Chapter 4
Statistical Models for Prediction

Background In this chapter, we consider statistical models for different types of outcomes: binary, unordered categorical, ordered categorical, continuous, and survival data. We discuss common statistical models in medical research such as the linear, logistic, and Cox regression model, and also simpler approaches and more flexible extensions, including regression trees and neural networks. Details of the methods are found in many excellent texts. We focus on the most relevant aspects of these models in a prediction context. All models are illustrated with case studies. In Chap. 6, we will discuss aspects of choosing between alternative statistical models.

4.1 Continuous Outcomes

Continuous outcomes have traditionally received most attention in texts on regression modelling, with the ordinary least square model ("linear regression") as the reference statistical model.[64,137,232,281,472] Continuous outcome are quite common in medical, epidemiological, and economical studies, but not so often considered for clinical prediction models.

The linear regression model can be written as

$$y = \alpha + \beta_i \times \mathrm{x}_i + \text{error},$$

where α refers to the intercept, β_i to the set of regression coefficients that relate one or more predictors x_i to the outcome y. The error is calculated as observed y – predicted y ($\hat{y}$). This difference is also known as the residual for the prediction of y. We assume that the residuals have a normal distribution, and do not depend on x_i ("homoscedasticity").

The outcome y is hence related to a *linear combination* of the x_i variables with the estimated regression coefficients β_i. This is an important property, which is also seen in *generalized* linear models, such as the logistic regression model.

E.W. Steyerberg, *Clinical Prediction Models*,
DOI:10.1007/978-0-387-77244-8_4,

*4.1.1 Examples of Linear Regression

An example of a medical outcome is blood pressure. We may want to predict the blood pressure after treatment with an anti-hypertensive or other intervention.[241,460] Also, quality of life scales may be relevant to evaluate.[242] Such scales are strictly speaking only ordinal, but can for practical purposes often be treated as continuous outcomes. A specific issue is that quality of life scores have ceiling effects, because minimum and maximum scores apply.

4.1.2 Economic Outcomes

Health economics is another important field where continuous outcomes are considered, such as length of stay in hospital, or length of stay at a specific ward (e.g. the intensive care unit), or total costs for patients.[88]

Cost data are usually not normally distributed. Such economic data have special characteristics, such as patients without any costs (zero), and a long tail because some patients having considerable costs. We might consider the median as a good descriptor of the outcome. Interestingly, we are however always interested in the mean costs, since the expectation is what matters most from an economical perspective. Sometimes analyses have been performed to identify "high-cost" patients, after dichotomizing the outcome at some cost threshold.

*4.1.3 Example: Prediction of Costs

Many children in moderate climates suffer from an infection by the respiratory syncytial virus (RSV). Some children, especially premature children are at risk of a severe infection, leading to hospitalization. The mean RSV hospitalization costs were 3,110 euros in a cohort of 3,458 infants and young children hospitalized for severe RSV disease during the RSV seasons 1996–1997 to 1999–2000 in the Southwest of The Netherlands. RSV hospitalization costs were higher for some patient categories, e.g. those with lower gestational age or lower birth weight, and younger age. The linear regression model had an adjusted R^2 of 8%.[345] This indicates a low explanatory ability for predicting hospitalization costs of individual children. However, the model could accurately estimate the anticipated mean hospitalization costs of groups of children with the same characteristics. These predicted costs were used in decision analyses of preventive strategies for severe RSV disease.[46]

4.1.4 Transforming the Outcome

An important issue in linear regression is whether we should transform the outcome variable. The residuals $(y-\hat{y})$ from a linear regression should have a normal distribution with a constant spread ("homoscedasticity"). This can sometimes be achieved by,

e.g. a log transformation for cost data, but other transformations are also possible. As Harrell points out, transformations of the outcome may reduce the need to include transformations of predictor variables.[174] Care should be taken in backtransforming predicted mean outcomes to the original scale. Predicted medians and other quantiles are not affected by transformation. The log–normal distribution can be used for the mean on the original scale after a log transformation, but a more general, non-parametric, approach is to use "smearing" estimators.[341]

4.1.5 Performance: Explained Variation

In linear regression analysis, the total variance in y ("total sum of squares", TSS) is the sum of variability explained by one or more predictors ("model sum of squares", MSS) and the error ("residual sum of squares", RSS):

$$\text{TSS} = \text{MSS} + \text{RSS}$$
$$\text{var (regression on } x_i\text{)} + \text{var (error)} = \sum (\hat{y} - \text{mean } (y))^2 + \sum (y - \hat{y})^2$$

The estimates of the variance follow from the statistical fit of the model to the data, which is based on the analytical solution of a least squares formula. This fit minimizes the error term in the model, and maximizes the variance explained by *xi*. Better prediction models explain more of the variance in y. R^2 is defined as MSS / TSS.[472]

To appreciate values of R^2, we consider six hypothetical situations where we predict a continuous outcome y, which has a standard normal distribution (N(0,1), i.e. mean 0 and standard deviation 1) with one predictor x (N(0,1)). The regression coefficients for x are varied in simulations, such that R^2 is 95%, 50%, 20%, 10%, 5%, and 0% (Fig. 4.1). We note that an R^2 of 95% implies that observed outcomes are always very close to the predicted values, while gradually relatively more error occurs with lower R^2 values. When R^2 is 0%, no association is present.

To appreciate R^2 further, we plot the distributions of predicted values ($\hat{y}$). The distribution of $\hat{y}$ is wide when R^2 is 95%, and very small when R^2 is 5%, and near a single line when R^2 is 0% (Fig. 4.2). The distribution of y is always normal with mean 0 and standard deviation 1.

4.1.6 More Flexible Approaches

The generalized additive model (GAM) is a more flexible variant of the linear regression model.[180, 181, 472] A GAM allows for more flexibility especially for continuous predictors. It replaces the usual linear combination of continuous predictors with a sum of smooth functions to capture potential non-linear effects: $y = b_0 + f_i(x_i)$ + error, where f_i refers to functions for each predictor, e.g. loess smoothers.

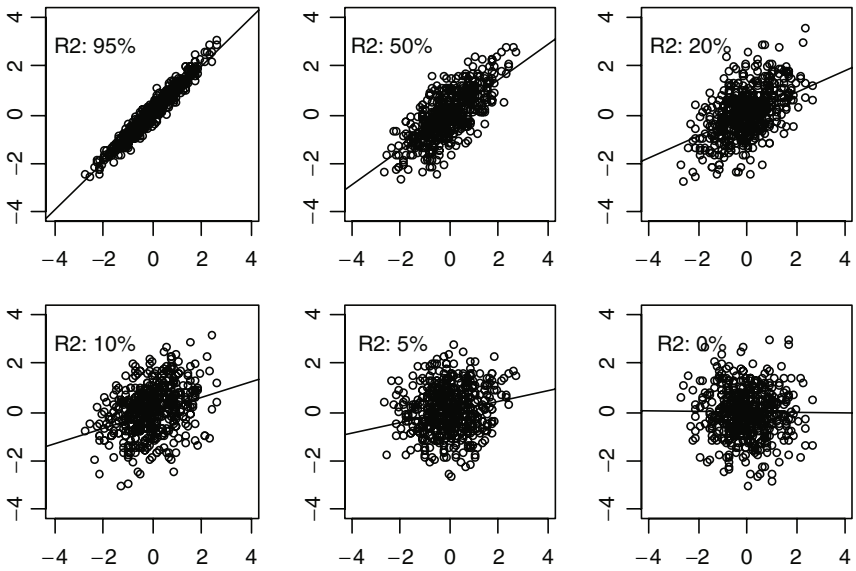

Fig. 4.1 Linear regression analysis with true regression models with $y = \beta \times x + \text{error}$, where $\text{sd}(y) = \text{sd}(x) = 1$. The outcome y is shown on the y-axis, x on the x-axis

