CHAPTER 5

Inferential Statistics and Predictive Analytics

Inferential statistics draws valid inferences about a population based on an analysis of a representative sample of that population. The results of such an analysis are generalized to the larger population from which the sample originates, in order to make assumptions or predictions about the population in general. This chapter introduces linear, logistics, and polynomial regression analyses for inferential statistics. The result of a regression analysis on a sample is a predictive model in the form of a set of equations.

The first task of sample analysis is to make sure that the chosen sample is representative of the population as a whole. We have previously discussed the one-way chi-square goodness-of-fit test for such a task by comparing the sample distribution with an expected distribution. Here we present the chi-square two-way test of independence to determine whether significant differences exist between the distributions in two or more categories. This test helps to determine whether a candidate independent variable in a regression analysis is a true candidate predictor of the dependent variable, and to thus exclude irrelevant variables from consideration in the process.

We also generalize traditional regression analyses to Bayesian regression analyses, where the regression is undertaken within the context of the Bayesian inference. We present the most general Bayesian regression analysis, known as the Gaussian process. Given its similarity to other decision tree learning techniques, we save discussion of the Classification and Regression Tree (CART) technique for the later chapter on ML.

To use inferential statistics to infer latent concepts and variables and their relationships, this chapter includes a detailed description of principal component and factor analyses. To use inferential statistics for forecasting by modeling time series data, we present survival analysis and autoregression techniques. Later in the book we devote a full chapter to AI- and ML-oriented

techniques for modeling and forecasting from time series data, including dynamic Bayesian networks and Kalman filtering.

5.1 CHI-SQUARE TEST OF INDEPENDENCE

The one-way Chi-Square (χ^2) goodness-of-fit test (which was introduced earlier in the descriptive analytics chapter) is a non-parametric test used to decide whether distributions of categorical variables differ significantly from predicted values. The two-way or two-sample chi-square test of independence is used to determine whether a significant difference exists between the distributions of two or more categorical variables. To determine if Outlook is a good predictor of Decision in our play-tennis example in Appendix B, for instance, the null hypothesis H0 is that two distributions are not equal; in other words, that the weather does not affect if one decides to play or not. The Outlook vs. Decision table is shown below in TABLE 5.1. Note that the row and the column subtotals must have equal sums, and that total expected frequencies must equal total observed frequencies.

TABLE 5.1: : Outlook vs. Decision table

Outlook	Decision play	Decision don't play	Row Subtotal
sunny	2	3	5
overcast	4	0	4
rain	3	2	5
Column Subtotal	9	5	Total = 14

Note also that we are computing expectation as follows with a view that the observations are assumed to be representative of the past

$$
\begin{aligned}
&Exp\left(Outlook = sunny \;\&\; Decision = play\right)\\
&\quad = 14 \times p\left(Outlook = sunny \;\&\; Decision = play\right)\\
&\quad = 14 \times p\left(Outlook = sunny\right) \times p\left(Decision = play\right)\\
&\quad = 14 \times \left(\sum_{Decision} \left(p\left(Outlook = sunny\right) \times p\left(Decision\right)\right)\right) \times\\
&\qquad \left(\sum_{Outlook} \left(p\left(Decision = play\right) \times p\left(Outlook\right)\right)\right)\\
&\quad = 14 \times \left(p\left(sunny\right) \times p\left(play\right) + p\left(sunny\right) \times p\left(don't\, play\right)\right) \times\\
&\qquad \left(p\left(play\right) \times p\left(sunny\right) + p\left(play\right) \times p\left(overcast\right) + p\left(play\right) \times p\left(rain\right)\right)\\
&\quad = 14 \times (\text{Row subtotal for } sunny/14) \times (\text{Column subtotal for } play/14)\\
&\quad = \left(5 \times 9\right)/14
\end{aligned}
$$

The computation of Chi-square statistic is shown in TABLE 5.2.

TABLE 5.2: : Computation of Chi-square statistic

Joint Variable	Observed (O)	Expected (E)	$(O-E)^2/E$
sunny & play	2	3.21	0.39
sunny & don't play	3	1.79	0.82
overcast & play	4	2.57	0.79
overcast & don't play	0	1.43	1.43
rainy & play	3	3.21	0.01
rainy & don't play	2	1.79	0.02

Therefore, Chi-square statistic = $\sum_i \frac{(O-E)^2}{E} = 3.46$

The degree of freedom is $(3-1) \times (2-1)$, that is, 2. With 95% as the level of significance, the critical value from the Chi-square table is 5.99. Since the value 3.46 is less than 5.99, so we would reject the null hypothesis that there is significant difference between the distributions in Outlook and Decision. Hence the weather does affect if one decides to play or not.

5.2 REGRESSION ANALYSES

In this section, we begin with simple and multiple linear regression techniques, then present logistic regression for handling categorical variables as the dependent variables, and, finally, discuss polynomial regression for modeling nonlinearity in data.

5.2.1 Simple Linear Regression

Simple linear regression models the relationship between two variables X and Y by fitting a linear equation to observed data:

$$Y = a + bX$$

where X is called an *explanatory variable* and Y is called a *dependent variable*. The *slope* b and the *intercept* a in the above equation must be estimated from a given set of observations. "Least-squares" is the most common method for fitting equations, wherein the best-fitting line for the observed data is calculated by minimizing the sum of the squares of the vertical deviations from each data point to the line.

Suppose the set $(y_1, x_1), ..., (y_n, x_n)$ of n observations are given. The expression to be minimized is the sum of the squares of the residuals (i.e., the differences between the observed and predicted values):

$$\sum_{i=1}^{n} (y_i - a - bx_i)^2$$

By solving the two equations obtained by taking partial derivatives of the

above expression with respect to a and b and then equating them to zero, the estimations of a and b can be obtained.

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{j=1}^{n} y_j}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} = \frac{Cov(X,Y)}{Var(X)}$$

$$\hat{a} = \bar{Y} - \hat{b}\bar{X}$$

The plot in FIGURE 5.1 shows the observations and linear regression model (the straight line) of the two variables Temperature (Fahrenheit degree) and Humidity (%), with Temperature as the dependent variable. For any given observation of Humidity, the difference between the observed and predicted value of Temperature provides the residual error.

Temperature (°F)	Humidity (%)
75	70
80	90
85	85
72	95
69	70
72	90
83	78
64	65
81	75
71	80
65	70
75	80
68	80
70	96

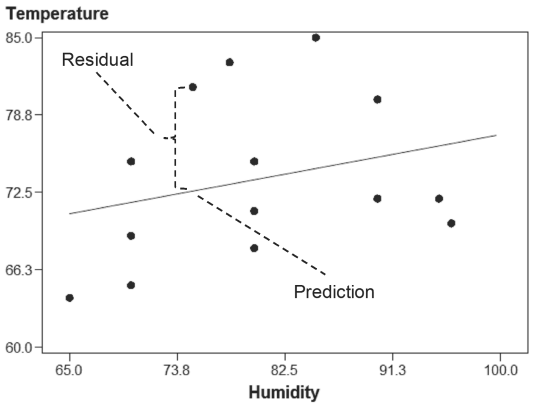


FIGURE 5.1 : Example linear regression

The correlation coefficient measure between the observed and predicted values can be used to determine how close the residuals are to the regression line.

