Chapter 17
Sensitivity Analysis and Model Validation

Justin D. Salciccioli, Yves Crutain, Matthieu Komorowski and Dominic C. Marshall

Learning Objectives

- Appreciate that all models possess inherent limitations for generalizability.
- Understand the assumptions for making causal inferences from available data.
- Check model fit and performance.

17.1 Introduction

Imagine that you have now finished the primary analyses of your current research and have been able to reject the null hypothesis. Even after your chosen methods have been applied and robust models generated, some doubts may remain. "*How confident are you in the results? How much will the results change if your basic data is slightly wrong? Will that have a minor impact on your results? Or will it give a completely different outcome?*" Causal inference is often limited by the assumptions made in study design and analysis and this is particularly pronounced when working with observational health data. An important approach for any investigator is to avoid relying on any single analytical approach to assess the hypothesis and as such, a critical next step is to test the assumptions made in the analysis.

Sensitivity Analysis and Model Validation are linked in that they are both attempts to assess the appropriateness of a particular model specification and to appreciate the strength of the conclusions being drawn from such a model. Whereas model validation is useful for assessing the model fit within a specific research dataset, sensitivity analysis is particularly useful in gaining confidence in the results of the primary analysis and is important in situations where a model is likely to be used in a future research investigation or in clinical practice. Herein, we discuss

MIT Critical Data, *Secondary Analysis of Electronic Health Records*,
DOI 10.1007/978-3-319-43742-2_17

concepts relating to the assessment of model fit and outline broadly the steps relating to cross and external validation with direct application to the arterial line project. We will discuss briefly a few of the common reasons why models fail validity testing and the potential implications of such failure.

17.2 Part 1—Theoretical Concepts

17.2.1 Bias and Variance

In statistics and machine learning, the bias–variance trade-off (or dilemma) is the problem of simultaneously minimizing two sources of error that prevent supervised learning algorithms from generalizing beyond their training set. A model with high bias fails to accurately estimate the data. For example, a linear regression model would have high bias when trying to model a quadratic relationship—no matter how the parameters are set (as shown in Fig. 17.1). Variance, on the other hand, relates to the stability of your model in response to new training examples. An algorithm that fits the training data very well but generalizes poorly to new examples (showing over-fitting) is said to have high variance.

Some common strategies for dealing with bias and variance are outlined below.

- High bias:
  - Adding features (predictors) tends to decrease bias, at the expense of introducing additional variance.
  - Adding training examples will not fix high bias, because the underlying model will still not be able to approximate the correct function.
- High variance:
  - Reducing model complexity can help decrease variance. Dimensionality reduction and feature selection are two examples of methods to decrease model parameters and thus reduce variance (parameter selection is discussed below).
  - A larger training set tends to decrease variance.

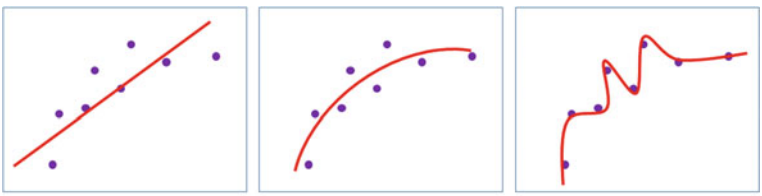

Fig. 17.1 Comparison between bias and variance in model development

17.2.2 Common Evaluation Tools

A variety of statistical techniques exist to quantitatively assess the performance of statistical models. These techniques are important, but generally beyond the scope of this textbook. We will, however, briefly mention two of the most common techniques: the R^2 value used for regressions and the Receiver Operating Characteristic (ROC) curve used for binary classifier (dichotomous outcome).

The R^2 value is a summary statistic representing the proportion of total variance in the outcome variable that is captured by the model. The R^2 has a range from 0 to 1 where values close to 0 reflect situations where the model does not appreciably summarise variation in the outcome of interest and values close to 1 indicate that the model captures nearly all of the variation in the outcome of interest. High R^2 values means that a high proportion of the variance is explained by the regression model. In R programming, the R^2 is computed when the linear regression function is used. For an example of R-code to produce the R^2 value please refer to the "R^2" function.