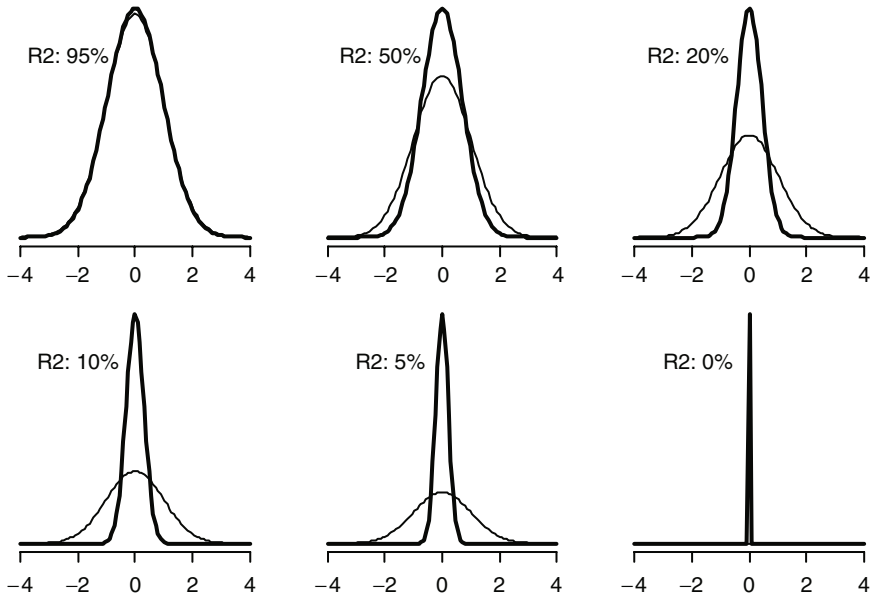

Fig. 4.2 Probability density functions for observed and predicted values ("fitted values", $\hat{y}$). For the first graph ($R^2 = 95\%$), the distribution of predicted values (*thick line*) is nearly identical to the distribution of observed y values (*thin line*), while for the last graph all predictions are for the average of 0

Loess smoothers are based on locally weighted polynomial regression.[75] At each point in the data set a low-degree polynomial is fit to a subset of the data, with data values near the point where the outcome *y* is considered. The polynomial is fitted using weighted least squares, giving more weight to nearby points and less weight to points further away. The degree of the polynomial model and the weights can be chosen by the analyst.

The estimation of a GAM is more computationally demanding than for linear models, but this is no limitation anymore with modern computer power and software. A GAM assumes that the outcome is already appropriately transformed, and then automatically estimates the transformation of continuous predictors to optimize prediction of the outcome.

An even more flexible approach is the alternating conditional expectation method.[174, 181] Here, *Y* and *X*s are simultaneously transformed to maximize the correlation between the transformed *Y* and the transformed *X*s.

$g(y) = \alpha + f_i(x_i)$+ error, where *g* refers to a transformation of the outcome *y*, and f_i refers to functions for each predictor. For cost data, several other specific approaches have been proposed.[27, 341]

4.2 Binary Outcomes

For outcome prediction, we often consider diagnostic (presence of disease) or prognostic outcomes (e.g. mortality, morbidity, complications, see Chap. 2). The logistic regression model is the most widely used statistical technique nowadays for such binary medical outcomes.[174, 472] The model is flexible in that it can incorporate categorical and continuous predictors, non-linear transformations, and interaction terms. Many of the principles of linear regression also apply for logistic regression, which is an example of a *generalized* linear model. As in linear regression, the binary outcome *Y* is linked to a linear combination of a set of predictors and regression coefficients β. We use the logistic link function to restrict predictions to the interval <0,1>. The model is stated in terms of the probability that *y* = 1 ("*P*(*y*=1)"), rather than the outcome *Y* directly.

Specifically, we write the model as a linear function in the logistic transformation (logit), where logit(*P*(*y*=1)) = log(odds(*P*(*y*=1)), or log([*P*(*y*=1)/(*P*(*y*=1)+1)]):

$\text{Logit}(P(y=1)) = \propto_0 + \beta_i \times x_i = \text{lp}$, where logit indicates the logistic transformation, α the intercept, βi the estimated regression coefficients, x_i the predictors, and lp linear predictor.

The coefficients β_i are usually estimated by maximum likelihood in a standard logistic regression approach, but this is not necessarily the case. For example, we will discuss penalized maximum likelihood methods to shrink the β_i for predictive purposes (Chap. 13). The interpretation of the coefficients β_i is as for any regression model, that the coefficient indicates the effect of a 1-unit increase in x_i, keeping the other predictors in the model constant. When we consider a single predictor in a logistic model, β_i is an unadjusted, or univariate effect; with multiple predictors, it is an "adjusted" effect, conditional on the values of other predictors in the model. The exponent of the regression coefficient (e^{β}) indicates the odds ratio.

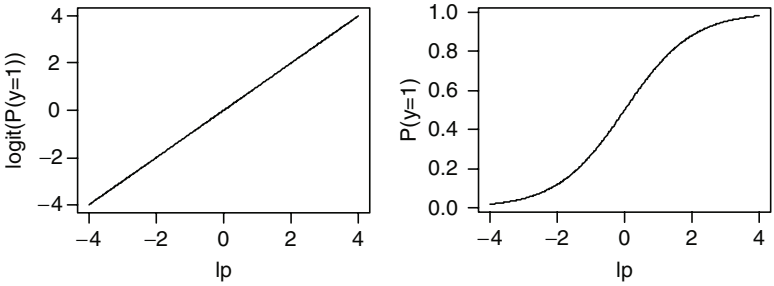

Fig. 4.3 Logistic function. The linear predictor lp is related to the predicted probability $P(y{=}1)$ as: Logit($P(y{=}1)$) = lp, or $P(y{=}1) = 1 / (1 + \exp(-\text{lp}))$

Predicted probabilities can be calculated by backtransforming: $p(y{=}1) = e^{lp} / (1+e^{lp}) = 1 / (1+e^{-lp})$. The quantity e^{lp} is the odds of the outcome. The logistic function has a characteristic sigmoid shape, as is bounded between 0 and 1 (Fig. 4.3). We note that a lp value of 0 corresponds to a probability of 50%. Low lp values correspond to low probabilities (e.g. lp −4, $p<2\%$), and high lp values correspond to high probabilities (e.g. lp +4, $p>98\%$).

4.2.1 R^2 in Logistic Regression Analysis

We learned from the linear regression examples that R^2 is related to the relative spread in predictions. When predictions cover a wider range, the regression model better predicts the outcome. This concept also applies to dichotomous outcomes, e.g. analyzed with a logistic regression model. Better prediction models for dichotomous outcomes have a wider spread in predictions, i.e. predictions close to 0% and close to 100%.

To illustrate this concept, we use the same simulated data as for the examples of linear regression models, but we now dichotomize the outcome y (if $y<0$, yd = 0, else yd = 1). The relationship between a standard normal variable x and the six yd outcomes is shown in Fig. 4.4.

4.2.2 Calculation of R^2 on the Log Likelihood Scale

Where the linear model is optimized with least squares estimation, the logistic model is usually optimized with maximum likelihood techniques. The likelihood refers to the probability of the data given the model, and enables estimation of parameters in various non-linear models. The natural logarithm of the likelihood (log likelihood, LL) is usually used for convenience in numerical estimation. The LL is calculated as the sum over all subjects of the distance between the natural log of the predicted probability p for the binary outcome to the actually observed outcome y:

$$\text{LL} = \sum y \times \log(p) + (1 - y) \times \log(1 - p),$$

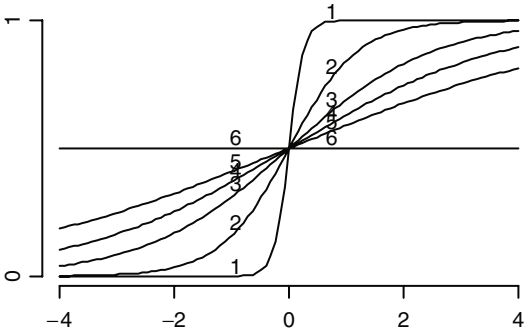

Fig. 4.4 Predicted probabilities of a 0/1 outcome by six logistic models according to a normally distributed *x* variable. The predictive strength varied, with Nagelkerke's R^2 decreasing from 87% (labelled "1") to 0% (label "6")

where y refers to the binary outcome and p the predicted probability for each subject.

If $y = 1$, the probability should be high (ideally 100%), such that $\log(p)$ gets close to 0. Then the term $(1-y)$ drops out. If $y=0$, the term $(1-y) = 1$, and p should be low (ideally 0%), such that $\log(1-p)$ gets close to zero. A perfectly fitting model would have an LL of zero. In medical problems, perfect predictions cannot be made, unless a fully deterministic model is identified. The LL is hence usually negative for a fitted logistic regression model. A better model will have an LL closer to zero.

As reference we consider the LL of a model with average predictions:

$$LL_0 = \sum y \times \log(\text{mean}(y/n)) + (1 - y) \times \log(1 - \text{mean}(y/n)),$$

where LL_0 refers to the log likelihood of the Null model, and mean(y/n) is the average probability of the binary outcome y. The LL_0 is negative, unless y/n is 0 or 1.

We can quantify the performance of a prognostic model by comparison with the Null model. We multiply by −2, since the difference on the −2 LL scale is a Likelihood Ratio statistic (LR), which follows a χ^2 distribution:

$$LR = -2\,(LL_0 - LL_1),$$

where LL_1 refers to the model with predictors, LL_0 to the Null model, and LR is the likelihood ratio. The LR statistic can be used for univariate analysis, but also for testing the joint importance of a larger set of predictors in the model ("global LR statistic"). We can also easily make comparisons between nested submodels, which contain a subset of the predictors in a larger model. For example, we can compare models with and without age as a predictor to determine the LR for age, or compare models with and without a block of predictors, e.g. with and without a set of patient history characteristics. Statistical testing is straightforward between such nested models.