5.2.2 Multiple Linear Regression

Multiple linear regression models the relationship between two or more response variables X_i and one dependent variable Y as follows:

$$Y = a + b_1 X_1 + ... + b_p X_p$$

The given n observations $(y_1, x_{11}, ..., x_{1p}), ..., (y_n, x_{n1}, ..., x_{np})$ in matrix form are

$$\begin{bmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{bmatrix} = \begin{bmatrix} a \\ a \\ ... \\ a \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ ... \\ b_p \end{bmatrix}^T \begin{bmatrix} x_{11} & x_{21} & ... & x_{n1} \\ x_{12} & x_{22} & ... & x_{n2} \\ ... & ... & ... & ... \\ x_{1p} & x_{2p} & ... & x_{np} \end{bmatrix}$$

Or in abbreviated form

$$\mathbf{Y}{=}A{+}B^T\mathbf{X}$$

The expression to be minimized is

$$\sum_{i=1}^{n} (y_i - a - b_1 x_{i1} - ... - b_p x_{ip})^2$$

The estimates of A and B are as follows:

$$\hat{B} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y} = \frac{Cov(\mathbf{X},\mathbf{Y})}{Var(\mathbf{X})}$$
$$\hat{A} = \bar{\mathbf{Y}} - \hat{B}\bar{\mathbf{X}}$$

5.2.3 Logistic Regression

The dependent variable in logistic regression is binary. In order to predict categorical attribute Decision in the play-tennis example in Appendix B from a new category Temperature, suppose the attribute Temp_0_1 represents a continuous version of the attribute Decision, with 0 and 1 representing the values "don't play" and "play" respectively. FIGURE 5.2 shows the scatter plot and a line plot of Temperature vs. Temp_0_1 (left), and a scatter plot and logistic curve for the same (right). The scatter plot shows that there is a fluctuation among the observed values, in the sense that for a given Temperature (say, 72), the value of the dependent variable (play/don't play) has been observed to be both 0 and 1 on two different occasions. Consequently, the line plot oscillates between 0 and 1 around that temperature. On the other hand, the logistic curve transitions smoothly from 0 to 1. We describe here briefly how logistic regression is formalized.

Since the value of the dependent variable is either 0 or 1, the most intuitive way to apply linear regression would be to think of the response as a probability value. The prediction will fall into one class or the other if the response crosses a certain threshold or not, and therefore the linear equation will be of the form:

$$p\left(Y = 1|X\right) = a + bX$$

However, the value of $a{+}bX$ could be > 1 or < 0 for some X, giving probabilities that cannot exist. The solution is to use a different probability representation. Consider the following equation with a ratio as the response variable:

$$\frac{p}{1-p} = a + bX$$

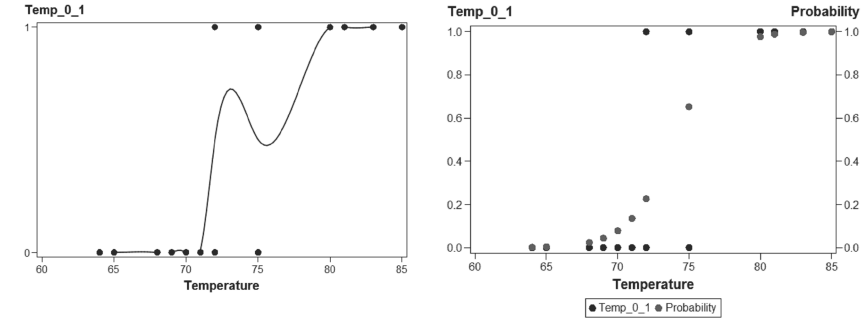


FIGURE 5.2 : (left) Scatter and line plots of Temperature vs. Temp_0_1, and (right) scatter plot and logistic curve for the same

The ratio ranges from 0 to ∞ for some X but the value of $a + bX$ would be below 0 for some X. The solution is to take the log of the ratio:

$$\log\left(\frac{p}{1-p}\right) = a + bX$$

The logit function above transforms a probability statement defined in $0 < p < 1$ to one defined in $-\infty < a + bX < \infty$. The value of p is as follows:

$$p = \frac{e^{a+bX}}{1 + e^{a+bX}}$$

A maximum likelihood method can be applied to estimate the parameters a and b of a logistic model with binary response:

- $Y = 1$ with probability p
- $Y = 0$ with probability $1 - p$

For each $Y_i = 1$ the probability p_i appears in the likelihood product. Similarly, for each $Y_i = 0$ the probability $1 - p_i$ appears in the product. Thus, the likelihood of the sample $(y_1, x_1), ..., (y_n, x_n)$ of n observations takes the following form:

$$\begin{aligned} & L(a, b; (y_1, x_1), ..., (y_n, x_n)) \\ &= \prod_{i=1}^{n} p^{y_i} (1-p)^{1-y_i} \\ &= \prod_{i=1}^{n} \left(\frac{e^{a+bx_i}}{1+e^{a+bx_i}}\right)^{y_i} \left(\frac{1}{1+e^{a+bx_i}}\right)^{1-y_i} \\ &= \prod_{i=1}^{n} \frac{\left(e^{a+bx_i}\right)^{y_i}}{1+e^{a+bx_i}} \end{aligned}$$

Alternatively, we can maximize the log likelihood, $\log(L(a, b; Data))$, to solve

for a and b. In the example above, the two values of a and b that maximize the likelihood are 45.94 and -0.62, respectively. Hence the logistic equation is:

$$\log\left(\frac{p}{1-p}\right) = 45.94 - 0.62X$$

FIGURE 5.2 (right) shows the logistics curve plotted for the sample.

5.2.4 Polynomial Regression

The regression model

$$Y = a_0 + a_1X + a_2X^2 + ... + a_kX^k$$

is called the k-th order polynomial model with one variable, where a_0 is the Y-intercept of X, a_1 is called the *linear effect parameter*, a_2 is called the *quadratic effect parameter*, and so on. The model is a linear regression model for $k = 1$. This model is non-linear in the X variable, but it is linear in the parameters a_0, a_1, a_2, ... and a_k. One can also have higher-order polynomial regression involving more than one variable. For example, a second-order or quadratic polynomial regression with two variables X_1 and X_2 is

$$Y = a_0 + a_1X_1 + a_2X_2 + a_{11}X_1^2 + a_{22}X_2^2 + a_{12}X_1X_2$$

FIGURE 5.3 (left) shows a plot of 7 noisy, evenly-spaced random training samples $(x_1, y_1), ..., (x_7, y_7)$ drawn from an underlying function f (shown in dotted line). Note that a dataset of $k+1$ observations can be modeled perfectly by a polynomial of degree k. In this case, a six-degree polynomial will fit the data perfectly, but will "over-fit" the data and will not generalize well for test data. To decide on the appropriate degree for a polynomial regression model, one can begin with a linear model and include higher-order terms one by one until the highest-order term becomes non-significant (determined by looking at p-values for the t-test for slopes). One could also start with a high-order model and exclude the non-significant highest-order terms one by one until the remaining highest-order term becomes significant. Here we measure fitness with respect to the test data of eight evenly-spaced samples as shown plotted in FIGURE 5.3 (right).

FIGURE 5.4 shows six different polynomial regressions of degrees 1 to 6 along with the sample data. Note that the polynomial regression of degree 6 goes through all seven points and fits perfectly.

The closest fit to the sample training data is based on the R2 measure in Excel. FIGURE 5.5 (left) shows this measure of fitness between the training data and the predictions of each polynomial model of certain degree. As shown in the figure, the polynomial of degree 6 has the perfect measure of fitness 1.0. The test error is based on the error measure

$$\sum_{i=1}^{8} |y_i - f(x_i)|$$

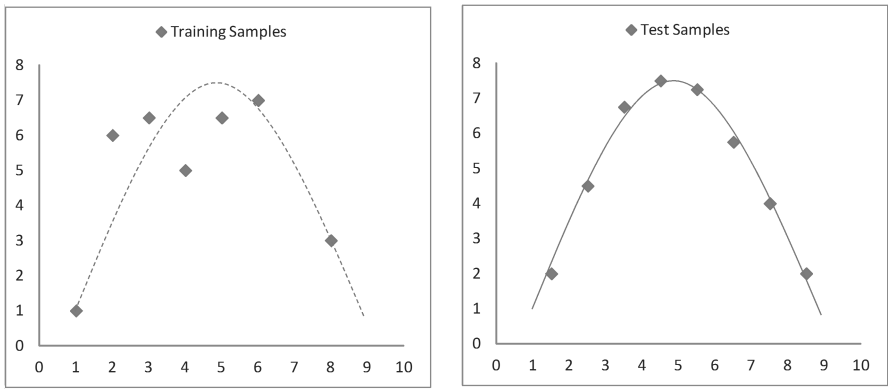


FIGURE 5.3 : Sample (left) and test data (right) from a function

between the test data and the six polynomial regressions corresponding to degrees 1 to 6. We can see in FIGURE 5.5 (left) that while the measure of fitness is steadily increasing towards 1.0, the test error in FIGURE 5.5 (right) reaches a minimum at degree 2 (hence the best fit) and then increases rapidly as the models begin to over-fit the training data.

