Chapter 5
Overfitting and Optimism in Prediction Models

Background If we develop a statistical model with the main aim of outcome prediction, we are primarily interested in the validity of the predictions for new subjects, outside the sample under study. A key threat to validity is overfitting, i.e. that the data under study are well described, but that predictions are not valid for new subjects. Overfitting causes optimism about a model's performance in new subjects. After introducing overfitting and optimism, we illustrate overfitting with a simple example of comparisons of mortality figures by hospital. After appreciating the natural variability of outcomes within a single centre, we turn to comparisons across centres. We find that we would exaggerate any true patterns of differences between centres, if we would use the observed average outcomes per centre as predictions of mortality.

A solution is presented, which is generally named "shrinkage." Estimates per centre are drawn towards the average to improve the quality of predictions. We then turn to overfitting in regression models, and discuss the concepts of selection and estimation bias. Again, shrinkage is a solution, which now draws estimated regression coefficients to less extreme values. Bootstrap resampling is presented as a central technique to correct overfitting and quantify optimism in model performance.

5.1 Overfitting and Optimism

To derive a model, we use empirical data from a sample of subjects, drawn from a population (Fig. 5.1). The sample is considered to be drawn at random. The data from the sample are only of interest in that they represent an underlying population.[13,409] We use the empirical data to learn about patterns in the population, and to derive a model that can provide predictions for new subjects from this population. In learning from our data an important risk is that the data under study are well described, but that the predictions do not generalize to new subjects outside the sample. We may capitalize on specifics and idiosyncrasies of the sample. This is referred to as "overfitting." In statistics, overfitting is sometimes defined as fitting a statistical model that has too many parameters, or as the "curse of dimensionality."[181] For prediction models, we may define overfitting more precisely as fitting a statistical

E.W. Steyerberg, *Clinical Prediction Models*,
DOI:10.1007/978-0-387-77244-8_5,

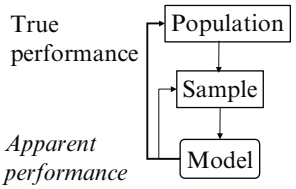

Fig. 5.1 Graphical illustration of optimism, which is defined as the difference between true performance and apparent performance. The apparent performance is determined on the sample where the model was derived from; true performance refers to the performance in the underlying population. The difference between apparent and true performance is defined as the optimism of a prediction model

model with too many degrees of freedom in the modelling process. Degrees of freedom are used by estimation of the coefficients in a regression model, but also by searching for the optimal model structure. The latter may include procedures to search for important predictors from a larger set of candidate predictors, optimal coding of predictors, and consideration of potential non-linear transformations.

Overfitting leads to a too optimistic impression of model performance that may be achieved in new subjects from the underlying population. Optimism is defined as true performance minus apparent performance, where true performance refers to the underlying population, and apparent performance refers to the estimated performance in the sample (Fig. 5.2). Put simply: "what you see may not be what you get."[23]

5.1.1 Example: Surgical Mortality in Oesophagectomy

Surgical resection of the oesophagus (oesophagectomy) may be performed for subjects with oesophageal cancer. It is among the surgical procedures that carry a substantial risk of 30-day mortality (see also Fig. 6.2).[125,213] Underlying differences in quality between hospitals may affect the 30-day mortality. A question is whether we can identify the better hospitals, and whether we can predict the mortality for a typical subject in a hospital.[260]

5.1.2 Variability within One Centre

We first illustrate the variability of mortality estimates within a single centre, according to different sample sizes. For oesophagectomy, we assume 10% as an average estimate of mortality among elderly patients, based on analyses of the

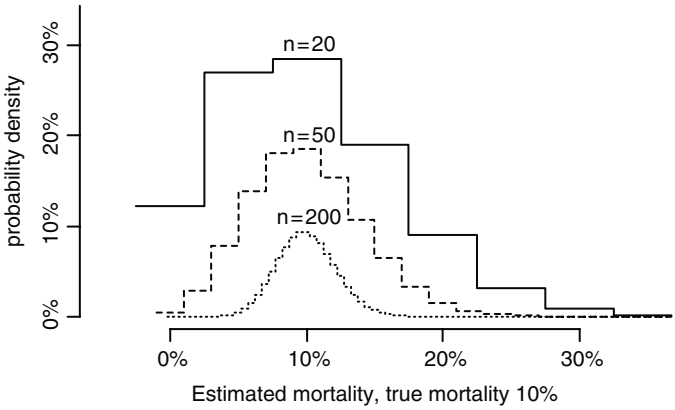

Fig. 5.2 Estimated mortality in relation to sample size. When the true mortality is 10% in samples of $n = 20$, around 30% of these will contain two deaths (estimated mortality, 10%). With larger sample sizes, observed mortalities are more likely close to 10%

SEER-Medicare registry data, where mortality exceeded 10%: 221 of 2,031 subjects had died within 30 days after surgery, or 10.9% [95% CI, 9.6%–12.3%].[423]

For illustration, we assume that case-mix is irrelevant, i.e. that all patients have the same true mortality risks. The observed mortality rate in a centre may then be assumed to follow a binomial distribution (Fig. 5.2). When the true mortality is 10% in samples of $n = 20$, around 30% of these will contain two deaths (estimated mortality, 10%). With larger sample sizes, observed mortalities are more likely close to 10%; e.g. when $n = 200$, mortality is estimated between 8% and 12% in 71% of the samples.