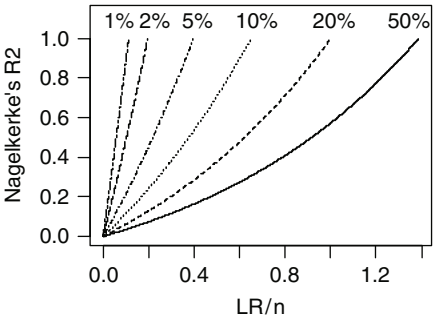

Fig. 4.5 Relationship between Nagelkerke's R^2 and the LR statistic for incidence of the outcome of 1–50%. The LR is divided by n to make the scale independent of sample size. We note a reasonably linear relationship, especially for lower incidences. Largest LRs per subject are possible with an incidence of 50%

The absolute value of the LR depends on n, the number of patients, similar to the sum of squares in linear regression analysis. Several attempts have been made to define an R^2 measure for generalized linear models, relating LR to $-2\ LL_0$. R^2 values ideally enable direct comparison across predictors, irrespective whether the predictor was categorical or continuous, and independent of the sample size. A nowadays popular definition of R^2 uses the LR and $-2\ LL_0$ as follows:

$$R^2 = (1 - \exp(-LR/n)) / (1 - \exp(- - 2\ LL_0 / n)),$$

where n is the number of patients.

This definition of R^2 was proposed by Nagelkerke, and has the advantage of being scaled between 0 and 100%.[309] For a perfect model, LR = $+2\ LL_0$, and R^2 = 100%. The relationship between the LR statistic and Nagelkerke's R^2 is more or less linear (Fig. 4.5).

We will use the Nagelkerke definition of R^2 throughout this book. The scaling between 0 and 100% makes it a natural measure to indicate how close we are with our predictions to the observed 0 and 1 outcomes (Fig. 4.6). The calculation is based on the LL scale, which is the scale used in the fitting process to optimize the model given the data. The calculation includes the LR, which is the theoretically preferred quantity for testing of significance in logistic models.

4.2.3 Models Related to Logistic Regression

Logistic regression can be viewed as an improvement over linear discriminant analysis, which is an older technique.[170] Discriminant analysis usually makes more

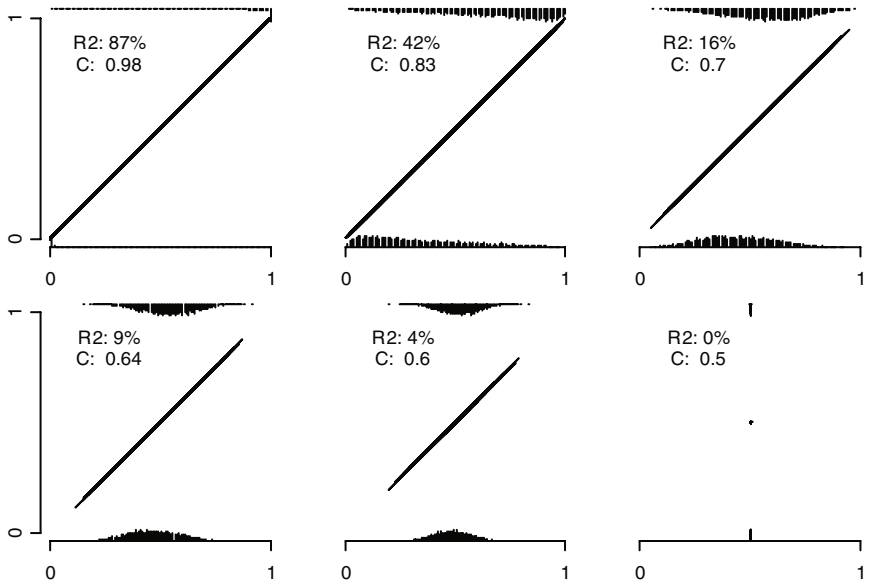

Fig. 4.6 Distribution of observed outcomes (0 or 1), in relation to predicted probabilities from logistic models relating y to a predictor x. The y variable was created from the linear regression example in Fig. 4.1 by dichotomization, and had an average incidence of 50%. We note that Nagelkerke's R^2 values for logistic regression are slightly smaller than the Pearson R^2 values for linear regression in Fig. 4.1. Discrimination is indicated by the c statistic (equivalent to the area under the receiver operating characteristic curve, see Chap. 15)

assumptions on the underlying data, for example multivariate normality, which is not the case in logistic regression. The data need to follow a binomial distribution, which is a natural assumption for 0/1 data. However, when correlations between outcomes exist, for example because of grouping of patients within hospitals, this assumption may be violated. Generalized estimation equations (GEE) are an extension of logistic regression for correlated data.[322, 472]

4.2.4 Bayes Rule

Bayes rule has often been used in a diagnostic context for the prediction of the likelihood of an underlying disease.[331] A prior probability of disease (p(D)) is considered before information becomes available (e.g. from history taking, or from a diagnostic test, denoted as predictor x). The information is used to calculate a posterior probability of disease (p($D|x$)).

This approach has been followed with some success in the 1970s by De Dombal in deriving diagnostic estimates for patients with abdominal pain.[92] Probabilities were estimated with a Bayesian approach, where the prior probability of a diagnosis was updated with information from a large database. This database contained data on the prevalence of signs and symptoms according to the outcome diagnosis. This information can efficiently be summarized with diagnostic likelihood ratios ("LR"). The diagnostic LR for a specific sign or symptom x is

$LR(x) = p(x|D) / p(x|!D)$, where D indicates presence of the disease (determined by a reference standard), and $!D$ no disease.

The combination of a prior probability of disease and LR is straightforward with Bayes' formula:

$Odds(D|x) = Odds(D) \times LR(x)$, where
$Odds(D)$ is the prior odds of disease, calculated as $p(D)/(1 - p(D))$.

In logit form the formula reads as:
$Logit(D|x) = Logit(D) + \log(LR(x))$

This looks very similar to the logistic model shown before. The intercept α is replaced by $Logit(D)$, the prior probability of disease, and $\beta_1 \times x_1$ is replaced by $\log(LR(x))$. The term "$\log(LR(x))$" has been referred to as "weight of evidence", since it indicates how much the prior probability changes by evidence from a test.[397]

For a test with a positive or negative result, there is a simple relationship between LR and OR:

$OR = LR(+)/LR(-)$, and
$\log(OR) = \text{coefficient} = \log(LR(+)/LR(-)) = \log(LR(+)) - \log(LR(-))$, where
$LR(+)$ and $LR(-)$ are the LRs for positive and negative test results, respectively.

In a logistic model with one predictor representing the test (+ or − result), the intercept α reflects the $\text{logit}(y)$ when the test is negative. When the test is positive, the change in logodds is given by the coefficient, and $\text{logit}(y)$ = intercept + coefficient.

*4.2.5 Example: Calculations with Likelihood Ratios

Suppose we have a test with 80% sensitivity and 90% specificity, and a prevalence of disease of 10%. For 1,000 patients, the cross-table may look like Table 4.1.

The $LR(+) = p(\text{Test } +|D)/p(\text{Test } +|!D) = 0.8 / 0.1 = 8$.
The $LR(-) = p(\text{Test } -|D)/p(\text{Test } -|!D) = 0.2 / 0.9 = 0.22$. We can calculate the posterior probabilities of disease with the formula $Odds(D|x) = Odds(D) \times LR(x)$. For a positive test, $Odds(D|x) = 100/900 \times 8 = 8/9$. The probability is calculated as $odds/(odds+1) = (8/9)/(8/9 + 1) = 47\%$.
For a negative test results, $Odds(D) \times LR(-) = 100/900 \times 0.2/0.9 = 2/81$, or a probability of 2.4% ((2/81) / (2/81 + 1)).

These numbers can also be calculated directly from the table: prior = 100/1,000 = 10%; posterior 80/180=47% and 20/920=2.4%. On logodds scale the change =

Table 4.1 Cross-tabulation of a test with + or – results with presence of disease (*D* or !*D*)

	D	!*D*	Total
Test +	80	90	180
Test –	20	810	920
Total	100	900	1,000

Table 4.2 Logistic regression analysis for example in Table 4.1

Variable	*b*	SE	OR [95% CI]
Intercept	–3.701	0.226	
Test	3.583	0.274	36 [21–62]

log(8) = +2.1 for a positive test and log(0.22) = −1.5 for a negative test result. The odds ratio is 8/0.22 = 36, and the log(OR) = 2.1 − 1.5 = 3.6.