5.3 BAYESIAN LINEAR REGRESSION

Bayesian linear regression views the regression problem as introduced above as an estimation of the functional dependence between an input variable $\mathbf{X}$ in $\Re^d$ and an output variable Y in $\Re$ as shown below:

$$\begin{aligned} & Y(\mathbf{X}) \\ & = \sum_{i=1}^{M} w_i \phi_i(\mathbf{X}) + \varepsilon \\ & = \mathbf{w}^T \Phi(\mathbf{X}) + \varepsilon \end{aligned}$$

where $\mathbf{w}^T \Phi(\mathbf{X})$ (e.g., $\Phi(\mathbf{X}) = (1, x, x^2, ...)$) is a linear combination of M predefined nonlinear basis functions $\phi_i(\mathbf{X})$ with input in $\Re^d$ and output in $\Re$. The observations are additively corrupted by i.i.d. noise with normal distribution

$$\varepsilon \sim N(0, \sigma_n^2)$$

that has zero mean and variance σ_n^2.

The goal of Bayesian regression is to estimate the weights w_i given a training set $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$ of data points. In contrast to classical regression, a Bayesian linear regression characterizes the uncertainty in $\mathbf{w}$ through a probability distribution $p(\mathbf{w})$. We use a multivariate normal distribution as prior on the weights as

$$p(\mathbf{w}) = N(\mathbf{0}, \Sigma_w)$$

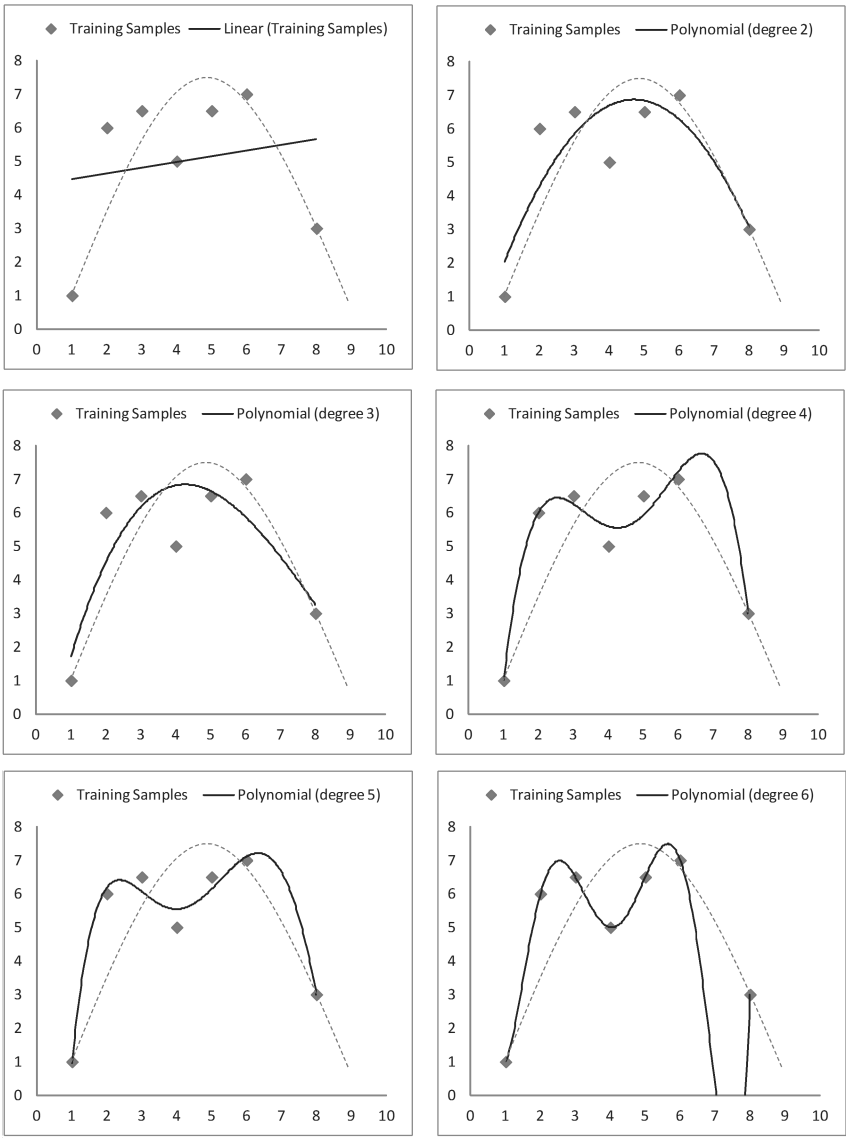


FIGURE 5.4 : Fitting polynomials of degrees 1 to 6

with zero mean and Σ_w as an $M \times M$ -sized covariance matrix. Further observations of data points modify this distribution using Bayes' theorem, with the assumption that the data points have been generated via the likelihood function. Let us illustrate Bayesian linear regression as

$$Y = \mathbf{w}^T \mathbf{X} + \varepsilon$$

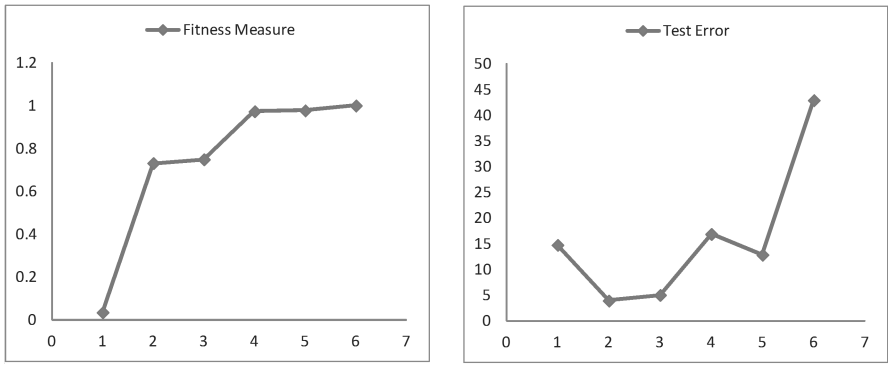


FIGURE 5.5 : Plots of fitness measures (left) and test errors (right) against polynomial degrees

Suppose $\mathbf{D}$ is the $d \times n$ matrix of input $\mathbf{x}$ vectors from this training set and $\mathbf{y}$ is the vector of y values. We have a Gaussian prior $p(\mathbf{w})$ of parameters $\mathbf{w}$, and the likelihood of the parameters is

$$p(\mathbf{y}|\mathbf{D}, \mathbf{w}) = N\left(\mathbf{w}^T\mathbf{D}, \sigma_n^2\mathbf{I}\right)$$

According to Bayes' rule, the posterior distribution over $\mathbf{w}$ is

$$\begin{aligned} & p(\mathbf{w}|\mathbf{y}, \mathbf{D}) \\ & \quad \propto p(\mathbf{y}|\mathbf{D}, \mathbf{w})\, p(\mathbf{w}) \\ & \quad = N\left(\tfrac{1}{\sigma_n^2}\mathbf{A}^{-1}\mathbf{D}\mathbf{y}, \mathbf{A}^{-1}\right), \text{ where } \mathbf{A} = \Sigma_{\mathbf{w}}^{-1} + \tfrac{1}{\sigma_n^2}\mathbf{D}\mathbf{D}^T \end{aligned}$$

The predictive distribution is

$$\begin{aligned} & p(Y|\mathbf{X}, \mathbf{y}, \mathbf{D}) \\ & \quad = \int_{\mathbf{w}} \left(\mathbf{w}^T\mathbf{X}\right) p(\mathbf{w}|\mathbf{y}, \mathbf{D})\, d\mathbf{w} \\ & \quad = N\left(\tfrac{1}{\sigma_n^2}\mathbf{X}^T\mathbf{A}^{-1}\mathbf{D}\mathbf{y}, \mathbf{X}^T\mathbf{A}^{-1}\mathbf{X}\right) \end{aligned}$$

The multivariate normal distribution above can be used for predicting Y given a vector input $\mathbf{X}$.

