring. The fixed-probability central credible interval forecast is an alternative forecast format that could be implemented more subtly. In this format, it is the probability contained in the forecast interval, rather than the width of the interval, that is specified in advance and is constant from forecast to forecast. This format makes the probability component of the credible interval forecast implicit, so the forecast consists of two numbers having the same physical dimensions as the quantity being forecast.

Figure 7.30 illustrates the relationship of 75% central credible intervals for two subjective distributions having the same mean. The shorter, broader distribution represents a relatively uncertain forecasting situation, where events fairly far away from the center of the distribution are regarded as having substantial probability. A relatively wide interval is therefore required to subsume 75% of this distribution's probability. On the other hand, the tall and narrow distribution describes considerably less uncertainty, and a much narrower forecast interval contains 75% of its density. If the variable being forecast is temperature, the 75% central credible interval forecasts for these two cases might be 10° to 20°, and 14° to 16°, respectively.

A strong case can be made for operational credible-interval forecasts (Grounds et al. 2017; Murphy and Winkler, 1974, 1979). Since nonprobabilistic temperature forecasts are already often specified as ranges, fixed-probability central credible interval forecasts could be introduced into forecasting operations quite unobtrusively. Forecast users not wishing to take advantage of the implicit probability information would notice little difference from the present forecast format, whereas those understanding the meaning of the forecast ranges would derive additional benefit. Even forecast users unaware that the forecast range is meant to define a particular interval of fixed probability might notice over time that the interval widths were related to the precision of the forecasts.

7.10.4. Assessing Discrete Probabilities

Experienced weather forecasters are able to formulate subjective probability forecasts that evidently quantify their uncertainty regarding future weather quite successfully. Examination of the error characteristics of such forecasts (see Chapter 9) reveals that they are largely free of the biases and inconsistencies sometimes exhibited in the subjective probability assessments made by less experienced individuals. Commonly, inexperienced forecasters produce probability forecasts exhibiting overconfidence (Murphy, 1985), or biases due to such factors as excessive reliance on recently acquired information (Spetzler and Staël von Holstein, 1975; Tversky, 1974).

Individuals who are experienced at assessing their subjective probabilities can do so in a seemingly subconscious or automatic manner. People who are new to the practice often find it helpful to use physical or conceptual devices that allow comparison of the uncertainty to be assessed with a situation that is more concrete and familiar (Garthwaite et al., 2005). For example, Spetzler and Staël von Holstein (1975) describe a physical device called a probability wheel, which consists of a spinner of the sort that might be found in a child's board game, on a background that has the form of a pie chart. This background has two colors, blue and orange, and the proportion of the background covered by each of the colors can be adjusted. The probability wheel is used to assess the probability of a dichotomous event (e.g., a PoP forecast) by adjusting the relative coverages of the two colors until the forecaster feels the probability of the event to be forecast is about equal to the probability of the spinner stopping in the orange sector. The subjective probability forecast is then read as the angle subtended by the orange sector, divided by 360°.

Conceptual devices can also be employed to assess subjective probabilities. For many people, comparison of the uncertainty surrounding the future weather is most easily assessed in the context of lottery games or betting games. Such conceptual devices translate the probability of an event to be forecast into more concrete terms by posing hypothetical questions such as "would you prefer to be given \$2 if precipitation occurs tomorrow, or \$1 for sure (regardless of whether or not precipitation occurs)?" Individuals preferring the sure \$1 in this lottery situation evidently feel that the relevant PoP is <0.5, whereas individuals who feel the PoP is >0.5 would generally prefer to receive \$2 on the chance of precipitation. A forecaster can use this lottery device by adjusting the variable payoff relative to the certainty equivalent (the sum to be received for sure) until the point of indifference, where either choice would be equally attractive. That is, the variable payoff is adjusted until the expected (i.e., probability-weighted average) payment is equal to the certainty equivalent. Denoting the subjective probability as p, the procedure can be written formally as

$$\text{Expected payoff} = p\,(\text{Variable payoff}) + (1-p)(\$0) = \text{Certainty equivalent} \tag{7.65a}$$

which leads to

$$p = \frac{\text{Certainty equivalent}}{\text{Variable payoff}}. \tag{7.65b}$$

The same kind of logic can be applied in an imagined betting situation. Here the forecasters ask themselves whether receiving a specified payment should the weather event to be forecast occurs, or suffering some other monetary loss if the event does not occur, is preferable. In this case the subjective probability is assessed by finding monetary amounts for the payment and loss such that the bet is a fair one, implying that the forecaster would be equally happy to be on either side of it. Since the expected payoff from a fair bet is zero, the betting game situation can be represented as

$$\text{Expected payoff} = p\,(\$\text{payoff}) + (1-p)(-\$\text{loss}) = 0, \tag{7.66a}$$

leading to

$$p = \frac{\$\text{loss}}{\$\text{loss} + \$\text{payoff}}. \tag{7.66b}$$

Many betting people think in terms of odds in this context. Equation 7.66a can be expressed alternatively as

$$\text{odds ratio} = \frac{p}{1-p} = \frac{\$\text{loss}}{\$\text{payoff}}. \tag{7.67}$$