From a logistic regression analysis, we obtain: intercept = −3.7, coefficient for test is 3.6; OR=36 (Table 4.2). So, the linear predictor is −3.7 for a negative test and −0.1 for a positive test, which corresponds to probabilities of 2.4% and 47%; as expected this is identical to the calculations with LRs, or as directly observed from Table 4.1.

Graphically, we can well illustrate how Bayes' formula works for a positive or negative test result to obtain a posterior probability from a prior probability (Fig. 4.7).

4.2.6 Prediction with Naïve Bayes

Bayes rule is a general scientific approach to handle conditional probabilities, e.g. to obtain p($D|x$) from p($x|D$). The p($x|D$) can sometimes easier be estimated than p($D|x$). For example, sensitivity and specificity of a dichotomous test are estimated conditional on disease status. For prediction, we are however interested in p($D|x$).

De Dombal and others have used a simple method to estimate posterior probabilities for combinations of signs and symptoms.[92] The posterior probability after considering x_1 is used as the prior when considering x_2, etc. This approach is reasonable if the x_1, x_2, etc. are conditionally independent. Usually positive correlations are however present which makes that the effect of x_2 is smaller once x_1 has already been considered, compared to considering x_2 unconditionally. Such violation of conditional independence makes that $\mathrm{LR}\,x_2|x_1(x) < \mathrm{LR}\,x_2(x)$.[217]

This sequential application of Bayes' rule is equivalent to using the univariate logistic regression coefficients in a linear predictor. Because of its simplicity and mathematical incorrectness, Naïve Bayes is sometimes referred to as "Idiot's Bayes".

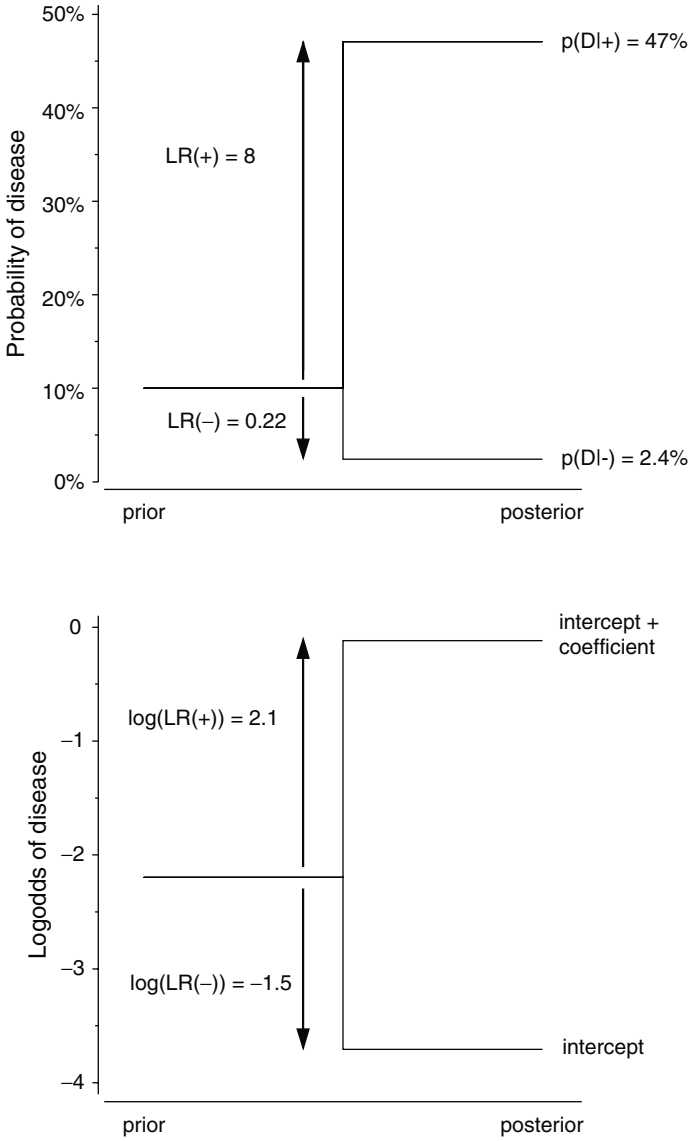

Fig. 4.7 Graphical illustration of Bayes's formula for a prior probability of disease of 10%. Diagnostic LRs of 0.22 and 8 change the posterior probability of disease to 2.4% and 47%, respectively. The second graph shows the probabilities on the logodds scale

The linear predictor reads like

$$\mathrm{Lp}_u = \beta_{1,u} \times x_1 + \beta_{2,u} \times x_2 + \ldots + \beta_{p,u} \times x_p,$$

where the subscript u indicates univariate estimates for the logistic regression coefficients.

*4.2.7 Examples of Naïve Bayes

A naïve Bayes modeling approach has been studied by Spiegelhalter, who found remarkably good performance for discrimination.[395] Also, the method has been applied in modelling the effects of genetic markers, where robustness in modelling is required at the expense of accepting bias in coefficients.[387]

*4.2.8 Calibration and Naïve Bayes

The problem with Naïve Bayes estimation is that correlations between the predictors are ignored. In the case of positive correlations, predictions will be too extreme, since the effects of predictors are overestimated. Both too low and too high predictions arise. This is reflected in a regression coefficient for the linear predictor ("calibration slope", β_{cal}) below 1 in the model: $y \sim lp_u$. A simple approach hence is to correct this calibration problem with a single coefficient for the linear predictor: Logit $(y) = \alpha + \beta_{cal} \times lp_u$.

In terms of multivariable OR (OR_m) or multivariable LR (LR_m), the exponent can be used for easy of notation: $OR_m = OR_u^{\beta cal}$ or $LR_m = LR_u^{\beta cal}$. The idea of recalibrating of the linear predictor comes back in Chap. 15 and 20.

*4.2.9 Logistic Regression and Bayes

The diagnostic LR can be used mathematically correct in a multivariable context. The key trick is to rescale test results. Instead of a "1" for positive and a "0" for negative, the univariate log(LR) values can be filled in for the test results.[395] In a multivariable model, the joined effects for the test results are subsequently estimated. Coefficients for the rescaled test results reflect to degree of correlation between test results from different tests. If there are no correlations, the coefficients of each test would be close to 1.

Multivariable diagnostic LRs can also be calculated by comparing models with and without the test of interest. The model without the test is the prior, and the model with the test included provides the posterior probabilities.[217] Subtracting these two equations provides the LRs.

*4.2.10 More Flexible Approaches to Binary Outcomes

Naïve Bayes estimation is an example of a more simplistic and robust method than logistic regression. A more flexible alternative model is a generalized additive model (GAM), as was already discussed for linear regression models.[180,181,472]

Another alternative is to consider generalized non-linear models. Here, the outcome is no more related to a mathematically simple linear combination of estimated regression coefficients and predictor values. Instead, non-linear combinations of predictors are possible. Generalized non-linear models are currently implemented as neural networks. Neural networks are often presented as fancy tools, "that represent the way our brain works," but it may be more useful to consider them as non-linear extensions of linear logistic models.[436,438]

The most common neural network model is the multilayer perceptron (Fig. 4.8). In such a network, the neurons are arranged in a layered configuration containing an input layer, usually one "hidden" layer, and an output layer. The values of input variables (patient characteristics) are imported into the network via the input layer and multiplied with the weights of the connections. These multiplied values constitute the input of the next (hidden) layer, from where the process is continued to produce the output variables (e.g. risk of mortality) in the output layer.

A neural network does not use any preliminary information about the links between the input and output variables; the relationships between input and output variables are determined by the data. It is hence not easily possible to explicitly force external knowledge into a model, e.g. that an age effect should be monotonically increasing. Neural networks learn by example; the errors from the initial prediction for the patients are fed back into the network and the weights for connections are adjusted to minimize the error; for the second time predictions are made and compared to the actual outcome. The process from input to output layer is repeated many times. However, to prevent "overtraining" the repetitions are usually stopped before the network is fully trained to the data.[410,436]

The hidden layer makes the network more flexible to recognize patterns in the data compared to a standard logistic regression model. The number of hidden layers and number of nodes are chosen by the analyst. A neural network without a hidden layer is equivalent to a logistic regression model.[436,438]

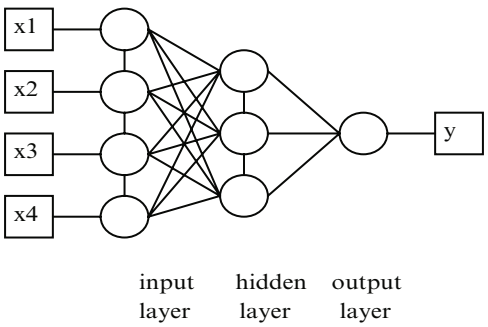

Fig. 4.8 A simple neural network with four input variables (predictors $x_1 - x_4$), one hidden layer with three nodes, and one output layer (outcome y)