5.3.1 Gaussian Processes

A Gaussian process is a collection of random variables, any finite subset of which has a joint Gaussian distribution. Gaussian processes extend multivariate Gaussian distributions to infinite dimensionality. A regression technique starts with a set of data points, $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$, consisting of input-output pairs. The task is then to predict or interpolate the output value y given an input vector $\mathbf{x}$. As we observe the output, the new input-output pair

is added to the observation data set. Thus the data set grows in size over time. Any number of observations $y_1, ..., y_n$ in an arbitrary data set can be viewed as a single point sampled from some multivariate Gaussian distribution. Hence, a regression data set can be partnered with a Gaussian process where the prediction always takes into account the latest observations.

We consider a more general form of regression function for interpolation

$$Y = f(\mathbf{X}) + N\left(0, \sigma_n^2\right)$$

where each observation $\mathbf{X}$ can be thought of as related to an underlying function f through some Gaussian noise model. We solve the above for the function f. In fact, given n data points and new input $\mathbf{X}$, our objective is to predict Y and not the actual f since their expected values are identical (according to the above regression function). We can obtain a Gaussian process from the Bayesian linear regression model:

$f(\mathbf{X}) = \mathbf{w}^T\mathbf{X}$ with $\mathbf{w} \sim N(0, \Sigma_{\mathbf{w}})$

where the mean is given by

$$E[f(\mathbf{X})] = E\left[\mathbf{w}^T\right]\mathbf{X} = \mathbf{0}$$

and the covariance is given by

$$E\left[f(\mathbf{x}_i)^T f(\mathbf{x}_j)\right] = \mathbf{x}_i^T E\left[\mathbf{w}\mathbf{w}^T\right]\mathbf{x}_j = \mathbf{x}_i^T \Sigma_{\mathbf{w}} \mathbf{x}_j$$

It is often assumed that the mean of this Gaussian process is zero everywhere, but one observation is related to another observation via the covariance function

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \times e^{\frac{-|\mathbf{x}_i - \mathbf{x}_j|^2}{2l^2}} + \sigma_n^2 \times \delta(\mathbf{x}_i, \mathbf{x}_j)$$

where the maximum allowable covariance is defined as σ_f^2, which should be high (and hence not zero) for functions covering a broad range on the Y axis. The value of the kernel function approaches its maximum if $\mathbf{x}_i \approx \mathbf{x}_j$. In this case $f(\mathbf{x}_i)$ and $f(\mathbf{x}_j)$ are perfectly correlated. This means the neighboring points yield very close functional values, making the function smooth, and distant observations will have a negligible effect during interpolation of f at a new $\mathbf{x}$ value. The length parameter$l$ determines the effect of this separation. $\delta(\mathbf{x}_i, \mathbf{x}_j)$ is known as the *Kronecker delta* function ($\delta(\mathbf{x}_i, \mathbf{x}_j) = 0$ if $\mathbf{x}_i \neq \mathbf{x}_j$ else 1).

For Gaussian process regression, suppose the observation set is $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$ and a new input observation is x_* We capture the covariance functions $k(\mathbf{x}_i, \mathbf{x}_j)$ for all possible $\mathbf{x}_i, \mathbf{x}_j$, and x_* in the following three matrices:

$$\mathbf{K} = \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & k(\mathbf{x}_1, \mathbf{x}_2) & ... & k(\mathbf{x}_1, \mathbf{x}_n) \\ k(\mathbf{x}_2, \mathbf{x}_1) & k(\mathbf{x}_2, \mathbf{x}_2) & ... & k(\mathbf{x}_2, \mathbf{x}_n) \\ ... & ... & ... & ... \\ k(\mathbf{x}_n, \mathbf{x}_1) & k(\mathbf{x}_n, \mathbf{x}_2) & ... & k(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix}$$

$$\mathbf{K}_* = \begin{bmatrix} k(\mathbf{x}_*, \mathbf{x}_1) & k(\mathbf{x}_*, \mathbf{x}_2) & \ldots & k(\mathbf{x}_*, \mathbf{x}_n) \end{bmatrix}$$
$$\mathbf{K}_{**} = k(\mathbf{x}_*, \mathbf{x}_*)$$

Note that $k(\mathbf{x}_i, \mathbf{x}_i) = \sigma_f^2 + \sigma_n^2$, for all i. As per the assumption of a Gaussian process, the data set can be represented as a multivariate Gaussian distribution as follows:

$$\begin{bmatrix} \mathbf{y} \\ y_* \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} \mathbf{K} & \mathbf{K}_*^T \\ \mathbf{K}_* & \mathbf{K}_{**} \end{bmatrix}\right)$$

We are interested in the conditional distribution $p(y_*|\mathbf{y})$ which is given below:

$$y_*|\mathbf{y} \sim N\left(\mathbf{K}_*\mathbf{K}^{-1}\mathbf{y}, \mathbf{K}_{**} - \mathbf{K}_*\mathbf{K}^{-1}\mathbf{K}_*^T\right)$$

Therefore the best estimate for y_* is the mean $\mathbf{K}_*\mathbf{K}^{-1}\mathbf{y}$ of the above distribution.

Example
To illustrate the Gaussian process, consider the sample data set of seven points in TABLE 5.3, a plot of which was shown earlier in FIGURE 5.3.

TABLE 5.3: : Sample data set

X	Y
1	1
2	4.5
3	6.5
4	6
5	6.7
6	7
8	3
7.5	?

Considering the covariance function:

$$k(\mathbf{x}_i, \mathbf{x}_j) = e^{\frac{-|\mathbf{x}_i - \mathbf{x}_j|^2}{2l^2}}$$

With the value of l as 0.8 in the function k above, we calculate the covariance matrix $\mathbf{K}$ as shown in TABLE 5.4.

TABLE 5.4: : Matrix of covariance functions

1	0.457833	0.043937	0.000884	3.73E-06	3.29E-09	2.37E-17
0.457833	1	0.457833	0.043937	0.000884	3.73E-06	6.1E-13
0.043937	0.457833	1	0.457833	0.043937	0.000884	3.29E-09
0.000884	0.043937	0.457833	1	0.457833	0.043937	3.73E-06
3.73E-06	0.000884	0.043937	0.457833	1	0.457833	0.000884

TABLE 5.4: : Matrix of covariance functions

3.29E-09	3.73E-06	0.000884	0.043937	0.457833	1	0.043937
2.37E-17	6.1E-13	3.29E-09	3.73E-06	0.000884	0.043937	1

It is clear from the above table that the closer the X values are to each other, the higher the values of the covariance function are. We also have $\mathbf{K}_*$as shown in TABLE 5.5.

TABLE 5.5: : Vector of covariance functions

4.62E-15	5.45E-11	1.35E-07	6.98E-05	0.007576	0.172422	0.822578

The formula $\mathbf{K}_*\mathbf{K}^{-1}\mathbf{y}$ provides 3.23 as the mean of the predicted y-value for $x = 7.5$.

5.4 PRINCIPAL COMPONENT AND FACTOR ANALYSES

Principal component analysis (PCA) converts a set of measurements of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. PCA can be done by eigenvalue decomposition or SVD as introduced earlier in the background chapter.

Example

Consider the data set in FIGURE 5.6 with 6 variables and 51 observations of the US Electoral College votes, population, and area by state. The full data set is given in Appendix B.

Microsoft Excel - ElectoralPopArea.xls [Compatibility Mode]

	A	B	C	D	E	F	G	H
1	State	Electoral Votes	Population	Total Area	Water Area	Land Area	Density per Square Mile of	
2	Alabama	9	4,447,100	52,419.02	1,675.01	50,744.00	87.6	
3	Alaska	3	626,932	663,267.26	91,316.00	571,951.26	1.1	
4	Arizona	10	5,130,632	113,998.30	363.73	113,634.57	45.2	
5	Arkansas	6	2,673,400	53,178.62	1,110.45	52,068.17	51.3	
6	California	55	33,871,648	163,695.57	7,736.23	155,959.34	217.2	
7	Colorado	9	4,301,261	104,093.57	376.04	103,717.53	41.5	
8	Connecticut	7	3,405,565	5,543.33	698.53	4,844.80	702.9	
9	Delaware	3	783,600	2,489.27	535.71	1,953.56	401.1	
10	District of Columbia	3	572,059	68.34	6.94	61.4	9316.9	
11	Florida	27	15,982,378	65,754.59	11,827.77	53,926.82	296.4	

FIGURE 5.6 : United States Electoral College votes, population, and area by state

The correlation matrix in FIGURE 5.7 clearly indicates two groups of correlated variables. The fact that the number of electoral votes is proportional to the population gives rise to the first set of correlated variables. The second set of correlated variables is the set of all areas.