5.1.3 Variability between Centres: Noise vs. True Heterogeneity

We need to appreciate within centres variability when we want to make predictions of mortality by centre. For example, consider that 100 centres each reported mortality in 20 subjects, while the true mortality risk was 10% for every patient. On average two deaths are hence expected per centre (10% of 20). The expected distribution of the estimated mortality is as in Fig. 5.2: 12% of the centres will have 0% mortality, and 13% will report a 20% or higher mortality. An actual realization is shown in Fig. 5.3. A statistical test for differences between centres should be non-significant for most of such comparisons (for 95% of the cases when $p < 0.05$ is used as criterion for statistical significance).

Of more interest is the situation that the true mortality varies by centre. This can be simulated with a heterogeneity parameter, often refered to as τ (tau). Assuming a normal distribution for the differences across centres, we can write:

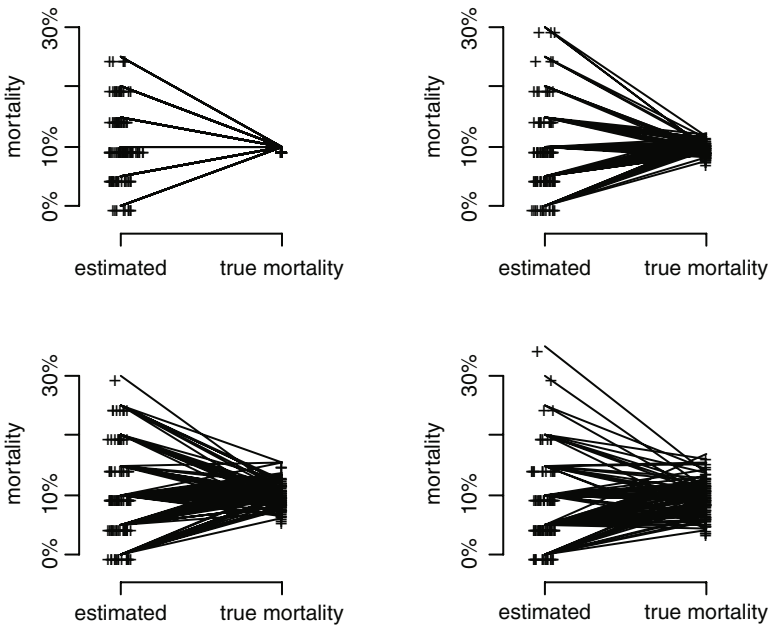

Fig. 5.3 Estimated and true mortality for 100 centres that analyzed 20 subjects each, while the average mortality was 10% for all (*upper left panel*), 10% ± 1% (*upper right panel*), 10% ± 2% (*lower left panel*), 10% ± 3% (*lower right panel*)

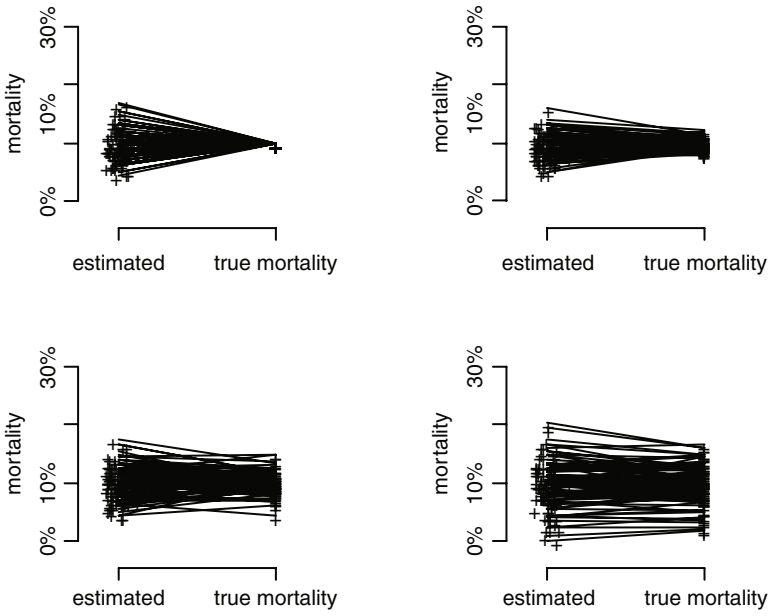

Fig. 5.4 Estimated and true mortality for 100 centres that had 200 subjects each, while the average was 10% for all (panel a), 10% ± 1% (panel b), 10% ± 2% (panel c), and 10% ± 3% (panel d)

true mortality ~ $N(10\%, \text{sd}= \tau)$. With $\tau = 1\%$, 95% of the centres have a mortality between 8% and 12%, while setting τ to 2% and 2.5% implies that 95% of the centres have a mortality between 6% and 14%, and between 5% and 15%, respectively. This underlying heterogeneity causes the estimated mortality to have more variability than expected from the binomial distribution with a single true mortality of 10%. This is recognized in the distributions of Fig. 5.3. Differences between centres can be tested, and will be identified as significant depending on the magnitude of the heterogeneity (τ), and the sample size (number of centres, sample size per centre).