4.2.11 Classification and Regression Trees

Recursive partitioning or Classification And Regression Tree ("CART") methods have been promoted by some as strong tools for predictive modeling. Recursive partitioning is a statistical method to construct binary trees.[57] The method is based on statistically optimal splitting ("partitioning") of the patients into pairs of smaller subgroups. Splits are based on cut-off levels of the predictors, which produce maximum separation among two subgroups and a minimum variability within these subgroups with respect to the outcome. The predictor causing the largest separation is situated at the top of the tree, followed by the predictor causing the next largest separation, and so on. Splitting continues until the subgroups reach a minimum size or until no improvement can be obtained. Several variants of recursive partitioning algorithms are available which use different criteria to construct a tree. Details of the statistical procedures can be found elsewhere.[57]

*4.2.12 Example: Mortality in Acute MI Patients

We illustrate the creation of a tree in patients with an acute myocardial infarction (MI). We use a data set from the GUSTO-I trial (see Chap. 22) which is labeled "sample5". It contains 429 patients, of whom 24 died by 30 days. We consider the predictors age (continuous) and Killip (4 categories, Fig. 4.9). An initial tree was quite complex, with many splits, especially at many age cut-offs. Some counter-intuitive patterns arose, such as a zero mortality among older patients within subbranches. A technique to construct better prediction trees is to prune a tree back to an "optimal" size. This can be achieved by using a cross-validation procedure (see also Chap. 17). Performance is determined in randomly drawn independent parts of the data for different tree sizes (Fig. 4.10). A pruned tree of size 3 was subsequently created (Fig. 4.11). So an enormous reduction in size was necessary to construct a more stable tree. Prediction of outcome for a new patient is accomplished by simply running that patient down the tree, according to the values of the predictors.[57]

4.2.13 Advantages and Disadvantages of Tree Models

An advantage of a tree is its simple presentation. Some claim that a tree represents how physicians think: starting with the most important characteristic, followed by another characteristic depending on the answer on the first, etc. Indeed, humans are remarkably quick in pattern recognition based on a few clues. However, humans have typically been outperformed by systematic prediction methods in experiments where a balanced, quantitative judgement was required, such as estimation of a probability based on a set of characteristics.[265] So, the fact that a tree may represent human thinking for classification does not argue in favour of the method for prediction. A true advantage may be that interaction effects are naturally incorporated in a tree, while a standard logistic regression

Fig. 4.9 Initial tree fitted in a small subsample of GUSTO-I ("sample5") with age and Killip class as predictors. Splits in the tree are labelled with the criterion for the split, e.g. Killip <2.5 indicates that patients with Killip class 1 or 2 go to the left in the tree and patients with Killip class 3 or 4 go to the right. The nodes are labeled with 30-day mortality as a fraction, e.g. 0.60 indicates a 60% mortality among those with Killip class 3 or 4. Vertical distances in the tree are based on the statistical improvement between parent and children nodes

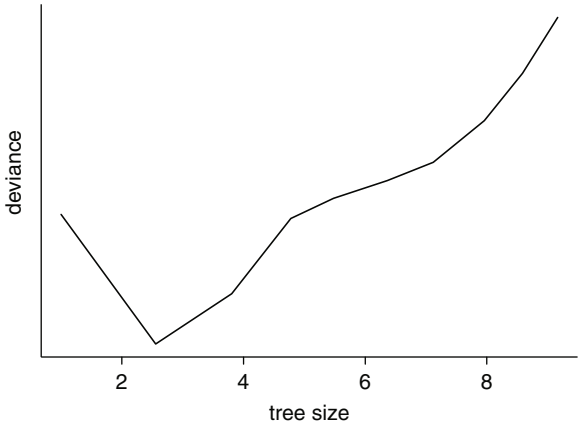

Fig. 4.10 Cross-validated deviance in relation to tree size; optimal size is around 3

model usually starts with main effects, that is one coefficient β_i per predictor. When multiple, high-order interactions are expected in a huge data set, and only categorical predictors are considered, a tree might be a good choice. Such situations may be rare in medical data, but may possibly be encountered in other areas of research.

Disadvantages of trees can be noted by considering a tree as a special case of linear logistic regression. First, all continuous variables have to be categorized, which implies a loss of information. As illustrated in Figs. 4.9 and 4.11, age is con-

Fig. 4.11 Pruned tree with size 3 for terminal nodes

sidered with different splits at different places in the tree, while the age effect could well be approximated with a single linear term in a logistic model (see Chap. 6). Moreover, these cut-points are determined from a search over all possible cut-points, which is well known to be very dangerous in a prediction context.[12]

Further, the tree assumes interactions between all predictors. After the first split, this interaction is of the first order, i.e. $x_1 \times x_2$. At the third level, second-order interactions are assumed (x1 × x2 × x3). In regression analysis, it is common practice to include main effects of predictors when interactions are considered; this principle is not followed in tree modelling. A higher-order interaction term is included to model the effect of a predictor in a specific branch, and simply omitted from the other branches. A predictor is typically selected in one branch of the tree and not in another. This poses a clear risk of testimation bias (Chap. 5): predictors are selectively considered when their effects are relatively large, and not if their effects are small.

*4.2.14 Trees as Special Cases of Logistic Regression Modelling

From a model selection viewpoint, trees have three distinctive characteristics compared to a logistic regression model when we consider a set of potential predictors.

1. In a logistic model, a default strategy is to include all predictors as main effects. This model can be extended with interaction terms if the power to examine these is sufficient. It is rare to study interactions that are more complex than considering three variables (second order). In contrast, trees by default assume that higher-order interaction are present, and cannot model main effects.

2. Continuous variables should not be categorized in regression models.[355] Trees do so by necessity, which causes a loss of information.
3. One might use a stepwise selection method in a logistic model, especially in larger data sets with sufficient power to select all relevant predictors. Generally a high *p*-value is advisable to prevent various problems (Chap. 11).[174] A tree however always needs to be selective in the inclusion of predictors, and quickly runs out of cases within branches. Limited power is a major problem in the development of trees.

As an example we write the linear predictor for the tree in Fig. 4.11 as:

$$\text{Lp} = \beta_1 \times \text{Killip} > 2 + \beta_2 \times \text{Killip} \leq 2 \times \text{age} \leq 67.5 + \beta_3 \text{ Killip} = 1 \times \text{age} > 67.5 + \beta_4 \times \text{Killip} = 2 \text{ age} > 67.5.$$

We estimate four parameters which identify the four terminal nodes. If we want a more standard formulation with an intercept we could write:

$$\text{Lp} = \alpha + \beta_1 \times \text{Killip} > 2 + \beta_2 \times \text{Killip} = 1 \times \text{age} > 67.5 + \beta_3 \times \text{Killip} = 2 \times \text{age} > 67.5,$$

where the intercept term refers to patients with Killip 1 or 2, and age = 67.5.

In this formulation, it is clear that age is ignored among those with Killip class > 2, and that a dichotomized age variable is used in interaction with patients in Killip class 1 or 2.

In a logistic regression model, we could combine Killip class 3 and 4 (representing "shock"), and omit the interaction of Killip with age:

$$\text{Lp} = \alpha + \beta_1 \times \text{age} + \beta_2 \times \text{Killip}=1 + \beta_3 \times \text{Killip}=2 + \beta_4 \times \text{Killip} > 2.$$

Even simpler, we could include Killip as a linear rather than as a categorized predictor:

$$\text{Lp} = \alpha + \beta_1 \times \text{age} + \beta_2 \times \text{Killip}.$$

We could extend this model to allow for age × Killip interaction:

$$\text{Lp} = \alpha + \beta_1 \times \text{age} + \beta_2 \times \text{Killip} + \beta_3 \times \text{age} \times \text{Killip}.$$

*4.2.15 Other Methods for Binary Outcomes

Various other methods are available or under development. Such methods include multivariate additive regression splines (MARS) models. These form a kind of hybrid between generalized additive models and classication trees.[129] MARS models aim to find low-order additive structure as well as interactions between risk factors.

A support vector machine (SVM) performs classification tasks by constructing hyperplanes in a multidimensional space that separates cases of different class labels. SVM supports both regression and classification tasks and can handle multiple continuous and categorical variables.[464] Specialized texts are available that discuss these and other statistical models for binary data.[181]

4.2.16 Summary on Binary Outcomes

In sum, logistic regression provides a quite flexible model to derive predictions from empirical data. Interactions and nonlinearity can be incorporated. Some other models, such as GAM, neural nets (GNLM), MARS, can be seen as extensions, with the default linear logistic model as a special case. Naïve Bayes is a simplified version of logistic regression, ignoring correlations between predictors. Trees can be seen as special cases of logistic regression, requiring categorizations of continuous variables and assuming higher order interactions.