Thus a forecaster being indifferent to an even-money bet (1:1 odds) harbors an internal subjective probability of $p = 0.5$. Indifference to being on either side of a 2:1 bet implies a subjective probability of 2/3, and indifference at 1:2 odds is consistent with an internal probability of 1/3.

The same kind of thinking as these lottery and betting games for individual probability elicitation can also be applied to probability evaluation through “prediction markets” (Wolfers and Zitzewitz, 2004). However, in the case of prediction markets the tacit probability assessments are made through a *consensus forecasting* process, where the judgments of multiple individuals are aggregated. Experience in meteorological contexts (Baars and Mass, 2005; Hamill and Wilks, 1995; Sanders, 1963; Thompson, 1977; Vislocky and Fritsch, 1995a) has shown that such consensus forecasts typically improve on the performance of the individual forecasts that are combined into the consensus.

7.10.5. Assessing Continuous Distributions

The same kinds of lotteries or betting games just described can also be used to assess quantiles of a subjective continuous probability distribution using the *method of successive subdivision*. Here the approach is to identify quantiles of the subjective distribution by comparing event probabilities that they imply with the reference probabilities derived from conceptual money games. Use of this method in an operational setting is described in Krzysztofowicz et al. (1993).

The easiest quantile to identify is the median. Suppose the distribution to be identified pertains to tomorrow’s maximum temperature. Since the median divides the subjective distribution into two equally probable halves, its location can be assessed by evaluating a preference between, say, \$1 for sure and \$2 if tomorrow’s maximum temperature is warmer than 14 °C. The situation is the same as that described in Equation 7.65. Preferring the certainty of \$1 implies a subjective probability for the event {maximum temperature warmer than 14 °C} that is smaller than 0.5. A forecaster preferring the chance at \$2 evidently feels that the probability for this event is larger than 0.5. Since the cumulative probability, p, for the median is fixed at 0.5, we can locate the threshold defining the event {outcome above median} by adjusting it to the point of indifference between the certainty equivalent and a variable payoff equal to twice the certainty equivalent.

The quartiles can be assessed in the same way, except that the ratios of certainty equivalent to variable payoff must correspond to the cumulative probabilities of the quartiles, that is, 1/4 or 3/4. At what temperature T_{LQ} are we indifferent to the alternatives of receiving \$1 for sure, or \$4 if tomorrow's maximum temperature is below T_{LQ}? The temperature T_{LQ} then estimates the forecaster's subjective lower quartile. Similarly, the temperature T_{UQ}, at which we are indifferent to the alternatives of \$1 for sure or \$4 if the temperature is above T_{UQ}, estimates the upper quartile.

Especially when someone is inexperienced at probability assessments, it is a good idea to perform some consistency checks. In the method just described, the quartiles were assessed independently, but together define a range—the 50% central credible interval—in which half the probability should lie. Therefore a good check on their consistency would be to verify that we are indifferent to the choices between \$1 for sure, and \$2 if $T_{LQ} \leq T \leq T_{UQ}$. If we prefer the certainty equivalent in this comparison the quartile estimates T_{LQ} and T_{UQ} are apparently too close. If we prefer the chance at the \$2 they apparently subtend too much probability. Similarly, we could verify indifference between the certainty equivalent, and four times the certainty equivalent if the temperature falls between the median and one of the quartiles. Any inconsistencies discovered in checks of this type indicate that some or all of the previously estimated quantiles need to be reassessed.

7.11. EXERCISES

7.1 a. Derive a simple linear regression equation using the data in Table A.3, relating June temperature (as the predictand) to June pressure (as the predictor).
b. Explain the physical meanings of the two parameters.
c. Formally test whether the fitted slope is significantly different from zero.
d. Compute the R^2 statistic.
e. Estimate the probability that a predicted value corresponding to $x_0 = 1013$ mb will be within 1 °C of the regression line, using Equation 7.23.
f. Repeat (e), assuming the prediction variance equals the MSE.

7.2 Consider the following partial ANOVA table, describing the results of a regression analysis:

Source	df	SS	MS
Total	26	318.2874	
Regression	_____	316.6065	_____
Residual	25	_____	_____

a. Fill in the 4 blanks in the table.
b. How many predictor variables are in the equation?
c. What is the sample variance of the predictand?
d. What is the R^2 value?
e. Estimate the probability that a prediction made by this regression will be within ±0.2 units of the actual value.
f. Formally test whether the estimated slope $b = 0.69$ is significantly different from zero.