5.1.4 Predicting Mortality by Centre: Shrinkage

We recognize that the estimated mortalities are too extreme as predictions compared with the distribution of the true mortalities (Fig. 5.3). Predictions other than 10% are by definition too extreme when there is no heterogeneity. Too extreme predictions also occur when there is underlying variability across centres (e.g. true mortality between 6 and 14%). Per centre, the estimated mortality is an unbiased estimator of the true mortality in each centre. But the overall distribution of estimated mortality suffers from the low numbers per centre, which makes that chance severely affects our predictions.

The phenomenon in Fig. 5.3 is an example of regression to the mean.[301] It is a motivation for shrinkage of predictions to the average, a principle that is also important in more complex regression models.[81,459] We should shrink the individual centre's estimates towards the overall mean to make better predictions overall.

We can also say that predictions tend to be overfitted: They point at very low and very high risk hospitals, while the truth will be more in the middle. The identification of extreme hospitals will be unreliable with small sample size. With larger sample size, e.g. 200 subjects per centre, the overfitting problem is reduced (Fig. 5.4). Empirical Bayes and random effects methods have been proposed to make better predictions (see Chap. 21).[22,458]

5.2 Overfitting in Regression Models

5.2.1 Model Uncertainty: Testimation

Overfitting is a major problem in regression modelling. It arises from two main issues: model uncertainty and parameter uncertainty (Table 5.1). Model uncertainty is caused by specification of the structure of our model, such as which characteristics are included as predictors, on information of the data set under study. The model structure is therefore uncertain. This model uncertainty is

Table 5.1 Causes and consequences of overfitting in prediction models

Issue	Characteristics
Causes of overfitting	
Model uncertainty	The structure of a model is not pre-defined, but determined by the data under study. Model uncertainty is an important cause of overfitting
Parameter uncertainty	The predictions from a model are too extreme because of uncertainty is the effects of each predictor (model parameters)
Consequences of overfitting	
Testimation bias	Overestimation of effects of predictors because of selection of effects that withstood a statistical test
Optimism	Decrease in model performance in new subjects compared with performance in the sample under study

usually ignored in statistical analyses, which falsely assume that the model was pre-specified.[69,101,194]

The result of model uncertainty is selection bias.[26,82,365,407] Note that selection bias here refers to the bias caused by selection of predictors from a larger set of predictors, in contrast to the selection of subjects from an underlying population in standard epidemiological texts. Suppose that we investigate 20 potential predictors for inclusion in a prognostic model. If these are all noise variables, the true regression coefficients are zero. On average one variable will be statistically significant at the $p<0.05$ level. The estimated effect will be relatively extreme, since otherwise the effect would not have been significant. If this one variable is included in the model, it will have a quite small or quite large effect (Fig. 5.5, left panel). On average the effect of such a noise variable is still zero.

If some of the 20 variables are true predictors, they will sometimes have a relatively small and sometimes a relatively large effect. If we only include a predictor when it has a relatively large effect in our model, we are overestimating the effect of such a predictor. This phenomenon is referred to as *testimation bias*: Because we test first, the effect estimate is biased.[26,69]

In the example of a predictor with true regression coefficient 1 and Standard Error (SE) 0.5, the effect will be statistically significant if estimated as lower than $-1.96 \times \text{SE} = -0.98$, or exceeding $+1.96 \times \text{SE} = +0.98$ (52% of the estimated coefficients, Fig. 5.5, right panel). The average of the estimated coefficients in these 52% cases is 1.39 rather than 1. Hence, a bias of +39% occurs. In formal terms, we can state: if b is significant, then $b=b$, else $b=0$. Instead of considering the whole distribution of predictor effects, we only consider a selected part.

Testimation bias is a pervasive problem in medical statistics and predictive modelling.[174] The bias is large for relatively weak effects, as is common in medical research. Selection bias is not relevant if we have a huge sample size, or consider