The R^2 value is an overall measure of strength of association between the model and the outcome and does not reflect the contribution of any single independent predictor variable. Further, while we may expect intuitively that there is a proportional relationship between the number of predictor variables and the overall model R^2, in practice, adding predictors does not necessarily increase R^2 in new data. It is possible for an individual predictor to decrease the R^2 depending on how this variable interacts with the other parameters in the model.

For the purpose of this discussion we expect the reader to be familiar with the computation and utility of the values of sensitivity and specificity. In situations such as developing a new diagnostic test, investigators may define a single threshold value to classify a test result as positive. When dealing with a dichotomous outcome, the Receiver Operating Characteristic (ROC) curve is a more complete description of a model's ability to classify outcomes. The ROC curve is a common method to show the relationship between the sensitivity of a classification model and its false positive rate (1 - specificity). The resultant Area Under the Curve of the ROC reflects the prediction estimate of the model, can take values from 0.5 to 1 with values of 0.5 implying near random chance in outcomes and values nearer to 1 reflecting greater prediction. For an example of ROC curves in R, please refer to the "ROC" function in the accompanying code. For further reading on the ROC curve, see for example the article by Fawcett [1] (Fig. 17.2).

17.2.3 Sensitivity Analysis

Sensitivity analysis involves a series of methods to quantify how the uncertainty in the output of a model is related to the uncertainty in its inputs. In other words, sensitivity analysis assesses how "sensitive" the model is to fluctuations in the parameters and data on which it is built. The results of sensitivity analysis can have

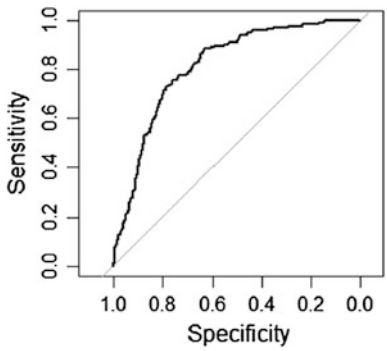

Fig. 17.2 Example of receiver operator characteristic (ROC) curve which may be used to assess the ability of a model to discriminate between dichotomous outcomes

important implications at many stages of the modeling process, including for identifying errors in the model itself, informing the calibration of model parameters, and exploring more broadly the relationship between the inputs and outputs of the model.

The principles of a sensitivity analysis are: (a) to allow the investigator to quantify the uncertainty in a model, (b) to test the model of interest using a secondary experimental design, and (c) using the results of the secondary experimental design to calculate the overall sensitivity of the model of interest. The justification for sensitivity analysis is that a model will always perform better (i.e. over-perform) when tested on the dataset from which it was derived. Sub-group analysis is a common variation of sensitivity analysis [2].

17.2.4 Validation

As discussed in Chap. 16—Data Analysis validation is used to confirm that the model of interest will perform similarly under modified testing conditions. As such, it is the primary responsibility of the investigator to assess the suitability of model fit to the data. This may be accomplished with a variety of methodological approaches and for a more detailed discussion of model fit diagnostics the reader is referred to other sources [3]. Although it is beyond the scope of this textbook to discuss validation in detail, the general theory is to select a model based on two principles: model parsimony and clinical relevance. A number of pre-defined model selection algorithm-based approaches including Forward selection, Backward, and Stepwise selection, but also lasso and genetic algorithms, available in common statistical packages. Please refer to Chap. 16 for further information about model selection.

Cross validation is a technique used to assess the predictive ability of a regression model. The approach has been discussed in detail previously [4]. The concept of cross-validation relies on the principle that a large enough dataset can

split into two or more (not necessarily equally sized) sub-groups, the first being used to derive the model and the additional data set(s) reserved for model testing and validation. To avoid losing information by training the model only on a subset of available data, a variant called k-fold cross validation exist (not discussed here).

External validation is defined as testing the model on a sample of subjects taken from a population different than the original cohort. External validation is usually a more robust approach for testing the derived model in that the maximum amount of information has been used from the initial dataset to derive a model and an entirely independent dataset is used subsequently to verify the suitability of the model of interest. Although external validation is the most rigorous and an essential validation method, finding a suitably similar albeit entirely independent cohort for external validation is challenging and is often unavailable for researchers. However, with the increasing amount of healthcare data being captured electronically it is likely that researchers will also have increasing capacity for external validation.