4.3 Categorical Outcomes

Categorical outcomes without a clear ordering are common in diagnostic medical problems. The diagnostic process starts with considering presenting signs and symptoms of a patient. This leads the physician towards a set of differential diagnoses. Each diagnosis has a probability given the patient's clinical and nonclinical profile. Usually, one of these differential diagnoses is defined as the working diagnosis or target disease, to which the diagnostic work-up is primarily directed. Consequently, diagnostic studies commonly focus on the ability of tests to include or exclude the presence of this target disease. The alternative diagnoses (which may all direct different treatment decisions) are thus included in the outcome category "target disease absent." After dichotomization of the diagnostic outcome, we may develop diagnostic prediction rules with logistic regression analysis. However, considering only the target disease is a simplification of clinical practice.

Table 4.4 Characteristics of some statistical models for binary outcomes

Categories	Interactions	Linearity	Selection	Estimation
Linear logistic regression	Possible	Flexible	Flexible	Standard ML or penalization
Idiot's Bayes	No	Often categories for diagnostic outcome	Flexible	Univariate effects (+ calibration slope)
GAM	Possible	Highly flexible	Flexible	Nonparametric, close to penalized ML
GLNM, neural net	Assumed	Highly flexible	Flexible	Backpropagation, early stopping to prevent overfitting
Trees	Assumed	Categorization	Assumed	Various splitting methods

4.3.1 Polytomous Logistic Regression

Several studies discussed the use of polytomous logistic regression to accommodate simultaneous prediction of three or more unordered outcome categories.[29,484] The model for j outcome categories can be written as:

$$\text{Logodds}(y=j \text{ vs. } y=\text{reference}) = \alpha_j + \beta_{i,j} \times x_{i,j} = \text{lp}_j$$

where $j - 1$ models are fitted each with separate sets of intercept α_j and regression coefficients β_t. We illustrate the polytomous model for prediction of three diagnostic outcome categories in a detailed case study.

*4.3.2 Example: Histology of Residual Masses

After chemotherapy, patients with nonseminomatous testicular germ cell tumor may have residual masses of metastases.[425] These residual masses may contain benign tissue, mature teratoma, or cancer cells. Surgery is not necessary for benign tissue. Mature teratoma can grow and hence cause problems during follow-up. The most serious diagnosis is residual cancer, where a direct benefit from surgery is plausible.

We consider three outcome categories with varying therapeutic benefit: no benefit for benign tissue, some for teratoma, and most benefit for surgical removal of residual cancer.[35] We have proposed to weigh the benefit as 1:3:8 based on expert estimates of the prognosis of unresected vs. resected masses.[422] This ordering in severity of the outcome was not used in the modeling, since biological knowledge was available that implied that prognostic relationships would be very different for the different histologies. For example, some histologies are known to produce certain tumor markers while others do not. Masses with teratoma masses are not expected to decrease substantially in size by chemotherapy, while cancer is usually responsive. Hence, a substantial decrease would make residual cancer unlikely.

Polytomous logistic regression analysis requires that one of the outcome categories is chosen as reference category. For the other outcome categories the polytomous logistic regression analysis fits simultaneously submodels that compare the outcome categories with the chosen reference. Thus, for each outcome category, different regression coefficients are estimated for each predictor. These submodels together comprise the polytomous model and can be used to estimate the probability of presence of each diagnostic outcome. In our example study, the reference diagnosis was viable cancer. Hence, we fitted a polytomous regression model, consisting of two submodels, one for benign tissue compared to viable cancer, and one for mature teratoma compared to viable cancer. These models take a similar form as the binary logistic model:

$$\begin{aligned}&\text{Logit(benign vs. cancer)} = \alpha_b + \beta_{1,b} \times x_1 + \beta_{2,b} \times x_2 + \ldots + \beta_{p,b} \times x_p = \beta_{i,b} \times X = \text{lp}_b;\\ &\text{Logit(teratoma vs. cancer)} = \alpha_t + \beta_{1,t} \times x_1 + \beta_{2,t} \times x_2 + \ldots + \beta_{p,t} \times x_p = \beta_{i,t} \times X = \text{lp}_t.\end{aligned}$$

The subscript b indicates that we predict the odds of benign tissue, and subscript t for teratoma with p predictors.

The interpretation of the regression coefficients is similar as for dichotomous logistic regression, i.e., the logodds of the outcome (benign tissue or mature teratoma) relative to cancer per unit change in the predictor values. The probabilities of benign and teratoma tissue can be calculated by:

$$P(\text{benign tissue}) = \exp(\text{lp}_b) / [1 + \exp(\text{lp}_b) + \exp(\text{lp}_t)]$$
$$P(\text{mature teratoma}) = \exp(\text{lp}_t) / [1 + \exp(\text{lp}_b) + \exp(\text{lp}_t)].$$

As probabilities need to sum to 1, the probability of cancer can then be calculated by:

$P(\text{cancer}) = 1 - P(\text{benign tissue}) - P(\text{mature teratoma})$.

We fitted a multivariable polytomous logistic regression model with six predictors to enable estimation of the probabilities of benign tissue, mature teratoma, and viable cancer. Variable selection was not applied; we simply included all six of the available predictors.

*4.3.3 Alternative Models

For comparison reasons, we may fit consecutive multivariable dichotomous logistic models. In our example, we make one model to predict benign tissue (vs. mature teratoma or viable cancer). The second, consecutive, model aimed to predict the odds of mature teratoma vs. viable cancer in patients who did not have benign tissue.

$$\text{Logit}(\text{benign vs. teratoma/cancer}) = \alpha_b + \beta_1 \times x_1 + \beta_2 \times x_2 + \ldots + \beta_p \times x_p = \beta_i \times X = \text{lp};$$
$$\text{Logit}(\text{teratoma vs. cancer}) = \alpha_t + \beta_{1,t} \times x_1 + \beta_{2,t} \times x_2 + \ldots + \beta_{p,t} \times x_p = \beta_{i,t} \times X = \text{lp}_t.$$

The latter formula is identical to a previous formula for the polytomous model, but the coefficients are estimated differently. In the polytomous model, all coefficients are estimated jointly. In the consecutive logistic model, a selection of patients is made to estimate the coefficients.

With these two binary logistic models the diagnostic probabilities are calculated by:

$P(\text{benign tissue}) = \exp(\text{lp}) / (1 + \exp(\text{lp}))$
$P(\text{mature teratoma}) = (1 - P(\text{benign tissue})) \times \exp(\text{lp}_t) / [1 + \exp(\text{lp}_t)]$
$P(\text{cancer}) = 1 - P(\text{benign tissue}) - P(\text{mature teratoma})$

In our example, we use the same six predictors, but in principle we could select different predictors for lp and lp_t. Also, we could have considered different transformations of the continuous predictors related to LDH and mass size.

In both approaches, 14 parameters were estimated: 2 intercepts (α) and 2 sets of 6 regression coefficients ($\beta_{1:6}$). The performance of the two approaches was very similar according to discrimination (area under ROC curve) and R^2 measures. See Biesheuvel et al. for a more detailed decription of this case study.[35] Further discussion of approaches to unordered outcomes is provided in other reports.[352, 425]

*4.3.4 Comparison of Modelling Approaches

We considered a total of 1,094 patients, where 425 (39%) had benign tissue, 535 (49%) mature teratoma, and 134 (12%) viable cancer. Table 4.5 shows the distributions of the six predictors across the three diagnostic outcome categories and in the total study population.

The odds ratios for the predictors are shown in Table 4.6, considering a polytomous regression model, and a consecutive logistic model. We note that the odds ratios for teratoma vs. cancer differ slightly between these modeling approaches. The odds ratios for necrosis vs. cancer are larger for most predictors than for necrosis vs. other histology.

4.4 Ordinal Outcomes

Ordinal outcomes are quite common in medical and epidemiological studies. Often, such scales are either simplified to binary outcomes, or treated as continous outcomes. As an example, we consider the Glasgow Outcome Scale (GOS).[430] This scale has five levels (Table 4.7).

This scale has often been dichotomized as mortality vs. survival, or a unfavorable (GOS 1, 2 or 3) vs. favorable (GOS 4 or 5) outcome. However, we can also explore the use of the full GOS. A practical consideration is that the GOS 2 category is very small, and that some may debate whether vegetative state is better than death. Therefore we combine the GOS categories 1 and 2, such that an outcome with four ordered levels is formed.