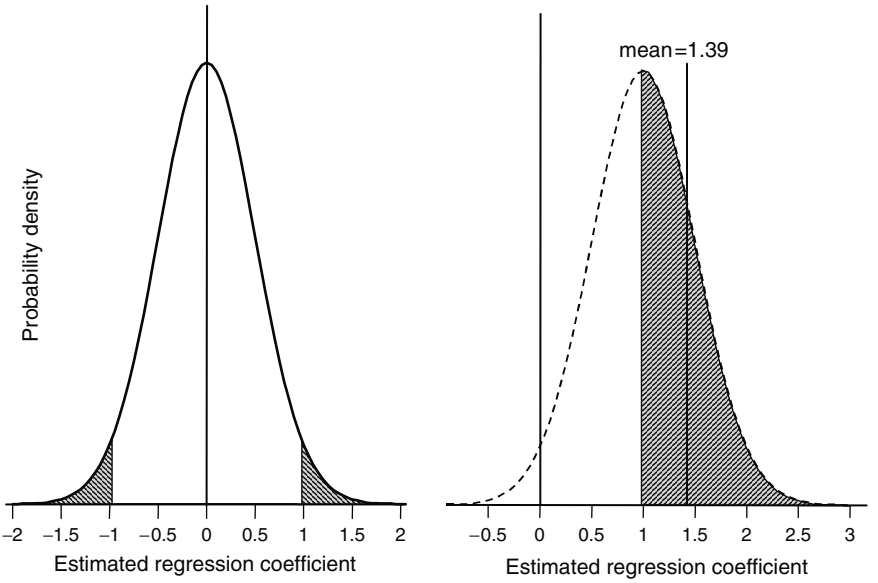

Fig. 5.5 Illustration of testimation bias. In case of a noise variable, the average of estimated regression coefficients is zero, and 2.5% of the coefficients is below − 0.98 (1.96 × SE of 0.5), and 2.5% of the coefficients is larger than +0.98 (1.96 × SE of 0.5). In case of a true coefficient of 1, the estimated coefficients are statistically significant in 52%. For these cases, the average of estimated coefficients is 1.39 instead of 1

predictors with underlying large effects, since these predictors will anyway be selected for a prediction model. Neither does selection bias occur if we pre-specify the prediction model ("full model").[174]

5.2.2 Other Biases

A well-known problem in prediction is bias by selection of an "optimal" cut-point for a continuous predictor.[12,117,355] A similar problem occurs if we examine different transformations for predictor variables as a check for linearity. For example, we may add a square term to a linear term, and omit the square term if not statistically significant.[148] More subtle variants occur when we less formally assess alternative model specifications. For example, we may consider different transformations of the outcome variable in a linear model, and visually judge the best transformation for use in further modelling. Or we examine different coding variants of a categorical predictor, with merging of groups with what we consider to have "similar outcomes." These issues are discussed in more detail in Chap. 9 and 10 on coding of predictors, and Chap. 11 and 12 on selection of predictors.

5.2.3 Overfitting by Parameter Uncertainty

It appears that even when the structure of our model is fully pre-specified, predictions are too extreme when multiple predictors are considered. This is because parameters, such as regression coefficients, are estimated in the model with uncertainty. This surprising finding has been the topic of much theoretical research.[81,459] An intuitive explanation is related to how we create a linear predictor in regression models. Hereto, the regression coefficients of multiple predictors are multiplied with the predictor values. With default estimation methods (e.g. least squares for linear regression and maximum likelihood for logistic or Cox regression), each of the coefficients is estimated in a (nearly) unbiased way. But each coefficient is associated with uncertainty, as reflected in the estimated standard error and 95% confidence interval (CI). This uncertainty tends us to overestimate predictions at the extremes of a linear predictor, i.e. low predictions will on average be too low, and high predictions will on average be too large. This is an example of regression to the mean. We can shrink coefficients towards zero to prevent this overfitting problem.[81,174,459]

This phenomenon is related to "Stein's paradox": biased estimates rather than unbiased estimates are preferable in multivariable situations to make better predictions.[107,398] Shrinkage introduces bias in the multivariable regression coefficients, but if we shrink properly the gain in precision of our predictions more than offsets the bias. The issue of bias–variance trade-off is central in prediction modelling,[181] and will be referred to throughout this book. Estimation with shrinkage methods is discussed in more detail in Chap. 13.

5.2.4 Optimism in Model Performance

Overfitting can visually be appreciated from the distributions of estimated mortality as in Figs. 5.3 and 5.4, but also from model performance measures. For example we may calculate Nagelkerke's R^2 for a logistic model that includes 20 centres (coded as a factor variable, with 19 dummy variables indicating the effect of 19 centres against a reference hospital). If the true mortality in all hospitals was 10%, the estimated R^2 was 9.4% when each hospital contained 20 subjects (Table 5.2). In fact, R^2 was 0%, since no true differences between centres were present. The estimated 9.4% is based on pure noise. We refer to the difference between 9.4% and 0% as the optimism in the apparent performance (Fig. 5.1). With larger sample sizes, the optimism decreases, e.g. to 0.1% for 20 centres with 2,000 subjects each (total 40,000 subjects, 4,000 deaths on average). Statistical testing of the between centre differences was by definition not significant in 95% of the simulations. We might require statistical significance of this overall test before trying to interpret between centre differences.

When true differences between centres were present (e.g. a range of 6–14% mortality, $\tau = 2\%$), the true R^2 was close to 1% ($n = 2{,}000$). With small sizes per centre, the estimated R^2 was 10.1%, which is again severely optimistic (Table 5.2).

