MODEL PERFORMANCE ANALYSIS AND MODEL VALIDATION IN LOGISTIC REGRESSION

R. Arboretti Giancristofaro, L. Salmaso

1. INTRODUCTION

Regression models are powerful tools frequently used to predict a dependent variable from a set of predictors. They are widely used in a number of different contexts. An important problem is whether results of the regression analysis on the sample can be extended to the population the sample has been chosen from. If this happens, then we say that the model *has a good fit* and we refer to this question as a *goodness-of-fit analysis*, *performance analysis* or *model validation analysis* for the model (Hosmer and Lemeshow, 2000; D'Agostino *et al.*, 1998; Harrell *et al.*, 1996; Stevens, 1996). Application of modelling techniques without subsequent performance analysis of the obtained models can result in poorly fitting results that inaccurately predict outcomes on new subjects.

We deal with how to measure the quality of the fit of a given model and how to evaluate its performance in order to avoid poorly fitted models, i.e. models which inadequately describe the above mentioned relationship in the population.

First we state an important preliminary assumption and the aim of our work, and we introduce the concept of goodness-of-fit and the principle of optimism. Then we illustrate a brief review of the diverse techniques of model validation. Next, we define a number of properties for a model to be considered "good," and a number of quantitative performance measures. Lastly, we describe a methodology for the assessment of the performance of a given model.

2. GOODNESS-OF-FIT ASSESSMENT

We begin our discussion of methods for assessing the fit of an estimated logistic regression model, with the assumption that we are at least preliminarily satisfied with our efforts at the model building stage. By this we mean that, to the best of our knowledge, the model contains those variables (main effects as well as interactions) that should be in the model and that those variables have entered in the correct functional form. That is to say, we are not concerned with any model building issue nor are we comparing competing models - each model is taken as it is and its performance is evaluated for validation purposes only. Furthermore, we are not investigating how possible changes of the set of covariates included in the model may result in larger overall significance or higher performance.

We illustrate our approach with reference to logistic regression and we apply it to evaluate and validate logistic regression equations. Anyway, because our approach is not concerned with the development of the model but treats it as a "black box", it can be applied to any statistical model that generates predicted probabilities such as, for example, classification trees and neural networks.

As we have already mentioned, our aim is to evaluate and measure how effectively the model we have describes the outcome variable both in the sample and in the population. Before dealing with methods for assessing the measure of model performance, we want to stress the reason why this phase is so important in any statistical analysis.

We have seen that after deciding which covariates to include in the logistic model, coefficients are estimated in order to maximize the overall likelihood of the sample. The maximum likelihood technique used for this estimation is a mathematical maximization procedure, and hence there is a great opportunity for capitalization on chance, especially if the number of subjects is not large compared to the number of variables (Stevens, 1996). As we shall see later, the resulting regression equation often outperforms on the sample used to fit the model, but it does not generalize to other samples from the same population. That is to say, the logistic regression equation computes the best possible event predictions on the sample used to fit the model, but its predictive power drops away when applied to an independent sample from the same population (i.e. to similar subjects not included in the original sample).

If this is the case, the model will very likely work better for the data used to fit it than for any other data. Analysts usually refer to this fact as the *principle of optimism* (Picard and Cook, 1984). As a consequence, when assessing the goodness-of-fit of a regression equation, it is not enough to evaluate how well it predicts on the sample used to fit the model: it is of crucial importance for the researcher to obtain some quantitative and objective measures of how well the regression equation will predict on an independent sample of similar data in order to determine its degree of generalizability.

If the predictive power drops off sharply, then the results of the analysis (i.e. the predictive equation) are sample-specific and have no generalizability, thus being of limited scientific utility.

At a first glance, one could think of assessing the goodness-of-fit of the model by fitting the model to the sample and then using the model to predict the event probability for each subject in the sample. Any measure of the distance between the observed values of the response variable and the predicted event probabilities would then be a good indicator of the performance of the model.