Table 4.5 Distribution of predictors across outcome categories in the total study population (n = 1,094)

	Benign	Mature teratoma	Viable cancer	Total
	N (%)	*N* (%)	*N* (%)	*N* (%)
Predictors				
No teratoma in primary tumor	279 (55)	170 (34)	54 (11)	503 (46)
Normal AFP level	200 (59)	112 (33)	27 (8)	339 (31)
Normal HCG level	184 (49)	154 (41)	40 (10)	378 (35)
Standardized value of LDH*	1.5 (0.39–70)	1.2 (0.12–21)	1.8 (0.34–64)	1.4 (0.12–70)
Postchemotherapy size (mm)*	18 (2–300)	30 (2–300)	40 (2–300)	28 (2–300)
Reduction in size (%)*	60 (−150–100)	20 (−150–100)	43 (−250–100)	43 (−250–100)
Outcome				
Histology at resection	425 (39)	535 (49)	134 (12)	1,094 (100)

* Median (range)

AFP Alpha-fetoprotein, *HCG* Human chorionic gonadotropin, *LDH* Lactate dehydrogenase

Table 4.6 Results of the multivariable polytomous and consecutive dichotomous logistic regression analysis. Values represent odds ratios with 95% confidence intervals

	Polytomous regression		Consecutative dichotomous regression	
Predictor	Benign vs. cancer	Teratoma vs. cancer	Benign vs. other	Teratoma vs. cancer
No teratoma in primary tumour	2.2 (1.4–3.3)	0.66 (0.44–0.99)	3.0 (2.2–4.0)	0.61 (0.40–0.92)
Normal AFP serum level	2.8 (1.7–4.6)	0.94 (0.57–1.5)	2.9 (2.1–4.0)	0.90 (0.54–1.5)
Normal HCG serum level	1.4 (0.89–2.3)	0.72 (0.46–1.1)	1.9 (1.3–2.6)	0.70 (0.44–1.1)
Log of standardized value of LDH	1.2 (0.84–1.6)	0.58 (0.42–0.78)	1.7 (1.4–2.2)	0.60 (0.44–0.81)
Square root of postchemotherapy mass size	0.79 (0.71–0.88)	0.91 (0.84–0.99)	0.85 (0.77–0.92)	0.89 (0.82–0.98)
Reduction in mass size (per 10%)	1.14 (1.06–1.22)	0.97 (0.92–1.02)	1.18 (1.12–1.24)	0.96 (0.92–1.0)

Table 4.7 Definition of the Glasgow Outcome Scale

Category	Label	Definition
1	Dead	–
2	Vegetative	Unable to interact with environment; unresponsive
3	Severe disability	Conscious but dependent
4	Moderate disability	Independent, but disabled
5	Good recovery	Return to normal occupational and social activities; may have minor residual deficits

4.4.1 Proportional Odds Logistic Regression

A standard logistic regression model can be used for each of the three possible dichotomous categorizations of the GOS: 12 (dead/vegetative) vs. 345, 123 vs. 45 (favorable), 1234 vs. 5 (good recovery). A straightforward extension of the logistic model is the proportional odds logistic model. Here, a common set of regression coefficients is assumed across all levels of the outcome, and intercepts are estimated for each level. So, in our example we have three intercepts α, but only one set of β, instead of three sets of β coefficients when fitting a polytomous logistic model. The common set of β coefficients can be thought of as an average over the three separate sets of βs estimated at each possible dichotomization. As an example we consider a simple model with age, motor score, and pupillary reactivity in a model to predict 6-month outcome in data from two RTCs in traumatic brain injury.[203]

An advantage of the proportional odds model is its parsomony in dealing with an ordered outcome. The price we pay is the assumption of proportionality of the

odds. This assumption is equivalent to saying that any cut-point on the outcome scale would lead to the same logistic regression coefficient. The model further has very similar assumptions as the usual logistic model. We can graphically check the proportionality assumption in univariate analyses for each predictor (Fig. 4.12). Distances between points should be identical on the logit scale within each category of a predictor (looking horizontally), or equivalently, the effects of predictors should be the same for every point (looking vertically). The assumption of proportional odds can formally be assessed with a score test. One could also develop usual logistic models by each categorization, and check for systematic trends in the estimated odds ratios (Table 4.8). There is considerable overlap in patients in such evaluations, but clear deviations from proportional odds should become visible. In

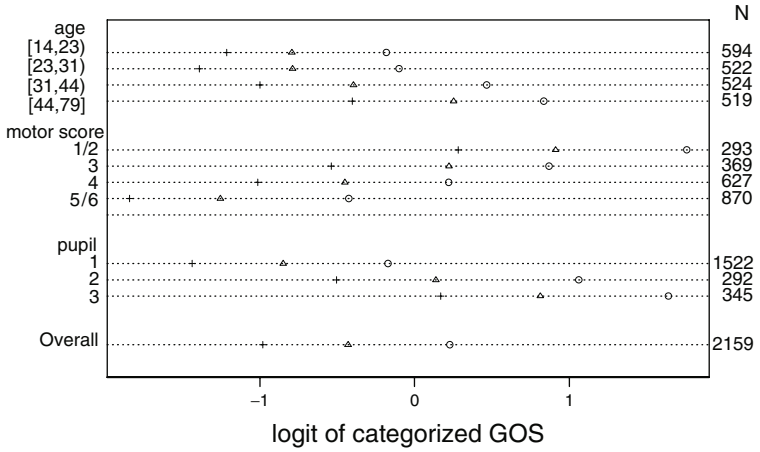

Fig. 4.12 Assessment of the proportional odds assumption for each of three predictors (univariate analysis) to predict for GOS at 6 months after traumatic brain injury. Data from the Tirilazad trials (n=2,159). The *circle*, *triangle*, and *plus* sign correspond to the GOS categorizations 12 vs. 345, 123 vs. 45, and 1234 vs. 5. For example, the overall logit of the last categorization is −1, or a probability of 27% (589/2,159 patients). The proportional odds assumption is well satisfied, since the horizontal distance between the points is constant within each category

Table 4.8 Logistic and proportional odds models for GOS at 6 months after traumatic brain injury in 2,159 patients from the Tirilazad trials[203]

Categorization	12 vs. 345	123 vs. 45	1234 vs. 5	Proportional
Age (per decade)	1.36	1.47	1.45	1.43
Motor 1/2	5.88	6.50	6.18	5.86
3	2.98	3.82	3.00	3.15
4	1.95	1.95	1.62	1.82
5/6	1	1	1	1
Pupils 2 reactive	1	1	1	1
1 reactive	1.73	1.81	2.51	2.01
Nonreactive	3.26	3.53	4.23	3.55

our example, the ORs per categorization are reasonably constant, and the proportional odds ratio provides a nice summary measure over the three categorizations.

*4.4.2 Alternative: Continuation Ratio Model

An alternative to the proportional odds model is the continuation ratio model. This model is related to the Cox proportional hazards model and allows predictors to have different effects on different levels of the ordinal outcome. An extensive illustration is provided by Harrell et al.[174,178]

4.5 Survival Outcomes

Survival analysis is appropriate for outcomes that occur during follow-up of patients. The outcome may for example be death or another event, such as recurrence of disease in cancer, or a complication after implantation after a heart valve. A key characteristic of survival data is that the follow-up of patients is typically incomplete. For example, some patients may have been followed for 1 year, others for 3 years, etc., while we may be interested in estimates of 5-year survival. Patients with such incomplete data are called censored observations. Because of censoring, logistic regression for the outcome (a binary variable) is inappropriate. One could think of linear regression on the survival time (a continuous outcome), but again censoring makes such an analysis usually meaningless.

4.5.1 Cox Proportional Hazards Regression

In medical and epidemiological studies, the Cox proportional hazard model is the most often used method for survival outcomes.[85] It is the natural extension of the logistic model to the survival setting. Indeed, the Cox model is equivalent to conditional logistic regression, with conditioning at times where events occur.[251] In the logistic model, we use an intercept in the linear predictor, while in the Cox model a baseline hazard function is used. The hazard function indicates the risk of the outcome during follow-up. The baseline hazard is nonparametric in the Cox model. As for the logistic model, simpler and more extensive methods exist, which can be seen as special cases or extensions of the Cox model.

The Cox regression model is often stated as a function of the hazard function[472]:

$$\lambda(t|X) = \lambda(t)\, e^{\beta X},$$

Where $\lambda(t)$ is the hazard at time t, and is usually estimated at the mean values of the predictors and βX is the linear predictor, $\beta_1 \times x_1 + \beta_2 \times x_2 + \ldots + \beta_p \times x_p$.

The linear predictor is usually centered at the mean values of the predictors, and $e^{\beta X}$ then indicates the hazard ratio compared to the average risk profile. Note that the linear predictor relates to the log of the hazard:

$$\log(\lambda(t|X)) = \log(\lambda(t)) + \beta_1 \times x_1 + \beta_2 \times x_2 + \ldots + \beta_p \times x_p.$$

The Cox regression model is semiparametric. It makes a parametric assumption on the effect of predictors, i.e., proportionality of effect during follow-up. The baseline hazard function $\lambda(t)$ is nonparametric. This is an advantage of the model, especially when we focus on the effect of predictors. Regression coefficients β_i can readily be estimated. The quantity e^{β_i} is the hazard ratio, similar to the odds ratio in logistic regression.