Correlation Matrix						
	Electoral Votes	Population	Total Area	Water Area	Land Area	Density per Square Mile of Land
Electoral Votes	1.0000	0.9996	0.1090	0.0510	0.1148	-.0757
Population	0.9996	1.0000	0.1100	0.0528	0.1156	-.0779
Total Area	0.1090	0.1100	1.0000	0.8276	0.9960	-.1679
Water Area	0.0510	0.0528	0.8276	1.0000	0.7739	-.0692
Land Area	0.1148	0.1156	0.9960	0.7739	1.0000	-.1783
Density per Square Mile of Land	-.0757	-.0779	-.1679	-.0692	-.1783	1.0000

FIGURE 5.7 : Correlation matrix for the data in FIGURE 5.6

All six eigenvalues and eigenvectors are shown in FIGURE 5.8, of which the first three are dominating as expected, given the two groups of correlated variables as shown in FIGURE 5.7 and the only remaining variable for density. The first principal component PRIN1 shows the domination of the coefficients corresponding to the three variables related to the areas. The second principal component PRIN2 shows the domination of the coefficients corresponding to Electoral Votes and Population.

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	2.83831733	0.90006486	0.4731	0.4731
2	1.93825247	0.97029922	0.3230	0.7961
3	0.96795324	0.71288651	0.1613	0.9574
4	0.25506673	0.25465650	0.0425	0.9999
5	0.00041023	0.00041023	0.0001	1.0000
6	0.00000000		0.0000	1.0000

Eigenvectors						
	PRIN1	PRIN2	PRIN3	PRIN4	PRIN5	PRIN6
Electoral Votes	0.195980	0.676606	0.058914	-.019591	-.707066	-.000000
Population	0.196790	0.676325	0.057041	-.023601	0.707141	0.000000
Total Area	0.570839	-.159819	0.044632	0.304680	0.000564	-.744163
Water Area	0.510240	-.187209	0.150926	-.818979	-.002452	0.105344
Land Area	0.562499	-.150401	0.026248	0.474510	0.001028	0.659639
Density per Square Mile of Land	-.142331	-.039750	0.983776	0.101702	0.001665	0.000000

FIGURE 5.8 : Eigenvalues and eigenvectors of the correlation matrix in FIGURE 5.7

FIGURE 5.9 shows the principal component plot of the data set in FIGURE 5.6 with the first two components. It is clear from the plot that the component PRIN1 is about the area and PRIN2 is about the population. This is the reason why the state of Alaska has a large PRIN1 value but very little PRIN2, and DC is just opposite. The state of California has large values for both PRIN1 and PRIN2.

Factor analysis (Anderson, 2003; Gorsuch, 1983) helps to obtain a small set of independent variables, called *factors* or *latent variables*, from a large set of correlated observed variables. Factor analysis describes the variability among observed variables in order to gain better insight into categories or to provide a simpler prediction structure. For example, factor analysis can reduce a large number of financial ratios into categories of financial ratios on

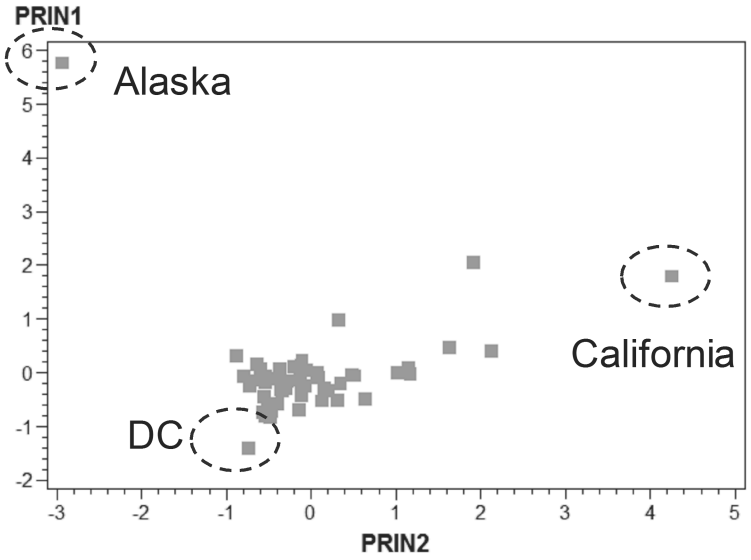


FIGURE 5.9 : Principal component plot of the data set in FIGURE 5.6

the basis of empirical evidence. It can help with finding contributing factors affecting, for example, prices of groups of stocks, GDPs of countries, and water and air qualities. For *exploratory factor analysis* (EFA), there is no predefined idea of the structure or dimensions in a set of variables. On the other hand, a *confirmatory factor analysis* (CFA) tests specific hypotheses about the structure or the number of dimensions underlying a set of variables.

The factor model proposes that observed responses $X_1, ..., X_n$ are partially influenced by underlying common factors $F_1, ..., F_m$ and partially by underlying unique factors $e_1, ..., e_n$.

$$\begin{aligned} X_1 &= \lambda_{11}F_1 + \lambda_{12}F_2 + ... + \lambda_{1m}F_m + e_1 \\ X_2 &= \lambda_{21}F_1 + \lambda_{22}F_2 + ... + \lambda_{2m}F_m + e_2 \\ &... \\ X_n &= \lambda_{n1}F_1 + \lambda_{n2}F_2 + ... + \lambda_{nm}F_m + e_n \end{aligned}$$

The coefficients λ_{ij} are called the *factor loadings*, so that λ_{ij} is the loading of the ith variable on the jth factor. Factor loadings are the weights and correlations between each variable and the factor. The higher the loading value, the more relevant the variable is in defining the factor's dimensionality. A negative value indicates an inverse impact on the factor. Thus a given factor influences some measures more than others, and this degree of influence is determined by loadings. The error terms $e_1, ..., e_n$ serve to indicate that the hypothesized relationships are not exact. The ith error term describes the residual variation specific to the ith variable X_i. The factors are often called

the *common factors* while the residual variables are often called the *unique* or *specific factors*.

The number of factors m should be substantially smaller than n. If the original variables $X_1, ..., X_n$ are at least moderately correlated, the basic dimensionality of the system is less than n. The goal of factor analysis is to reduce the redundancy among the variables by using a smaller number of factors.

To start an EFA, we first extract initial factors, using principal components to decide on the number of factors. Eigenvalue is the amount of variance in the data described by the factor, and eigenvalues help to choose the number of factors. In principal components, the first factor describes most of the variability. We then choose the number of factors to retain, and rotate axes to spread variability more evenly among factors. Redefining factors that loadings tend to make very high (-1 or 1) or very low (0) makes sharper distinctions in the interpretations of the factors.

Example

We apply the principal component-based factoring method on the data in FIGURE 5.6. The eigenvalues and the factor patterns are shown in FIGURE 5.10.

Eigenvalues of the Correlation Matrix: Total = 6 Average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	2.83831733	0.90006486	0.4731	0.4731
2	1.93825247	0.97029922	0.3230	0.7961
3	0.96795324	0.71288651	0.1613	0.9574
4	0.25506673	0.25465650	0.0425	0.9999
5	0.00041023	0.00041023	0.0001	1.0000
6	0.00000000		0.0000	1.0000

Factor Pattern

	Factor1	Factor2	Factor3	Factor4	Factor5
Electoral Votes	0.33017	0.94198	0.05796	-0.00989	-0.01432
Population	0.33154	0.94159	0.05612	-0.01192	0.01432
Total Area	0.96171	-0.22250	0.04391	0.15388	0.00001
Water Area	0.85962	-0.26063	0.14849	-0.41362	-0.00005
Land Area	0.94766	-0.20939	0.02582	0.23965	0.00002
Density per Square Mile of Land	-0.23979	-0.05534	0.96788	0.05136	0.00003

FIGURE 5.10 : Eigenvalues and factor patterns for the data in FIGURE 5.6

Now we apply the orthogonal Varimax rotation (maximizes the sum of the variances of the squared loadings) to obtain rotated factor patterns, as shown in FIGURE 5.11, and the revised distribution of variance explained by each factor, as shown in FIGURE 5.12. The total variance explained remains the same and gets evenly distributed between the major two factors.