17.3 Case Study: Examples of Validation and Sensitivity Analysis

This case study used the dataset produced for the "IAC study", which evaluated the impact of inserting an arterial line in intensive care patients with respiratory failure. Three different sensitivity analyses were performed:

1. Test the effects of varying the inclusion criteria of time to mechanical ventilation and mortality;
2. Test the effects of changes in caliper level for propensity matching on association between arterial catheter insertion and the mortality;
3. Hosmer-Lemeshow Goodness-of-Fit test to assess the overall fit of the data to the model of interest.

A number of R packages from CRAN, were used to conduct these analyses: Multivariate and Propensity Score Matching [5], analysis of complex survey samples [6], ggplot2 for generating graphics [7], pROC for ROC curves [8] and Twang for weighting and analyzing non-equivalent groups [9].

17.3.1 Analysis 1: Varying the Inclusion Criteria of Time to Mechanical Ventilation

The first sensitivity analysis evaluates the effect of varying the inclusion criteria of time to mechanical ventilation and mortality. Mechanical ventilation is one of the more common invasive interventions performed in the ICU and the timing of intervention may serve as a surrogate for the severity of critical illness, as we might

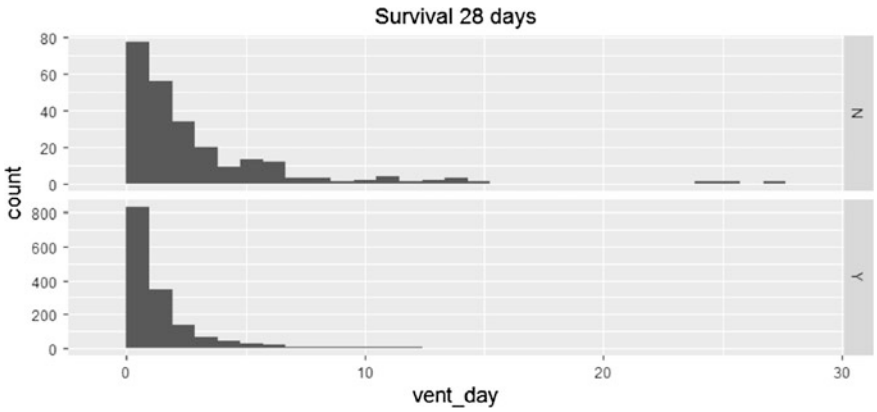

Fig. 17.3 Simple sensitivity analysis to compare outcomes between groups by varying the inclusion criteria. Modification of the inclusion criteria for subjects entered into the model is a common sensitivity analysis

expect patients with worse illness to require assisted ventilation earlier in the course of intensive care. As such, mechanical ventilation along with indwelling arterial catheter (IAC), another invasive intervention, may both be related to the outcome of interest, 28-day mortality. An example of R-code to inspect the distribution across groups of patients by ventilation status is provided in the "Cohort" function, in the accompanying R functions document (Fig. 17.3).

By modifying the time of first assisted mechanical ventilation we may also obtain important information about the effect of the primary exposure on the outcome. An example of R-code for this analysis is provided in the "Ventilation" function.

17.3.2 Analysis 2: Changing the Caliper Level for Propensity Matching

The second sensitivity analysis performed tests the impact of different caliper levels for propensity matching on the association between arterial catheter and the mortality. In this study, the propensity score matches a subject who did not received an arterial catheter with a subject who did. The matching algorithm creates a pair of two independent subjects whose propensity scores are the most similar. However, the investigator is responsible for setting a maximum reasonable difference in propensity score which would allow the matching algorithm to generate a suitable match; this maximum reasonable difference is also known as the propensity score 'caliper'. The choice of caliper for the propensity score match will directly influence the variance bias trade-off such that a wider caliper will result in matching of subjects which are more dissimilar with respect to likelihood of treatment. An