A well-known presentation of optimism is to visualize the trade-off between model complexity and model performance.[181] We illustrate this trade-off in Fig. 5.6,

Table 5.2 R^2 for a logistic model predicting mortality in 20 centres. True mortality was 10% in the first series of simulations, and R^2 reflects pure noise. True mortality varied between 6% and 14% ($\tau = 2\%$) in the second series of simulations

True mortality	Sample size	R^2_{app}	R^2_{adj}	$R^2_{bootstrap}$
10%	$20 \times n = 20$	9.4	−0.1	NA
	$20 \times n = 200$	1.0	0	−0.5
	$20 \times n = 2{,}000$	0.1	0	0
10% ± 2%	$20 \times n = 20$	10.1	0.3	NA
	$20 \times n = 200$	1.9	0.9	0.3
	$20 \times n = 2{,}000$	1.0	0.9	0.8

Nagelkerke's R^2 calculated in logistic regression models,[309] averaged over 500 repetitions. R^2_{app}, R^2_{adj}, $R^2_{bootstrap}$ refer to the apparent, adjusted and bootstrap-corrected estimates of R^2. The R^2_{adj} included "LR – *df*" instead of "LR" in the formula. Note that not all coefficients could directly be estimated, since some hospitals had 0% estimated mortality with $n = 20$; for these we used 1% as the estimated mortality (adding one subject as dead, with a weight of $1\% \times 20 = 0.2$). Bootstrapping with these weighted samples was not readily possible.

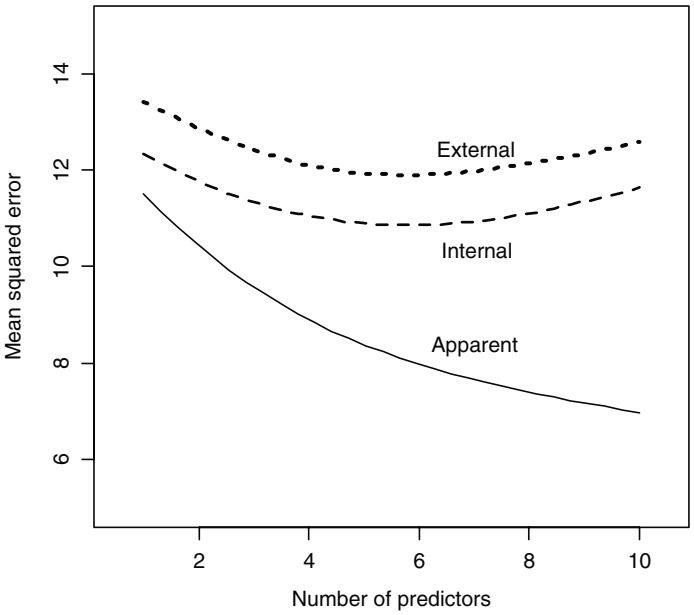

Fig. 5.6 Mean squared error of predictions from models with increasing complexity (1,000 simulated samples with $n = 50$). Apparent performance improves with more predictors, but internal and external performances worsen with more than five predictors

where we considered a simple linear regression model with 1 to 10 predictors. The model performance is evaluated by the mean squared error (mean $(y\text{-}\hat{y})^2$) for the underlying population (internal validation), and for a population where the true regression

coefficients were slightly different (external validation). With 50 subjects per sample for estimation of the model (1,000 simulations), we note that the apparent error decreases with more predictors considered. But the internal and external performances do not improve after approximately five predictors are included. Overfitting occurs after approximately five predictors, and optimism increases from modest for one predictor to substantial for models with ten predictors.

*5.2.5 Optimism-Corrected Performance

In linear regression analysis, an adjusted version of R^2 is available, which compensates for the degrees of freedom used in estimation of a model. Such an adjusted version can also be considered for Nagelkerke's R^2, which we consider e.g. for logistic and Cox models. We could subtract the degrees of freedom used to estimate the LR of the model in the calculation:

$$R^2_{\text{adjusted}} = (1 - e^{(-(LR-df)/n)})/(1 - e^{(-(-2LL0)/n)}),$$

where LR refers to the difference in −2 log likelihood (−2LL) of the model with and without the predictor, *df* are the degrees of freedom of the predictors in the model, *N* is the sample size, and LL0 is the log likelihood of the Null model (without predictors).

This adjusted version is not standard in most current software however. When we apply this formula for the simulated centre outcome as shown in Figs. 5.3 and 5.4, the average adjusted R^2 for noise differences is 0, with approximately half of the adjusted R^2 values being negative (Table 5.2). The adjustment made the R^2 estimates a bit conservative for small samples. For example, when true differences existed, the adjusted R^2 was 0.3% rather than 0.9% (Table 5.2).