Example

Here is an artificial example to check the validity and robustness of factor analysis. The data from the Thurstone box problem (Thurstone, 1947), as shown in FIGURE 5.13, measures 20 different characteristics of boxes, such as individual surface areas and box inter-diagonals. If these measurements are only linear combinations of the height, width, and depth of the box, then the

	Factor1	Factor2	Factor3	Factor4	Factor5
Electoral Votes	0.05001	0.99821	-0.02916	0.00444	-0.01432
Population	0.05057	0.99811	-0.03131	0.00657	0.01433
Total Area	0.98552	0.05696	-0.07507	0.14095	0.00000
Water Area	0.74351	0.01089	0.00056	0.66864	0.00001
Land Area	0.99306	0.06252	-0.08477	0.05223	0.00000
Density per Square Mile of Land	-0.09193	-0.04217	0.99487	-0.00145	-0.00001

FIGURE 5.11 : Rotated factor patterns from factors in FIGURE 5.10

Factor1	Factor2	Factor3	Factor4	Factor5
2.5237450	2.0016895	1.0044198	0.4697354	0.0004102

FIGURE 5.12 : Variance explained by each factor in FIGURE 5.11

data set could be reproduced by knowing only these dimensions and by giving them appropriate weights. These three dimensions are considered as factors.

i	x	y	z	i	x2	y2	z2	xy	xz	yz	sqrt(x2+y2	sqrt(x2+z2	sqrt(y2+z2	2(x+y)	2(x+z)	2(y+z)	logx	logy	logz	xyz	sqrt(x2+y2+z2)	ex	ey	ez
1	3	2	1	1	9	4	1	6	3	2	3.605551	3.162278	2.236068	10	8	6	0.477121	0.30103	0	6	3.741657387	20.08554	7.389056	2.718282
2	3	2	2	2	9	4	4	6	6	4	3.605551	3.605551	2.828427	10	10	8	0.477121	0.30103	0.30103	12	4.123105626	20.08554	7.389056	7.389056
3	3	3	1	3	9	9	1	9	3	3	4.242641	3.162278	3.162278	12	8	8	0.477121	0.477121	0	9	4.358898944	20.08554	20.08554	2.718282
4	3	3	2	4	9	9	4	9	6	6	4.242641	3.605551	3.605551	12	10	10	0.477121	0.477121	0.30103	18	4.69041576	20.08554	20.08554	7.389056
5	3	3	3	5	9	9	9	9	9	9	4.242641	4.242641	4.242641	12	12	12	0.477121	0.477121	0.47712	27	5.196152423	20.08554	20.08554	20.08554
6	4	2	1	6	16	4	1	8	4	2	4.472136	4.123106	2.236068	12	10	6	0.60206	0.30103	0	8	4.582575695	54.59815	7.389056	2.718282
7	4	2	2	7	16	4	4	8	8	4	4.472136	4.472136	2.828427	12	12	8	0.60206	0.30103	0.30103	16	4.898979486	54.59815	7.389056	7.389056
8	4	3	1	8	16	9	1	12	4	3	5	4.123106	3.162278	14	10	8	0.60206	0.477121	0	12	5.099019514	54.59815	20.08554	2.718282
9	4	3	2	9	16	9	4	12	8	6	5	4.472136	3.605551	14	12	10	0.60206	0.477121	0.30103	24	5.385164807	54.59815	20.08554	7.389056
10	4	3	3	10	16	9	9	12	12	9	5	5	4.242641	14	14	12	0.60206	0.477121	0.47712	36	5.830951895	54.59815	20.08554	20.08554
11	4	4	1	11	16	16	1	16	4	4	5.656854	4.123106	4.123106	16	10	10	0.60206	0.60206	0	16	5.744562647	54.59815	54.59815	2.718282
12	4	4	2	12	16	16	4	16	8	8	5.656854	4.472136	4.472136	16	12	12	0.60206	0.60206	0.30103	32	6	54.59815	54.59815	7.389056
13	4	4	3	13	16	16	9	16	12	12	5.656854	5	5	16	14	14	0.60206	0.60206	0.47712	48	6.403124237	54.59815	54.59815	20.08554
14	5	2	1	14	25	4	1	10	5	2	5.385165	5.09902	2.236068	14	12	6	0.69897	0.30103	0	10	5.477225575	148.4132	7.389056	2.718282
15	5	2	2	15	25	4	4	10	10	4	5.385165	5.385165	2.828427	14	14	8	0.69897	0.30103	0.30103	20	5.744562647	148.4132	7.389056	7.389056
16	5	3	2	16	25	9	4	15	10	6	5.830952	5.385165	3.605551	16	14	10	0.69897	0.477121	0.30103	30	6.164414003	148.4132	20.08554	7.389056
17	5	3	3	17	25	9	9	15	15	9	5.830952	5.830952	4.242641	16	16	12	0.69897	0.477121	0.47712	45	6.557438524	148.4132	20.08554	20.08554
18	5	4	1	18	25	16	1	20	5	4	6.403124	5.09902	4.123106	18	12	10	0.69897	0.60206	0	20	6.480740698	148.4132	54.59815	2.718282
19	5	4	2	19	25	16	4	20	10	8	6.403124	5.385165	4.472136	18	14	12	0.69897	0.60206	0.30103	40	6.708203932	148.4132	54.59815	7.389056
20	5	4	3	20	25	16	9	20	15	12	6.403124	5.830952	5	18	16	14	0.69897	0.60206	0.47712	60	7.071067812	148.4132	54.59815	20.08554

FIGURE 5.13 : 20 variable box problem data set (Thurstone, 1947)

As shown in FIGURE 5.14, the three dimensions of space are approximately discovered by the factor analysis, despite the fact that the box characteristics are not linear combinations of underlying factors but are instead multiplicative functions. Initial loadings and components are extracted using PCA.

The question is how many factors to extract in a given data set. Kaiser's criterion suggests that it is only worthwhile to extract factors which account for large variance. Therefore, we retain those factors with eigenvalues equal to or greater than 1.

There are 20 observations, each a function of x, y or z or one of their combinations. In FIGURE 5.14, Proportion indicates the relative weight of each factor in the total variance. For example, $12.6149/20 = 0.6307$. So the first factor explains about 63% of the total variance. Cumulative shows the total amount of variance explained, and the first six eigenvalues explain almost 99.6% of the total variance. From a factor analysis perspective the first three eigenvalues suggest a factor model with three common factors. This is because the first two eigenvalues are greater than unity and the third one is closer to unity and together they explain over 98% of the total variance.

	Eigenvalue	Difference	Proportion	Cumulative
1	12.6149281	8.5779579	0.6307	0.6307
2	4.0369702	1.0555741	0.2018	0.8326
3	2.9813961	2.8658323	0.1491	0.9817
4	0.1155639	0.0243452	0.0058	0.9874
5	0.0912187	0.0161426	0.0046	0.9920
6	0.0750761	0.0212835	0.0038	0.9958
7	0.0537926	0.0336045	0.0027	0.9984
8	0.0201881	0.0096911	0.0010	0.9995
9	0.0104970	0.0102092	0.0005	1.0000
10	0.0002878	0.0002478	0.0000	1.0000
11	0.0000400	0.0000124	0.0000	1.0000
12	0.0000276	0.0000164	0.0000	1.0000

		Factor1	Factor2	Factor3
x2	x2	0.65429	-0.64243	0.39829
y2	y2	0.73494	-0.07609	-0.67280
z2	z2	0.66029	0.69102	0.27622
xy	xy	0.87500	-0.34749	-0.32152
xz	xz	0.83048	0.34552	0.41593
yz	yz	0.83741	0.53072	-0.04473
sqrt(x2+y2)	sqrt(x2+y2)	0.86147	-0.50000	-0.06145
sqrt(x2+z2)	sqrt(x2+z2)	0.84659	-0.26776	0.45159
sqrt(y2+z2)	sqrt(y2+z2)	0.87070	0.28522	-0.38836
2(x+y)	2(x+y)	0.88297	-0.42891	-0.18507
2(x+z)	2(x+z)	0.88467	0.03195	0.46089
2(y+z)	2(y+z)	0.87804	0.40991	-0.23776
logx	logx	0.66268	-0.63904	0.36671