The idea of the comparison is good - the distance measure would be a simple and intuitive indicator of the reliability of the prediction. Of course, we would like the predicted probabilities to accurately reflect the distribution of the positive events in the sample. The problem here is that we are using the same sample twice - to fit the model and to evaluate its performance. As we have already stated, each model is mathematically optimised to best fit the data on which it is built. Thus, any performance indicator measured on the same sample used to fit the model is biased in favour of the model, i.e. indicates the highest possible performance. On the contrary, we are interested in evaluating how well the model can predict the probability of the positive event for subjects not included in the original sample on which the model was built.

As a consequence, the use of independent data to fit and test the model is preferable when we are interested in demonstrating the generalizability of the model in order to use it to predict outcomes for future subjects (Hosmer and Lemeshow, 2000; Harrell *et al.*, 1996; Stevens, 1996). This type of assessment is often called *model validation*. Here follows a brief review of the main possible approaches.

In some situations it may be possible to obtain a new sample of data from the same population or from a similar population. This new sample can then be used to assess the goodness-of-fit of a previously developed model by applying the model as it is to the new sample. This type of assessment is called *external validation*; as stated by Harrell *et al.,* (1996). External validation is the most stringent and unbiased test for the model and for the entire data collection process.

Most of the times it is not possible to obtain a new independent external sample from the same population or a similar one. It may then be possible to internally validate the model. The most accredited methods for obtaining a good *internal validation* of a model's performance are *data-splitting*, *repeated data-splitting*, *jack-knife technique* and *bootstrapping*. The core concept of these methods is similar in order to exclude a sub sample of observations, develop a model based on the remaining subjects, and then test the model in the originally excluded subjects.

The methods differ from one another in the way they identify the sub-sample used to build the model.

i) Data-splitting

A random portion of the sample, usually between two thirds and three quarters, is used for model development. The obtained model is then "frozen" and applied as it is to the remaining portion of the sample for assessing the model's performance (Harrell *et al.,* 1996; Picard and Berk, 1990).

ii) Repeated data-splitting

This consists in repeating the previous analysis many times. Each iteration of the analysis splits the original sample in a different way. Such a repeated analysis is far more accurate than a simple data-splitting (Harrell *et al.,* 1996).

iii) Jack-knife technique

The jack-knife technique is very similar to data-splitting. The only difference is in the size of the two sub samples. Here, the model is fitted on all but one of the subjects in the original sample and is then tested on the set-aside case (Mosteller and Tukey, 1977; Stone, 1974). This procedure can be repeated as many times as the number of subjects in the original sample.

iv) Bootstrapping

Bootstrapping is an alternative method of internal validation which involves taking a large number of simple random samples *with replacement* from the original sample (Harrell *et al.*, 1996) and fitting the model on each of these samples.

The sampling procedure is the main difference between bootstrapping and data-splitting – bootstrapping samples are taken *with* replacement while data-splitting samples are taken *without* replacement. As a consequence, a duplication of the subjects in the original sample is possible.

The replication allows us to obtain large samples even from small original samples. However, if the original sample is bad, the samples obtained using these techniques will be bad as well. What is more, the interpretation of bootstrapping results is controversial.

A detailed review of this technique can be found in Efron and Tibshirani (1993).

After an in-depth analysis of the abovementioned techniques, we have opted for *repeated data-splitting*. Although bootstrapping appears to be a very powerful technique, we choose not to use it in order to obtain easily interpretable results. We prefer data-splitting to the jack-knife technique because the former provides us with more data for model testing purposes. If the model is tested on a single subject (as in the jack-knife technique) it is not possible to assess one of the most important dimensions of model's performance, i.e. calibration, which we shall define in one of the following paragraphs. We repeat the data-splitting to obtain more accurate results.

At this point of our analysis, we remind the reader that our aim is to assess the goodness-of-fit of a given model, and to determine whether the model can be used to predict the outcome of a subject not included in the original sample. Given this aim, we would need to have some specific ideas about what it means to say that a model fits.

In the following paragraph firstly we distinguish between fitting and validation samples, then we introduce the concepts of calibration and discrimination as the two main dimensions of model performance. Next we describe three measures for evaluating model calibration and discrimination and finally, we develop a procedure to evaluate the overall performance and validate a number of specific logistic regression models.