4.5.2 Predicting with Cox

When we want to make predictions, we need to consider the risk over time, for example by using the cumulative hazard, or survival function. The standard formulation of the predicted survival at time t, given a set of predictors X, is as

$$S(t|X) = S(t)^{e(\beta X)},$$

Where $S(t|X)$ denotes the predicted survival at time t, given a set of predictors X, $S(t)$ is the baseline survival, usually estimated at the mean values of the predictors, and βX is the linear predictor.

The baseline survival is estimated from the nonparametric baseline hazard function as

$$S(t) = e^{-\Lambda(t)}$$

where $\Lambda(t)$ is the cumulative hazard at time t.

Note that $\log(\Lambda(t))$ can range between [−inf, +inf]; $\Lambda(t)$ [0, inf]; $S(t)$ [1, 0]. This is very similar to the behavior of quantities in logistic regression: logit, odds, and probability. The baseline survival in the development data determines the precise time points where we can make predictions for, which is not very natural for application of the model in new subjects.

4.5.3 Proportionality Assumption

The effect of predictors is assumed to be constant in time or more precisely stated: the hazards are assumed to be proportional. The proportionality assumption can be assessed in a number of ways, including graphical and analytical methods. A general approach is to calculate interval specific hazard ratios. With proportional hazards, the hazard ratio should be similar across any interval considered. Follow-up time can also be considered as a continuous variable, where assessing interaction with log(time) may be a useful approach.[174]

If we find that the effect of a predictor is nonproportional, we can stratify for categorical variables in the baseline hazard. For example, we could estimate baseline hazards for males and females seperately. For continuous predictors, e.g., age, we could specify interactions with log(age) as the time variable. Nonproportionality can also be visualized in a more nonparametric approach, i.e., with Kaplan–Meier curves.

4.5.4 Kaplan–Meier Analysis

Kaplan–Meier analysis is a nonparametric approach to survival outcomes.[224] It adequately deals with censored data, and provides attractive graphs on the relationship between predictor values and the outcome over time. The method can be seen as an extension of a cross-table for survival data. More technically, it can be interpreted as a Cox model with stratification of the baseline hazard to all predictor levels. For example, we could make a Cox model with sex as a stratification variable for the baseline hazard, without any other variables, which is equivalent to a Kaplan–Meier analysis with sex as a predictor. Also, testing in a Kaplan–Meier analysis is usually done with a log-rank test, which is equivalent to the Score test in the Cox model.

Kaplan–Meier analysis often has a role in prognostic modeling at the start of the analysis, i.e., to show univariate relationships graphically or to compute survival fractions at a certain time of follow-up. Also at the end of a modeling process, Kaplan–Meier curves are often used to present the predictions from the model. It is then necessary to group patients by their predictions, since Kaplan–Meier analysis cannot handle continuous predictors. Kaplan–Meier curves are for survival analysis what cross-tables are for binary or categorical outcomes.

*4.5.5 Example: NFI After Treatment of Leprosy

Nerve-function impairment (NFI) commonly occurs during or after chemotherapy in leprosy. It is the key pathological process leading to disability and handicap. A simple clinical prediction rule was developed with 2,510 patients who were followed-up for 2 years in Bangladesh.[87] In total, 166 patients developed NFI (Kaplan–Meier 2-year estimate: 7.0% [95%CI 6.0–8.0%]. A Cox regression model included two strong predictors (Table 4.10). Patients with no, one, or two unfavora-

Table 4.10 Multivariable hazard ratios from Cox proportional hazard analysis.[87] Three risk groups could be formed based on presence of no, one, or two unfavorable predictive characteristics, since the hazard ratios were very similar

Predictor	Hazard ratio [95% CI]
Leprosy group (MB vs. PB)	7.5 (5.3–11.0)
Nerve-function loss at registration	8.1 (5.7–12.0)

ble predictive characteristics had 1.3% (95% CI 0.8–1.8%), 16.0% (12–20%), and 65% (56–73%) risks of developing NFI within 2 years of registration, respectively.

4.5.6 Parametric Survival

Whereas Kaplan–Meier analysis represents a more nonparametric approach, parametric survival models are less flexible than Cox regression in their dealing with the baseline hazard function. Parametric models typically assume proportionality of the predictor effects, but a more smoothed hazard in time. Examples of parametric models include the exponential model (or Poisson model, using a constant hazard) and the Weibull model (two parameters to let the hazard increase or decrease monotonically over time). The exponential and Weibull model can also be seen as examples of accelerated failure time (AFT) models. Here, the effects of predictors are not viewed as multiplicative on the hazards scale, but as multiplicative on the time axis (or additive at the log-time axis). Other examples of AFT models are the log-normal and log-logistic model.[174 472]

Regression coefficients in exponential or Weibull models are hazard ratios after exponentiating. In AFT models, they represent a change in the log-time. The advantage of parametric survival models is their concise, parsimonious formulation, and smoothing of the underlying hazard. This makes these models especially to be considered for prediction purposes. Extrapolation is readily possible with parametric models, but not with Cox or Kaplan–Meier analysis because of their nonparametric nature. Predictions at the end of the follow-up are quite unstable with Cox or Kaplan–Meier analysis, and more robust with parametric methods. For estimation of the effect of predictors, the Cox model is often more suitable, since this model is less restrictive than an exponential or Weibull model. However, log-logistic models have been useful in situations where predictors worked especially during an early, acute phase of the hazard, which would show as non–proportional hazards in a Cox model.[174] Note finally that some of the more flexible methods for binary data have also been extended to survival models, but are not commonly used yet (e.g., neural networks).[181]

*4.5.7 Example: Replacement of Risky Heart Valves

In Chap. 2, we presented an overview of the decision dilemma on Björk-Shiley convexo-concave (BScc) mechanical heart valves.[448] Poisson regression models were constructed to estimated survival and the risk of strut fracture.[415] Poisson regression was especially useful to disentangle the effects of increasing age of the patient during follow-up from the increasing time since implantation of the valve during follow-up. The follow-up time was divided in yearly intervals, each with an age and time since implantation. Time since implantation started at zero, and increased

Table 4.11 Common statistical models for survival outcomes

Categories	Proportionality	Baseline hazard
Cox proportional hazards	Assumed	Nonparametric
Kaplan–Meier	No	Nonparametric
Exponential and Weibull	Assumed	Parametric
Log-normal, log-logistic	No, but multiplicative in time	Parametric

in steps of 1 year during follow-up. Age started at the age at implantation, and also increased in steps of 1 year during follow-up. The Poisson model could easily estimate the effects of both predictors, which would have been more complicated in a Cox regression analysis. Moreover, extrapolation to longer time since implantation was readily possible with the Poisson model.

4.5.8 Summary on Survival Outcomes

In sum, the Cox regression model provides a default framework for prediction of long-term prognostic outcomes. Kaplan–Meier analysis provides a nonparametric method, but requires categorization of all predictors. It is the equivalent of cross-tables for categorical outcomes for a survival context. Parametric survival models may be useful for predictive purposes because of their parsimony and robustness, for example at the end of follow-up, or even beyond the observed follow-up.

4.6 Concluding Remarks

Regression models are available for several types of outcome that we may want to predict, such as continuous, binary, unordered categorical, ordered categorical, and survival outcomes. The corresponding default regression models are the linear, logistic, polytomous, proportional odds, and Cox regression models, respectively. Both more and less flexible methods are available. Flexible methods may fit particular patterns in the data better, but may on the other hand lead to overfitting (Chap. 5). It is therefore not immediately clear what kind of model is to be preferred in a specific prediction problem (Chap. 6).

Special types of data can be encountered that required specific types of analyses. Correlated outcome data may occur by the design of a study, for example by clustering per hospital. In survival analysis, repeated and correlated events may occur, asking for extensions of the Cox model. Also, we may want to consider competing risks in estimation of actual risk instead of actuarial risks.[124,158,159]

Questions

4.1 Explained variation
- (a) What is the difference between explained variation in linear and logistic regression models?
- (b) Is the choice of scale for explained variation natural in linear and logistic regression models?
- (c) Why are larger likelihood ratios seen with an incidence of 50% compared to 1% in Fig. 4.5?

4.2 Categorical and ordinal outcomes
- (a) What is the proportionality assumption in the proportional odds model?
- (b) Mention at least two ways how the proportionality assumption can be checked
- (c) Would the proportionality assumption hold in the testicular cancer case study (Table 4.6)?
- (d) We could also make two logistic regression models for the testicular cancer case study, with one model for benign vs. other and another for cancer vs. other. What would be the problem with predictions from these models?

4.3 Parametric survival models
- (a) Why may we label the Cox regression model "semiparametric"?
- (b) Do you agree that Kaplan–Meier analysis is a fully nonparametric model?
- (c) Why is the Weibull model attractive for making long-term predictions? At what price?