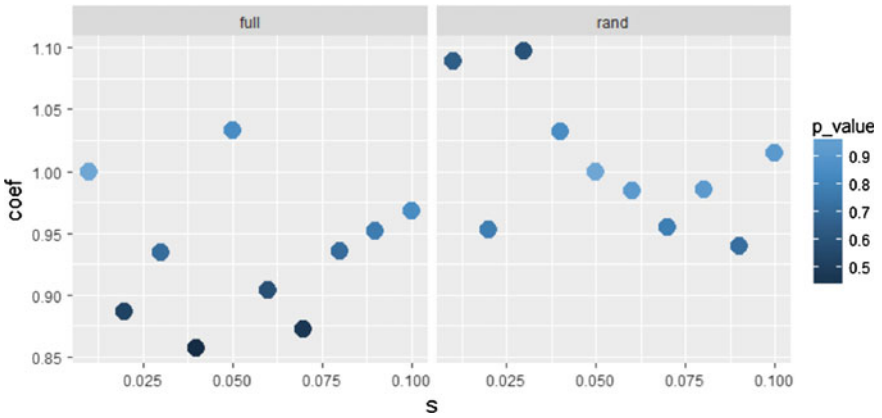

Fig. 17.4 A sensitivity analysis to assess the effect of modifying the propensity score caliper level

example of the R-code to produce a sensitivity analysis for varying the propensity score caliper level is provided in the accompanying R functions document as the "Caliper" function.

The Fig. 17.4 displays the effect of adjustments of the caliper level on the propensity score. The full model shows a lower coefficient due to the presence of additional variables.

17.3.3 Analysis 3: Hosmer-Lemeshow Test

The Hosmer-Lemeshow Goodness-of-Fit test may be used to assess the overall fit of the data to the model of interest [10]. For this test, the subjects are grouped according to a percentile of risk (usually deciles). A Pearson Chi square statistic is generated to compare observed subject grouping with the expected risk according to the model. An example of the R-code to conduct this test is provided in the accompanying R functions document as the "HL" function.

17.3.4 Implications for a 'Failing' Model

In the favorable situation of a robust model, each sensitivity analysis and validation technique supports the model as an appropriate summary of the data. However, in some situations, the chosen validation method or sensitivity analysis reveals an inadequate fit of the model for the data such that the model fails to accurately predict the outcome of interest. A 'failing' model may be the result of a number of different factors. Occasionally, it is possible to modify the model derivation

procedure in order to claim a better fit on the data. In the situations where modifying the model does not allow to achieve an acceptable level of error, however, it is good practice to renounce the investigation and re-start with an assessment of the a priori assumptions, in an attempt to develop a different model.

17.4 Conclusion

The analysis of observational health data carries the inherent limitation of unmeasured confounding. After model development and primary analysis, an important step is to confirm a model's performance with a series of confirmatory tests to verify a valid model. While validation may be used to check that the model is an appropriate fit for the data and is likely to perform similarly in other cohorts, sensitivity analysis may be used to interrogate inherent assumptions of the primary analysis. When performed adequately these additional steps help improve the robustness of the overall analysis and aid the investigator in making meaningful inferences from observational health data.

Take Home Messages

1. Validation and sensitivity analyses test the robustness of the model assumptions and are a key step in the modeling process;
2. The key principle of these analyses is to vary the model assumptions and observe how the model responds;
3. Failing the validation and sensitivity analyses might require the researcher to start with a new model.

Code Appendix

The code used in this chapter is available in the GitHub repository for this book: https://github.com/MIT-LCP/critical-data-book. Further information on the code is available from this website.

References

1. Fawcett T (2006) An introduction to ROC analysis. Pattern Recogn Lett 27(8):861–874
2. Brookes ST, Whitely E, Egger M, Smith GD, Mulheran PA, Peters TJ (2004) Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. J Clin Epidemiol 57(3):229–236
3. Pregibon D (1981) Logistic regression diagnostics. Ann Stat 9(4):705–724
4. Picard RR, Cook RD (1984) Cross-validation of regression models. J Am Stat Assoc 79 (387):575–583
5. Sekhon JS (2011) Multivariate and propensity score matching software with automated balance optimization: the matching package for R. J Stat Softw 42(i07)
6. Lumley T (2004) Analysis of complex survey samples. J Stat Softw 09(i08)
7. Wickham H (2009) ggplot2. Springer, New York
8. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M (2011) pROC: an open-source package for R and S + to analyze and compare ROC curves. BMC Bioinf 12:77
9. Ridgeway G, Mccaffrey D, Morral A, Burgette L, Griffin BA (2006) Twang: toolkit for weighting and analysis of nonequivalent groups. R package version 1.4-9.3. In: R Foundation for Statistical Computing, 2006. (http://www.cran.r-project.org). Accessed 2015
10. Hosmer DW, Lemesbow S (1980) Goodness of fit tests for the multiple logistic regression model. Commun Stat Theory Methods 9(10):1043–1069