3. FITTING AND VALIDATION SAMPLES

In accordance with the principles of data-splitting we distinguish between fitting and validation samples - they are samples from the same population, but are distinct and independent from one another. Furthermore, they constitute a total partition of the original sample - they are from the same population but, at the same time, are distinct and independent from one other.

Firstly we use the fitting sample to fit the model. Then we take the fitted model as it is, apply it to the validation sample, and evaluate the model's performance on it.

The sizes of the two sub-samples must be chosen in such a way as to have enough data in the fitting sample to fit the model and enough data in the validation sample to validate the model. This is a delicate point, which we shall discuss later.

Since the fitted model performs in an optimistic manner on the fitting sample, we are to expect a lower performance of the model on the validation sample.

Our focus is to measure the predictive performance of a model, i.e. its ability to accurately predict the outcome variable on new subjects. We are interested in quantifying the closeness of the model's probability estimates to the correct outcome values. For the fitting sample we are interested in evaluating how well the model fits the data on which it was developed. For the validation sample we evaluate how valid the model is as it attempts to predict an independent sample. Interest in the latter relates in particular to how much the performance ability decreases or degenerates from what was quantified in the fitting set. If the decrease is small, we can conclude that the model can be used to predict the outcome of subjects which are not in the original sample.

We now introduce the concepts of discrimination and calibration which are the two main dimensions of the model's performance.

3.1 *Discrimination*

Discrimination refers to the "ability of the model to distinguish correctly the two classes of outcomes" (D'Agostino *et al.,* 1998); in other words, it is a measure of the ability of the model to "separate subjects with different responses" (Harrell *et al.,* 1996). Let us consider a logistic regression model fitted on a given sample.

According to the observed values of the outcome variable, the sample can be divided into two sub-groups. The first, which we call *positive*, contains all the subjects with a positive observed outcome and the second, which we call *negative*, contains all the subjects with a negative observed outcome.

Using the model's regression equation, we can compute the predicted probabilities of the positive event for the subjects in both the sub-groups. Subsequently, we can plot the distributions of these predicted probabilities for each of the two sub-groups. Perfect discrimination would result in two non-overlapping distributions of predicted probabilities for the two sub-groups (D'Agostino *et al.,* 98). The idea behind this is intuitive. If the two curves do not overlap, then the predicted event probability allows us to distinguish between positive and negative subjects.

Figure 1 intuitively illustrates this point. Given the regression equation, we can select an adequate cut-off point and declare that a subject is positive if its predicted event probability is larger than the cut-off point. If the two curves do not overlap, the chance of a wrong classification is small. If this is the case, we say that the model discriminates between the positive and the negative subjects in the sample. Note that discrimination is not concerned with the probability values.

Given the estimated event probability for a subject, we can select a probability value, say p^*, as a cut-off point to decide whether the subject is declared to be more likely to have a positive or a negative outcome.

All subjects with predicted probabilities equal to or greater than p* will be declared as positive, the others as negative. Once this is done, a two by two table, such as the following, can be generated to assess the performance of this decision rule:

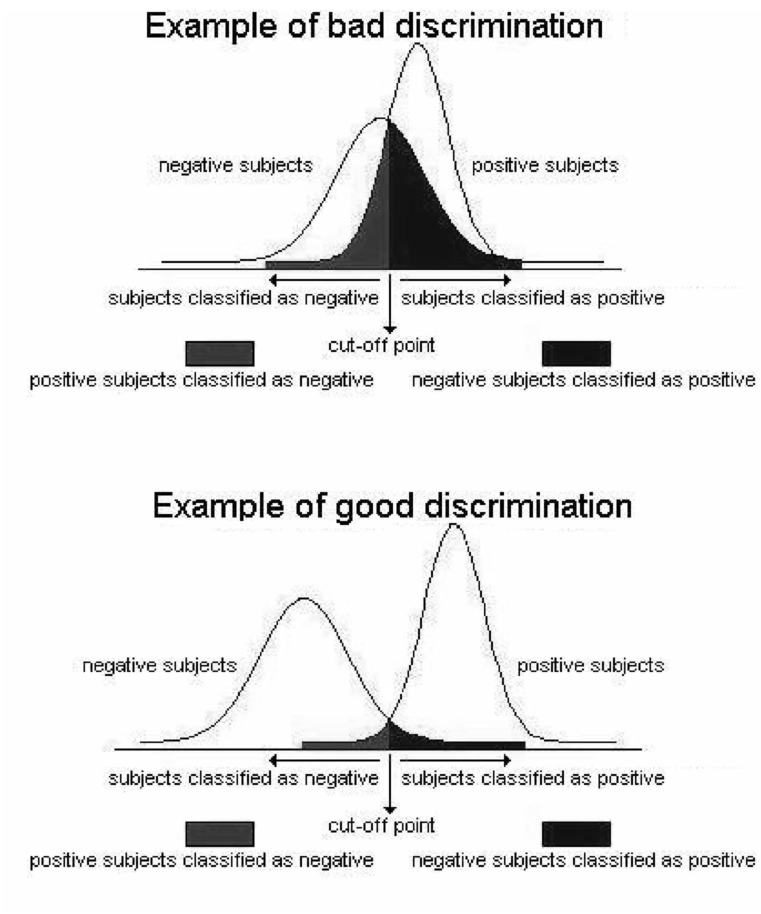
TABLE 1

Classification table

		predicted +	 -
	+	a	b
actual	-	c	d

We can define: $sensitivity = \dfrac{a}{a+b}$; $specificity = \dfrac{d}{c+d}$.

Sensitivity is a measure of the percentage of positive subjects which are classified as positive. Specificity is a measure of the percentage of negative subjects which are classified as negative. A classification table can be made for any possible choice of the cut-off point and the corresponding values of sensitivity and specificity can be computed. In general, a cut-off point with a high specificity tends to have a small sensitivity and vice versa.



Example of bad discrimination

negative subjects
positive subjects
subjects classified as negative
subjects classified as positive
cut-off point
positive subjects classified as negative
negative subjects classified as positive

Example of good discrimination

negative subjects
positive subjects
subjects classified as negative
subjects classified as positive
cut-off point
positive subjects classified as negative
negative subjects classified as positive

Figure 1 – Discrimination.

Sensitivity and specificity rely on a single cut-off point. A more complete description of classification accuracy is given by the area under the ROC (Receiver Operating Characteristic) curve. This curve plots the probability of correctly classifying a positive subject (sensitivity) against the probability of incorrectly classifying a negative subject (one minus specificity) for the entire range of possible cut-off points.

The area under the ROC curve, which ranges from zero to one, provides a measure of the model's ability to discriminate - the larger the area under the ROC curve, the more the model discriminates.

If we want to choose an optimal cut-off point for the purposes of classification, one might select a cut-off point that maximizes both sensitivity and specificity. This choice is facilitated by the use of the ROC curve area - the best choice for the cut-off point is approximately where the curve starts bending.

More intuitively, the interpretation of the ROC curve area can be viewed from a different angle. We can divide the sample into two sub-groups of positive and negative subjects. Let us call the n_1 subjects in the positive group $u_1, u_2, \dots, u_s, \dots, u_{n1}$ and the n_2 subjects in the negative group $v_1, v_2, \dots, v_t, \dots, v_{n2}$.

Let us consider a random pair of subjects (u_s, v_t), of which the first belongs to the positive group and the second belongs to the negative group.

Let us estimate $P(u_s)$ and $P(v_t)$. Since u_s has a positive outcome and v_t has a negative outcome, we would like the following inequality to be true:

$$P(u_s) > P(v_t) \tag{1}$$

If $P(u_s) > P(v_t)$, we say that the pair (u_s, v_t) is *concordant*; if $P(u_s) < P(v_t)$, we say that it is *discordant*; and if $P(u_s) = P(v_t)$, we say that the pair is a *tie*.