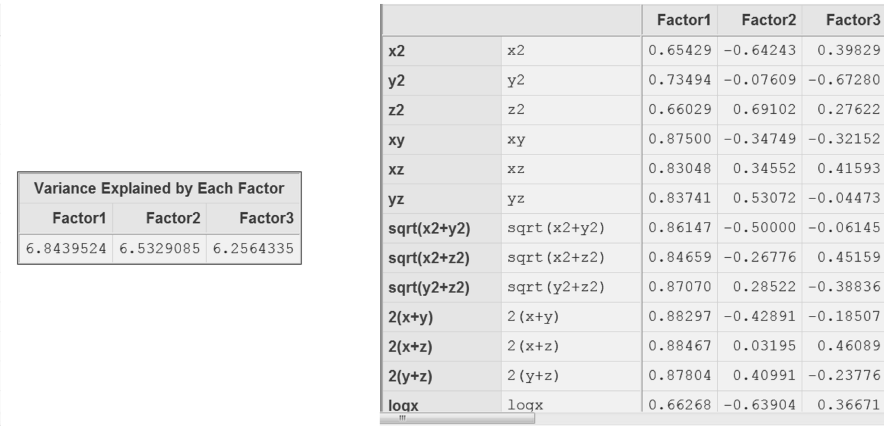
FIGURE 5.14 : Factor analysis of the data in FIGURE 5.13 shows the eigenvalues for variance explained by each factor and three retained factors

Rotating the components towards independence, rather than rotating the loadings towards simplicity, allows one to accurately recover the dimensions of each box and also to produce simple loadings. FIGURE 5.15 shows the factors of FIGURE 5.14 after an orthogonal Varimax rotation. The total of eigenvalues for the factors remains the same, but variability among factors is evenly distributed.

There are some differences between EFA and PCA and they will provide somewhat different results when applied to the same data. EFA assumes that the observations are based on the underlying factors, whereas in PCA the principal components are based on observations. The rotation of components is part of EFA but not PCA.

5.5 SURVIVAL ANALYSIS

Survival analysis (Kleinbaum and Klein, 2005; Hosmer et al., 2008) is a time-to-event analysis that measures the time from the beginning of a study to a terminal event, conclusion of the observation period, or loss of contact/with-



Variance Explained by Each Factor

Factor1	Factor2	Factor3
6.8439524	6.5329085	6.2564335

		Factor1	Factor2	Factor3
x2	x2	0.65429	-0.64243	0.39829
y2	y2	0.73494	-0.07609	-0.67280
z2	z2	0.66029	0.69102	0.27622
xy	xy	0.87500	-0.34749	-0.32152
xz	xz	0.83048	0.34552	0.41593
yz	yz	0.83741	0.53072	-0.04473
sqrt(x2+y2)	sqrt(x2+y2)	0.86147	-0.50000	-0.06145
sqrt(x2+z2)	sqrt(x2+z2)	0.84659	-0.26776	0.45159
sqrt(y2+z2)	sqrt(y2+z2)	0.87070	0.28522	-0.38836
2(x+y)	2(x+y)	0.88297	-0.42891	-0.18507
2(x+z)	2(x+z)	0.88467	0.03195	0.46089
2(y+z)	2(y+z)	0.87804	0.40991	-0.23776
logx	logx	0.66268	-0.63904	0.36671

FIGURE 5.15 : Factors of FIGURE 5.14 after an orthogonal Varimax rotation

drawal from the study. Survival data consist of a response variable that measures the duration of time (event time, failure time, or survival time) until a specified event occurs. Optionally, the data may contain a set of independent variables that are possibly associated with the failure time variable. Examples of survival analysis include determining the lifetime of a device (the time after installation until the device breaks down), the time until a company declares bankruptcy after its inception, the length of time an auto driver stayed accident-free since becoming insured, the length of time a person stayed on a job, the retention time of customers, and the survival time (or time until death) for organ transplant patients since transplant surgery.

The *censoring* problem in survival analysis arises due to incomplete observations of survival time during a period of study. Some subjects of study have censored survival times because they may not be observed for the full study period due to drop-out, loss to followup, or early termination of the study. A censored subject may or may not have an event of interest if it occurs before the end of the study but their data is incomplete.

Example

FIGURE 5.16 illustrates the survival histories of six subjects in an example of a study that could be measuring implanted device lifetimes or post-organ transplant outcomes.

In the figure, the terminal event of interest is "breakdown" or "death" which motivates a study of survival time. Not all devices will cease to work during the 325 days of the study period, but all will break down eventually. In the figure, the solid line represents an observed period at risk, while the broken line represents an unobserved period at risk. The letter "X" represents an

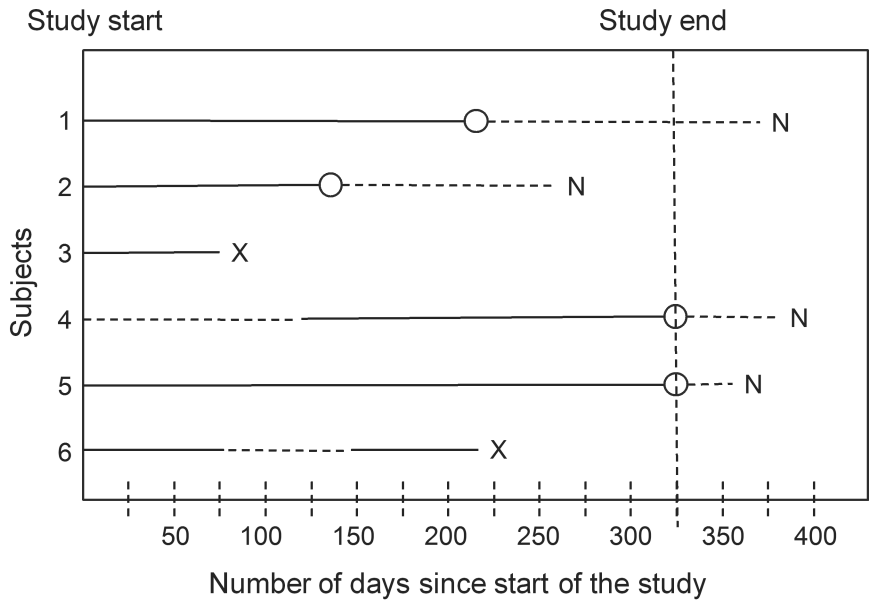


FIGURE 5.16 : Survival histories of subjects for analysis

observed terminal event, the open circle represents the censoring time, and the letter "N" represents an unobserved event.

An observation that is *right-censored* means the relevant event has not yet occurred at the time of observation. An observation that is *left-censored* means the relevant event has occurred before the time of observation but the exact time is not known. An observation is *interval-censored* if the event occurs at an unknown point in a time interval. Right-censored observations are the most common kind.

Example

In FIGURE 5.16, the observation of subject 5 is right-censored. Subject 4 joined the study late. Subject 6 is lost to observation for a while. Subject 2 joined the study on time but was lost to observation after some time, and died before the study period ended, but it is not known exactly when.

Consider the random variable T representing the survival time with density $f(t)$. The cumulative distribution function is $F(t) = p(T \leq t)$, which represents the probability that a subject survives no longer than time t. $S(t)$ is the survival function or the probability that a subject survives longer than

time t, that is,

$$S(t) = p(T > t) = 1 - F(t) = \int_t^{\infty} f(s)\, ds.$$

A typical question to be asked is "What is the probability of the device lasting past 300 days?" and the answer is $S(300)$. The hazard function is defined as

$$h(t) = \frac{f(t)}{S(t)}.$$

Methods for estimating survival function include *life table analysis*, *Kaplan-Meier product-limit estimator*, and *Cox's semi-parametric proportional hazard model*, of which the life table is the least complicated way to describe the survival in a sample. A straightforward multiple regression technique is not suitable for survival analysis because of the problem of censoring and the survival time dependent variable, as well as the fact that other independent variables are not normally distributed.