A more general optimism correction is possible with bootstrapping, which is explained in the next section. In Table 5.2, bootstrap-corrected performance was more conservative than the adjusted R^2 formula, which may be caused by a not fully normal distribution of the optimism in R^2.[401]

5.3 Bootstrap Resampling

Bootstrapping alludes to a German legend about Baron Münchhausen, who was able to lift himself out of a swamp by pulling himself up by his own hair. In later versions of the legend he was using his own bootstraps to pull himself out of the sea, which gave rise to the term *bootstrapping*. A bootstrap was a loop of leather sewn onto the back of each boot to hold onto when pulling boots onto one's feet. In statistics, bootstrapping is a method for estimating the sampling distribution of an estimator by resampling with replacement from the original sample.[486]

Bootstrapping mimics the process of sampling from the underlying population. Since we only have a sample from the population, this sampling is not truly

Table 5.3 Illustration of five bootstrap samples drawn with replacement from five ages

Original sample	Bootstrap samples
20, 25, 30, 32, 35	20, 20, 30, 32, 35
	20, 25, 25, 30, 35
	20, 25, 30, 30, 32
	25, 32, 35, 35, 35
	30, 30, 32, 35, 35
	…

For easier interpretation, values were sorted per sample

possible, similar to the legend about Baron Münchhausen. Bootstrap samples are drawn with replacement from the original sample to introduce a random element. The bootstrap samples are of the same size as the original sample, which is important for the precision of estimates in each bootstrap sample.

For example, the GUSTO-I subsample 5 includes 429 subjects (Chap. 24). When we draw bootstrap samples, these each contain 429 subjects, but some subjects may not be included, others once, others twice, others three times, etc. On average, a subject has 63.2% chance of being at least once selected for a bootstrap sample.[108] For illustration we consider the simple case of the age of five subjects who are 20-, 25-, 30-, 32-, and 35-years old. Bootstrap samples might look like these in Table 5.3.

5.3.1 Applications of the Bootstrap

Bootstrapping is a widely applicable, non-parametric method. It can provide valuable insight in the empirical distribution of a summary measure from a sample. Bootstrap samples are repeatedly drawn from the data set under study, and each analyzed as if they were an original sample.[108]

For some measures, such as the mean of a population, we can use a statistical formula for the standard deviation (SD = sqrt[var] = sqrt($[x_i - \text{mean}(x))^2 / (n-1)]$). We can use the SD to calculate 95% CI as ± 1.96 × SE or ± 1.96 × SD/sqrt(n). The bootstrap can be used to calculate the SE for any measure. For the mean, the bootstrap will usually result in a similar SE and 95% CI estimates as obtained from the standard formula. For other quantities, such as the median, no SE or 95% CI can be calculated with standard formulas, but the bootstrap can. See Harrell for an extensive illustration.[174]

5.3.2 Bootstrapping for Regression Coefficients

The bootstrap can assist in estimating distributions of regression coefficients, such as standard errors and CIs. The bootstrap can be useful in estimating distributions of related measures such as the difference between an adjusted and an unadjusted regression coefficient.[472] In the latter case, two regression coefficients would be estimated in each bootstrap sample. The difference would be

calculated in each sample, and the distribution over bootstrap samples would be interpreted as the sampling distribution. CIs can subsequently be calculated with three methods:

1. Normal approximation: The mean and SE are estimated from the distribution (note: the SD over bootstraps is the SE of the mean).
2. Percentile method: Quantiles are simply read from the empirical distribution. For example, 95% CIs are based on the 2.5% and 97.5% percentile, e.g. the 50th and 1,950th bootstrap estimate out of 2,000 replications.
3. Bias-corrected percentile method: Bias in estimation of the distribution is accounted for, based on the difference between the median of the bootstrap estimates and the sample estimate ("BCa").[108]

For reliable estimation of distributions, large numbers of replications are advisable, e.g. at least 2,000 for method 2 and 3. Empirical *p* values can similarly be based on bootstrap distributions, e.g. by counting the number of estimates smaller than zero for a sample estimate larger than zero (giving a one-sided empirical *p* value).[108]

5.3.3 Bootstrapping for Optimism Correction

A very important application of bootstrapping is in quantifying the optimism of a prediction model.[69,108,174,459] With a simple bootstrap variant, one repeatedly fits a model in bootstrap samples, and evaluates the performance in the original sample (Fig. 5.7).

The average performance of the bootstrap models in the original sample can be used as the estimate of future performance in new subjects. A more accurate estimate is however obtained in a slightly more complicated way.[108] The bootstrap is used to estimate the optimism: The decrease between performance in the bootstrap sample (Sample* Fig. 5.7) and performance in the original sample. This optimism is subsequently subtracted from the original estimate to obtain an "optimism-corrected" performance estimate.[174]

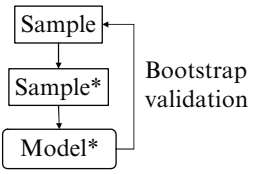

Fig. 5.7 Schematic representation of bootstrap validation for optimism correction of a prediction model. Sample* refers to the bootstrap sample that is drawn with replacement from the Sample (the original sample from an underlying population). Model* refers to the model constructed in Sample*

*5.3.4 Calculation of Optimism-Corrected Performance

Optimism-corrected performance is calculated as
Optimism-corrected performance = Apparent performance in sample − Optimism,

where

Optimism = Bootstrap performance − Test performance.
The exact steps are as follows:

1. Construct a model in the original sample; determine the apparent performance on the data from the sample used to construct the model;
2. Draw a bootstrap sample (Sample*) with replacement from the original sample (Sample, Fig. 5.7);
3. Construct a model (Model*) in Sample*, replaying every step that was done in the original Sample, especially model specification steps such as selection of predictors from a larger set of candidate predictors. Determine the bootstrap performance as the apparent performance of Model* on Sample*;
4. Apply Model* to the original Sample without any modification to determine the test performance;
5. Calculate the optimism as the difference between bootstrap performance and test performance;
6. Repeat steps 1–4 many times, at least 100, to obtain a stable estimate of the optimism;
7. Subtract the optimism estimate (step 5) from the apparent performance (step 1) to obtain the optimism-corrected performance estimate.