If we repeat this comparison for all the $n_1 \times n_2$ possible pairs, we can define the following statistic, which we call the *C statistic*:

$$C = \frac{\text{\# of concordant pairs} + 0.5 * \text{\# of ties}}{\text{\# of concordant pairs} + \text{\# of discordant pairs} + 0.5 * \text{\# of ties}}, \tag{2}$$

where $0 \leq C \leq 1$.

We can interpret the C statistic as the probability that a positive subject u_s and a negative subject v_t, results in $P(u_s) > P(v_t)$. That is, we are computing the number of times that a subject with a positive outcome has a higher event probability than a subject with a negative outcome. Half the ties are considered to be concordant pairs and half discordant pairs. This is clearly a measure of discrimination.

Hanley and McNeil (1982) showed that the C statistic is equivalent to the ROC curve area.

As a general rule (Hosmer and Lemeshow, 2000):

- if $C \geq 0.9$, the model is considered to have outstanding discrimination.
- if $0.8 \leq C < 0.9$, the model is considered to have excellent discrimination;
- if $0.7 \leq C < 0.8$, the model is considered to have acceptable discrimination;

- if C = 0.5, the model has no discrimination (predicted probabilities are *pure random:* we might as well flip a coin);
- If C < 0.5, the model has a *negative* discrimination, i.e. it is worse than random (Harrell *et al.*, 1996). Such a model tends to classify positive subjects as negative and negative subjects as positive.

We notice that the above mentioned levels should be adjusted according to the specific field of research. For instance, medical applications usually require higher values of the statistic than social studies or weather forecasting.

For a given model, we can compute the C statistic both for fitting and validation samples. We expect the C statistic to be larger for the former than for the latter. We can say that a model discriminates well on new subjects if the drop in value of the C statistic is quite small, and its value for the validation sample is still large enough.

3.2 *Calibration*

Calibration is a measure of how close the predicted probabilities are to the observed rate of the positive outcome for any given configuration of the independent variables of the model (D'Agostino *et al.*, 1998; Harrell *et al.*, 1996).

For a given configuration $\mathbf{X}$ of the independent variables, perfect calibration results in a prediction of the positive outcome that numerically agrees with the observed frequency of the event when that configuration occurs.

For example, if we have 8 statistical units in the sample that share the same configuration $\mathbf{X}_c$ of the independent variables and 6 of them have a positive outcome, we say that the model has a good calibration if $P(Y = 1 | \mathbf{X}_c) \cong 6/8 = .75$.

We define bias as:

$$\text{bias} = M(P_i) - M(Y) \tag{3}$$

where $M(P_i)$ denotes the mean of the predicted event probabilities P_i's in the sample and $M(Y)$ is the observed rate of positive outcomes in the sample. Bias is a crude but desirable measure of calibration (D'Agostino *et al.,* 1998).

Logistic regression is a mathematically unbiased method. Thus, it should be unbiased for the fitting sample and only slightly biased for the validation sample.

More accurate measures of calibration are often statistics which partition the data into groups and check how the average of the predicted probabilities compares with the outcome prevalence in each group.

Hosmer and Lemeshow (1980) have produced a widely used statistic to test the ability of a given model to calibrate.

Let the sample size be N. The most common version of this test arranges the subjects according to ascending predicted probabilities and divides them into Q groups of the same size so that the first group contains the N/Q subjects having the smallest estimated event probabilities, the second group contains the next N/Q subjects having the next smallest estimated event probabilities, and the Q-th group contains the N/Q subjects having the largest estimated event probabilities.

Usually the grouping is out of Q=10 deciles, but any other choice of number of groups is possible. It is suggested that each group contains enough subjects both on the fitting and the validation samples. The higher the number of covariates included in the model, the higher the number of subjects needed in each group for the test to have power.

Given the groups, the above mentioned test compares the observed number of positive outcomes (*prevalence* or *observed frequency*) with the mean of the predicted probabilities (*expected frequency*) in each group.

The more the groups' observed frequencies are close to the corresponding expected frequencies, the better the ability of the model to calibrate.