7.3 Derive an expression for the maximum likelihood estimate of the intercept b_0 in logistic regression (Equation 7.36), for the constant model in which $b_1 = b_2 = \ldots = b_K = 0$.

7.4 The 19 nonmissing precipitation values in Table A.3 can be used to fit the regression equation:

$$\ln[(\text{Precipitation}) + 1\,\text{mm}] = 499.4 – 0.512\,(\text{Pressure}) + 0.796\,(\text{Temperature})$$

The MSE for this regression is 0.701. (The constant 1 mm has been added to ensure that the logarithm is defined for all data values.)

a. Estimate the missing precipitation value for 1956 using this equation.

b. Construct a 95% prediction interval for the estimated 1956 precipitation.

7.5 Explain how to use cross-validation to estimate the prediction mean squared error, and the sampling distribution of the regression slope, for the problem in Exercise 7.1. If the appropriate computing resources are available, implement your algorithm.

7.6 Hurricane Zeke is an extremely late storm in a very busy hurricane season. It has recently formed in the Caribbean, the 500 mb height at gridpoint 37 (relative to the storm) is 5400 m, the 500 mb height at gridpoint 3 is 5500 m, and the 1000 mb height at gridpoint 51 is –200 m (i.e., the surface pressure near the storm is well below 1000 mb).

a. Use the NHC 67 model (see Table 7.7) to forecast the westward component of its movement over the next 12 h, if storm has moved 80 n.mi. due westward in the previous 12 hours.

b. What would the NHC 67 forecast of the westward displacement be if, in the previous 12 hours, the storm had moved 80 n.mi. westward *and* 30 n.mi. northward (i.e., $P_y = 30$ n. mi.)?

7.7 The fall (September, October, November) MOS equation for predicting maximum temperature (in °F) at Binghamton, New York, formerly used with a now-discontinued dynamical model, at the 60-h lead time was

$$\text{MAX } T = -363.2 + 1.541\,(850\,\text{mb T}) - .1332\,(\text{SFC} - 490\,\text{mb RH}) - 10.3\,(\text{COS DOY})$$

where:

(850 mb T) is the 48-h dynamical forecast of temperature (K) at 850 mb
(SFC-490 mb RH) is the 48-h forecast lower tropospheric RH in %
(COS DOY) is the cosine of the day of the year transformed to radians or degrees, that is, $= \cos(2\pi t/365)$ or $= \cos(360°\, t/365)$
and t is the day number of the valid time (the day number for January 1 is 1, and for October 31 it is 304)

Calculate what the 60-h MOS maximum temperature forecast would be for the following:

	Valid time	48-h 850 mb *T* fcst	48-h mean RH fcst
a.	September 4	278 K	30%
b.	November 28	278 K	30%
c.	November 28	258 K	30%
d.	November 28	278 K	90%

7.8 A MOS equation for 12–24 h PoP in the warm season might look something like

PoP = 0.25 + 0.0063(Mean RH) – 0.163(0–12 ppt [bin @ 0.1 in.]) – 0.165(Mean RH [bin @ 70%]).

where:

Mean RH (%) is the same variable as in Exercise 7.7 for the appropriate lead time

0–12 ppt is the model-forecast precipitation amount in the first 12 h of the forecast

[bin @ xxx] indicates use as a binary variable: =1 if the predictor is ≤ xxx

= 0 otherwise

Evaluate the MOS PoP forecasts for the following conditions:

	12-h mean RH	0–12 ppt
a.	90%	0.00 in.
b.	65%	0.15 in.
c.	75%	0.15 in.
d.	75%	0.09 in.

7.9 Explain why the slopes of the solid lines decrease, from Figure 7.27 to Figure 7.28a, to Figure 7.28b. What would the corresponding MOS equation be for an arbitrarily long lead time into the future?

7.10 A forecaster is equally happy with the prospect of receiving $1 for sure, or $5 if freezing temperatures occur on the following night. What is the forecaster's subjective probability for frost?

7.11 A forecaster is indifferent between receiving $1 for sure and any of the following: $8 if tomorrow's rainfall is >55 mm, $4 if tomorrow's rainfall is >32 mm, $2 if tomorrow's rainfall is >12 mm, $1.33 if tomorrow's rainfall is >5 mm, and $1.14 if tomorrow's precipitation is >1 mm.

a. What is the median of this individual's subjective distribution?

b. What would be a consistent 50% central credible interval forecast? A 75% central credible interval forecast?

c. In this forecaster's view, what is the probability of receiving more than one but no more than 32 mm of precipitation?