Note that the original sample is used for testing of Model*, while it contains largely the same subjects as the bootstrap sample (Sample*). Although this may seem invalid, both theoretical and empirical research supports this process. Alternative bootstrap validation procedures have been proposed.[1] Appealing variants are the .632 and .632+ methods, where the testing of the models from the bootstrap sample is on subjects from the original sample who were not included in the bootstrap sample.[109] On average, 63.2% of the subjects are selected in a bootstrap sample, giving the method its name. On average 36.8% of the subjects are left for testing of the model. These .632 and .632+ variants did however not have clear advantages over the bootstrap procedure described earlier in some empirical studies.[413,479]

We can apply the bootstrap approach to any performance measure, including the R^2, c statistic, and calibration measures such as calibration slope. A strong aspect of the bootstrap is that we can incorporate various complex steps from a modelling strategy. This is important since exact distributional results are virtually impossible

[1] The "simple bootstrap" compares the performance of the model from the original sample in bootstrap samples. This was less efficient than the procedure described here, where models from the bootstrap samples are tested in the original sample (see Efron).

to obtain, even for simple common selection algorithms.[336] The bootstrap can hence give insight in the relevance of model uncertainty, including both testimation bias and parameter uncertainty. In practice, however, it may be hard to fully validate a prediction model, including all steps made in the development of the model. For example, automated stepwise selection methods can be replayed in every bootstrap sample, leading to reasonably correct optimism-corrected performance estimates.[401] But more subtle modelling steps usually cannot fully be incorporated, such as choices on coding and categorization of predictors. The optimism-corrected estimate may then be an upper bound of what can be expected in future subjects. Only a fully specified modelling strategy can be replayed in every bootstrap sample.

It is often useful to calculate the optimism of a "full model," i.e. a prediction model, including all predictors without any fine-tuning such as deleting less-important predictors. The optimism estimate of such a full model may be a guide for further modelling decisions.[174] If the optimism is substantial, it is a warning that we should not base our model only on the data set at hand. Using external information may improve the future performance of the model.[164]

*5.3.5 Example: Stepwise Selection in 429 Patients

As an example, we consider a sample of 429 patients from the GUSTO-I study, which studied 30-day mortality in patients with acute myocardial infarction (details in Chap. 24). We first fitted a model with eight predictors, as specified in the TIMI-II study ("full model").[302] This model had a Nagelkerke R^2 of 23% as apparent performance estimate. In 200 bootstrap samples, the mean apparent performance was 25% (Table 5.4). When the models from each bootstrap sample were tested in the original sample, the R^2 decreased substantially (to 17%). The optimism hence was 25% − 17%=8%, and the optimism-corrected R^2, 23% − 8%=15%.

We can follow a backward stepwise selection procedure with $p<0.05$ for factors remaining in the model (Chap. 11). This leads to inclusion of only three predictors (age, hypotension, and shock). The apparent performance drops from 23% to 15% by excluding six of the eight predictors. The stepwise selection was repeated in every bootstrap sample, leading to an average apparent performance of 18%, which dropped to 12% when models were tested in the original sample (optimism, 6%; optimism-corrected R^2, 9%). When we falsely assume that the 3 predictor model was pre-specified, we would estimate the optimism as 3% rather than 6%. This discrepancy illustrates that optimism by selection bias was as important as the optimism due to parameter uncertainty in this example.

We note that the apparent performance in the bootstrap samples was higher than the apparent performance in the original sample (Table 5.4). This pattern is often noted in bootstrap model validation. It may be explained by the fact that some patients appear multiple times in the bootstrap sample. Hence, it is easier to predict the outcome, reflected in higher apparent performance. Further, we note that the

Table 5.4 Example of bootstrap validation of model performance, as indicated by Nagelkerke's R^2 in a subsample of the GUSTO-I data base (sample5, n=429)

Method	Apparent (%)	Bootstrap (%)	Test (%)	Optimism (%)	Optimism-corrected (%)
Full 8 predictor model	22.7	24.7	17.2	7.6	15.1
Stepwise, 3 predictors, p<0.05	17.6	18.7	12.7	5.9	11.7
Stepwise model falsely assumed to be pre-specified	17.6	18.2	15.4	2.9	14.7

optimism is smaller after model specification by stepwise selection (6% instead of 8%). However, the optimism-corrected performance of the stepwise model R^2 12% is clearly lower than the performance of the full 8 predictor model (R^2 15%). This pattern is often noted. A full model will especially perform better than a stepwise model when the stepwise selection eliminates several variables that are almost significant while they have some true predictive value. When a small set of dominant predictors is present, including only these would logically be sufficient. The bootstrap would show that these predictors are nearly always selected, and that other variables are most often excluded; the optimism would be relatively small and optimism-corrected performance similar to that of a full model. The leprosy case study is such an example (see Chap. 2). In the case that many noise variables are present in the full model, a selected submodel performs better than a full model. Careful pre-selection of candidate predictors is hence advisable, based on subject knowledge (literature, expert opinion), to prevent that pure noise variable are considered in the modelling process.