This closeness and thus the ability of the model to calibrate can be quantified using the Hosmer and Lemeshow χ^2 formula:

$$\chi^2 = \sum_{j=1}^{Q} \frac{(O_j - n_j P_j)^2}{n_j P_j (1 - P_j)} \tag{4}$$

where n_j, O_j, and P_j are respectively the number of observations, the number of positive outcomes and the average predicted probabilities for the j^{th} group.

Using an extensive set of simulations, Hosmer and Lemeshow (1980) demonstrated that, under the null hypothesis that the logistic regression model is the correct model, the statistic has approximately an asymptotic chi-squared distribution with Q-2 degrees of freedom. As a consequence, an observed value less than the critical value of the chi-squared distribution with Q-2 degrees of freedom at 0.05 α-level, indicates good calibration.

As for the choice of the number Q of groups, it is important that each group has enough subjects for the observed and expected event frequencies in the group to be significant. As a general rule, the higher the number of covariates included in the model, the higher the number of subjects needed in each group for the Hosmer and Lemeshow test to have enough power. As a basic indication, groups with less than ten observations could give a biased indication of calibration.

The advantage of the Hosmer and Lemeshow chi-squared test is that it provides analysts with a single, easily interpretable value that can be used to measure the calibration of a model.

We compute the Hosmer and Lemeshow chi-squared statistic for both the fitting and the development samples. For each group we compare the prevalence to the average predicted probability.

When applying the statistic to a validation set, the same number of groups should be used in order to compare the value of the statistic for the fitting and the validation samples. Figures 2 and 3 illustrates what we call *calibration bar plots*: they provide a visual interpretation of the above-mentioned statistic. If the model has a good calibration, the bars representing the prevalence and the average of predicted probabilities in each group should be very close to each other.

A reduction in calibration is to be expected as we pass from the fitting to the validation sample, but this reduction should be small. The bar plots in figure 2 and 3, for example, represent a model whose calibration ability pathologically degenerates when we pass from the fitting to the validation sample.

When the model generates predicted probabilities that are very close to 0 or to 1, the above mentioned statistic results in very large values because of the factor P_j (1-P_j) in the denominator. To avoid this problem and to better take into consideration the number Q of groups, the statistic has been modified to obtain the adjusted Hosmer and Lemeshow χ^2 formula:

$$\chi^2(adjusted) = \sum_{j=1}^{Q} \frac{(O_j - n_j P_j)^2}{n_j \left(P_j + \frac{1}{n_j} \right) \left(1 - P_j + \frac{1}{n_j} \right)} \quad (5)$$

The theory concerning the distribution of this statistic is not yet complete, but we can presume a distribution in some way similar to a chi-squared distribution with Q-2 degrees of freedom.

3.3 *Discrimination versus calibration*

A complete assessment of model performance should take into consideration *both* discrimination and calibration (Hosmer and Lemeshow, 2000). In any case, it is believed that discrimination is more important than calibration (Harrell *et al.,* 1996). If discrimination is good, the predictive model can in some way be recalibrated without sacrificing discrimination. If the predictive model has poor discrimination, no adjustment can correct the model.

Analysts should be aware that the importance of calibration relative to discrimination is strongly influenced by the field of analysis and the intended application. When predicting the efficacy of a new drug on a patient, for instance, not only are we interested in knowing if the drug has an effect or not (discrimination), but we are also concerned with an accurate prediction of the probability that the patient will recover from the disease within a defined period of time (calibration). In contrast, if our aim is to identify successful companies in order to have an insight into the best practices, then we might be more interested in a model with a very high discrimination, no matter what the value of calibration is. As a general rule, whenever the model is needed only to rank likely outcomes and not predict absolute probabilities, then calibration is much less important than discrimination (Harrell *et al.*, 1996).