In the life table analysis, the distribution of survival times is divided into a certain number of intervals. For each interval we can then compute the number and proportion of cases or objects that entered the respective interval, the number and proportion of terminal events, and the number of cases that were lost or censored in the respective interval. Several additional statistics can be computed based on these numbers and proportions, especially the estimated probability p_i of failure in the respective interval. This probability p_i of the ith interval is computed per unit of time as $(n_i - n_{i+1})/t_i$, where n_i is the estimated cumulative proportion surviving at the beginning of the ith interval, n_{i+1} is the cumulative proportion surviving at the end of the ith interval, and t_i is the width of the ith interval. Since the probabilities of survival are assumed to be independent across the intervals, the survival probability up to an interval is computed by multiplying out the probabilities of survival across all previous intervals.

The life table gives us a good indication of the distribution of survival over time. However, for predictive purposes it is often desirable to understand the shape of the underlying survival function in the population. The two major distributions that have been proposed for modeling survival or failure times are the exponential and the Weibull distribution.

The Kaplan-Meier product-limit estimator (1958) is a life table analysis in which each time interval contains exactly one case. The method arranges the data in increasing order of the observed values, noting the censored cases. It then computes the proportion of subjects who left the study after each change in the ordering. The advantage of using this estimator is that the resulting estimates do not depend on the grouping of the data into a certain number of time intervals. Cox's proportional hazards model determines the underlying hazard rate as a function of the independent variables.

Example

We undertake a retrospective analysis of 863 records (part displayed in FIGURE 5.17) of patients who underwent a kidney transplant during a certain period of time. The patient population examined contains males and females, and both black and white subjects.

```
Data on 863 kidney transplant patients
See Section 1.7
Data can be read in free format. The variables represented in the dataset are as follows:

Observation number
Time to death or on-study time
Death indicator (0=alive, 1=dead)
Gender (1=male, 2=female)
Race (1=white, 2=black)
Age in years
THE DATA
   1        1       0      1       1      46
   2        5       0      1       1      51
   3        7       1      1       1      55
   4        9       0      1       1      57
   5       13       0      1       1      45
   6       13       0      1       1      43
   7       17       1      1       1      47
   8       20       0      1       1      65
   9       26       1      1       1      55
  10       26       1      1       1      44
  11       28       1      1       1      49
  12       32       0      1       1      52
  13       32       0      1       1      31
  14       43       0      1       1      63
```

FIGURE 5.17 : Example data for survival analysis (Ref: http://www.mcw.edu/FileLibrary/Groups/Biostatistics/Publicfiles/DataFromSection/DataFromSectionTXT/Data_from_section_1.7.txt)

Survival studies were calculated using the Kaplan-Meier Product-Limit method. The outcome endpoints were alive and dead. The data were grouped by the gender and race variables and hence there were four groups. The survival functions for both white and black females are shown in FIGURE 5.18.

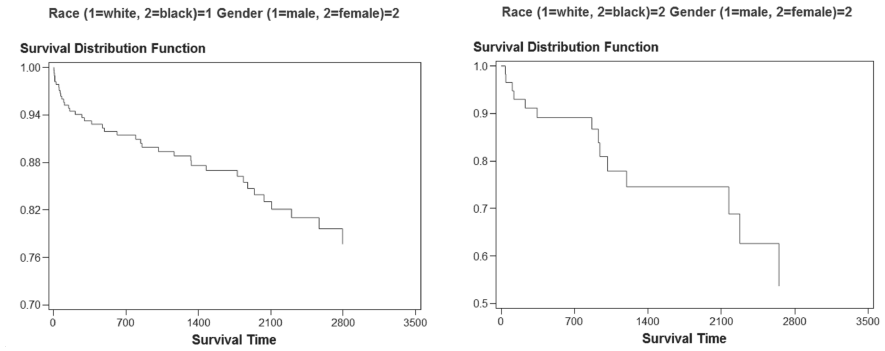


FIGURE 5.18 : An example survival analysis plots

It's clear from the figure that the average survival time is higher for the population representing the distribution function on the left than on the right. Similar other conclusions can be drawn by studying and comparing the two distribution functions.

5.6 AUTOREGRESSION MODELS

A process is stochastic if it evolves in time according to probabilistic laws. In this section, we introduce three types of stochastic processes, Autoregressive (AR), Moving Average (MA), and Autoregressive and Moving Average (ARMA), which are a special type of Gaussian process for modeling time-series data. A simple AR process, AR(1), is recursively defined as follows:

$$X_t = aX_{t-1} + \varepsilon_t$$

where a is the coefficient, ε_t is white noise with $\varepsilon_t \sim N\left(0, \sigma^2\right)$ and ε_i and ε_j are independent for $i \neq j$. By repeated substitution,

$$X_t = \varepsilon_t + a\varepsilon_{t-1} + a^2\varepsilon_{t-2} + ...$$

Therefore, $E\left[X_t\right] = 0$ and $Var\left[X_t\right] = \sigma^2\left(1 + a^2 + a^4 + ...\right) = \frac{\sigma^2}{1-a^2},\ if\ |a| < 1$. More generally, an autoregression model of order p, AR(p), is defined as :

$$X_t = a_1X_{t-1} + ... + a_pX_{t-p} + \varepsilon_t$$

where X_t can be obtained by linear regression from $X_{t-1}, ..., X_{t-p}$.

The MA process represents time series that are generated by passing the white noise through a non-recursive linear filter. The general MA process of order q, MA(q), is defined as follows:

$$X_t = \varepsilon_t + b_1\varepsilon_{t-1} + ... + b_q\varepsilon_{t-q}$$

where b_is are coefficients and ε_i is white noise with $\varepsilon_i \sim N\left(0, \sigma^2\right)$. AR(p) and MA(q) processes can be combined to define ARMA(p, q) as

$$X_t = \underbrace{a_1X_{t-1} + ... + a_pX_{t-p}}_{\text{Autoregressive (AR)}} + \underbrace{\varepsilon_t + b_1\varepsilon_{t-1} + ... + b_q\varepsilon_{t-q}}_{\text{MovingAvergae (MA)}}$$

Example

FIGURE 5.19 is an autoregressive plot of forecasts of the closing values of Intuit stock for seven days in the first half of January 2006. The order of the autoregressive process is 5 and the process employs a maximum likelihood estimator method. The figure also shows the 3 standard deviations from the mean as the Lower Control Limit (LCL) and Upper Control Limit (UCL) values.

Note from the figure that the autoregression process takes a time-step or two to predict a steep ascension or decline in the observed data.

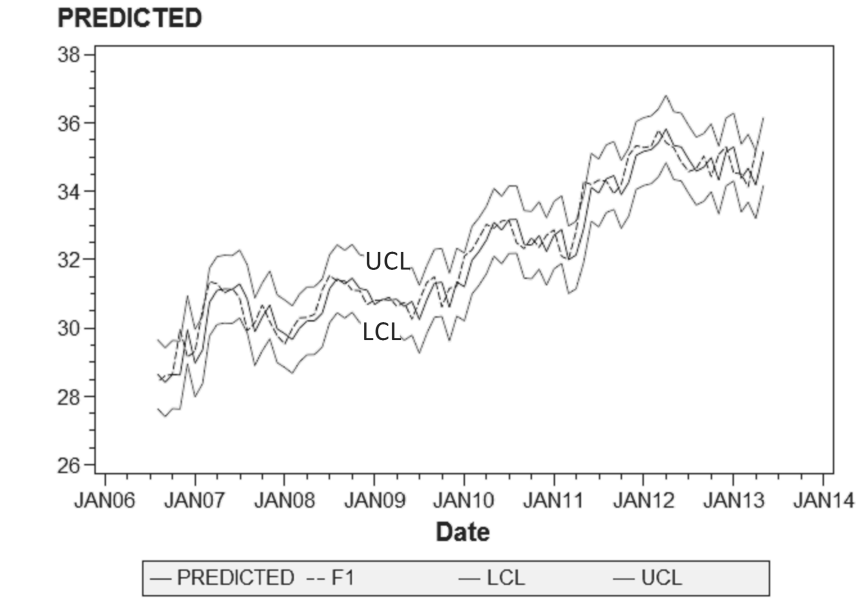


FIGURE 5.19 : Plot of forecasts by autoregression

5.7 FURTHER READING

Two very comprehensive books on applied regression analyses are (Gelman and Hill, 2006) and (Kleinbaum et al., 2007). A great tutorial on survival analysis can be found in (Kleinbaum and Klein, 2005).