5.4 Cost of Data Analysis

The development of a prediction model for outcome prediction is a constant struggle in weighing better fit to the data against generalizability outside the sample. The more we incorporate from a specific data set in a model, the less the model may generalize.[101] This has aptly been labelled the "cost of data analysis." On the other hand, we do not want to miss important properties of the data, such as a clearly non-linear relationship of a predictor to the outcome. A prediction model where underlying model assumptions are fulfilled will provide better predictions than a model where assumptions are violated. Therefore, it is natural to assess such assumptions as linearity of continuous predictor effects and additivity of effects (Chap. 12). However, if we test all assumptions of a model and iteratively adapt the model to capture even small violations, the model will be very specific for the data analyzed.

*5.4.1 Example: Cost of Data Analysis in a Tree Model

An interesting concept was proposed by Ye, who determined the "generalized degrees of freedom" (GDF) of a model selection and estimation procedure.[494] The GDF indicate the overfitting that was associated with a modelling strategy. For example, Ye showed that a stepwise selection strategy that selected a model with five predictors (apparent $df = 5$) had a GDF of 14.1. A regression tree had 19 nodes (apparent $df = 19$), but GDF of 76.[494]

An essential part of Ye's method is to determine the apparent performance of a model when developed with pure noise. In Table 5.2, we note that the optimism in R^2 in the pure noise simulations was indeed very similar to the optimism as determined with an adjusted R^2 or with bootstrapping when some true effects were present. For example, for $n = 200$, the optimism was 1% with pure noise or with true effects.

5.4.2 Practical Implications

In the development of prediction models, we have to be aware of the cost of all data analysis steps. The appropriateness of a modelling strategy is indicated by the generalizability of results to outcome prediction for new patients. Some practical issues are relevant in this respect.

- Sample size: With a small sample size we have to be prepared to make more assumptions about our data; the power to detect deviations from assumptions will anyway be small. If deviations from assumptions are detected, and the model is adapted, testimation bias will occur and the validity of predictions for new patients may not necessarily be improved (Chap. 13);
- Robust strategies: Some modelling strategies are more "data hungry" than other strategies. For example, fitting a pre-specified logistic regression model with age and sex uses only two degrees of freedom. If we test for linearity of the age effect, and interactions between age and sex, we spend more degrees of freedom. If we use a method such as regression tree analysis, we search for cut-points of age, and model interactions by default, making the method more data-hungry than logistic regression (Chap. 4). Similarly, stepwise selection asks more of the data than fitting a pre-specified model. Not only do we want to obtain estimates of coefficients, we also want to determine which variables to include as predictors (Chap. 11);
- Bootstrap validation: The bootstrap can assist in determining an appropriate level of fine-tuning of a model to the data under study. However, when many alternative modelling strategies are considered, the bootstrap results may become less reliable in determining the optimal strategy, since the optimum may again be very specific for the data under study. The bootstrap works best to determine optimism for a single, pre-defined strategy.

5.5 Concluding Remarks

In science in general, and in prediction modelling specifically, we need to seek a balance between curiosity and skepticism. On the one hand, we want to make discoveries and advance our knowledge, but on the other hand we must subject any discovery to stringent tests, such as validation, to make sure that chance has not fooled us.[23] It has been demonstrated that our scientific discoveries are often false, especially if we search hard and explore a priori unlikely hypotheses.[210] Overfitting and the resulting optimism are important concerns in prediction models.

Questions

5.1 Overfitting and optimism
- (a) What is overfitting and why is it a problem?
- (b) What are the two main causes of overfitting? What is the difference and give some examples?

5.2 Shrinkage for prediction (Figs. 5.3 and 5.4)

A solution against the consequence of overfitting is shrinkage. For example, estimates per centre can be drawn towards the average to improve the quality of predictions in Figs. 5.3 and 5.4.
- (a) Is the required shrinkage more, or less, in Fig. 5.4 compared with Fig. 5.3?
- (b) Is the underlying true heterogeneity more, or less, in Fig. 5.4 compared with Fig. 5.3?

5.3 Bootstrapping (Sect. 5.3)
- (a) How can a bootstrap sample be created? How is this done with the `sample` command in R?
- (b) How can the test sample for the .632 bootstrap variant by selected in *R*?
- (c) How can bootstrapping be used to derive optimism-corrected estimates of model performance, addressing the two main causes of overfitting?