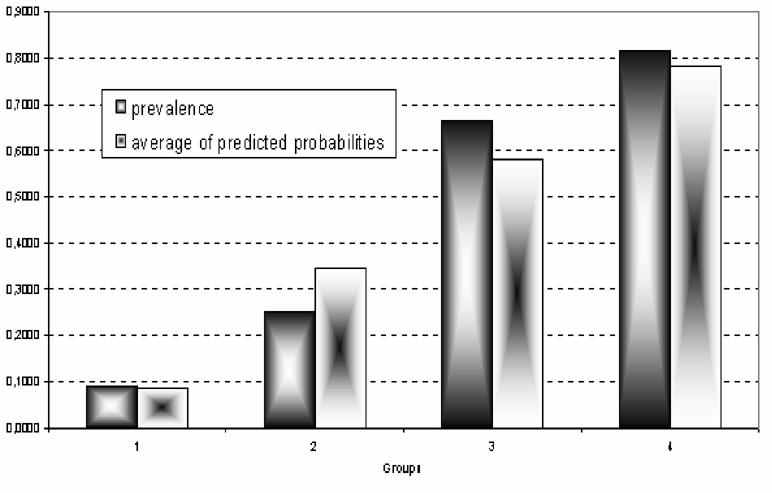


Figure 2 – Calibration bar plot – Fitting sample.

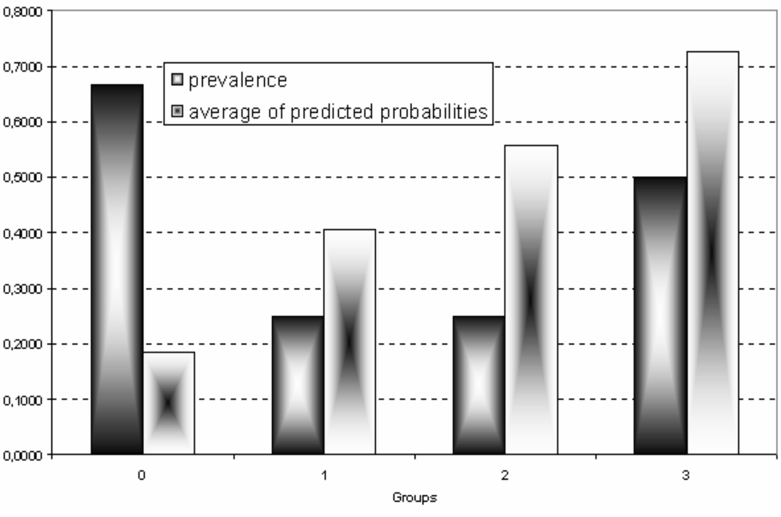


Figure 3 – Calibration bar plot – Validation sample.

Furthermore, note that a predictive model cannot have both a perfect calibration and a perfect discrimination (Diamond, 1992). A model that maximizes discrimination does so at the expense of calibration, and, on the other hand, a model that maximizes calibration does so at the expense of discrimination. As a consequence, a good model should balance between these two dimensions of performance in order to find the best trade-off.

4. MODEL VALIDATION PROCEDURE

In the previous paragraphs we have discussed what we mean by model validation and why it is a crucial phase of any modelling analysis. We have also introduced the main dimensions of model performance and a number of statistics used to quantitatively assess them. We now illustrate a possible strategy for model validation. The procedure is aimed at evaluating the performance of those models, with regards to their ability to predict the outcome variable. Because models are taken as they are, the procedure is very general in scope and applies to any statistical model that generates predicted probabilities.

As we have already mentioned and motivated, we follow the repeated data-splitting approach. Data-splitting is the act of partitioning available data into two portions for model validation purposes. The first portion of the data, called *fitting sample*, is used to fit the model and the latter, called *validation sample*, is used to evaluate its performance. The fitting sample is as large as 75% of the original sample and is extracted from the original sample through a simple random sampling without replacement (the choice of the cut-off point of 75% will be motivated later). The remaining data is used as a validation sample. That is, the split is purely random and there is no data duplication.

A possible objection to the use of splitting is the loss of information incurred in the model fitting. If we fit the model to the fitting sample rather than to all the original data, we "waste" some of the available data, i.e. we do not use all the available data to fit the model. Deriving the parameter estimates using only a portion of the data is a clear violation of the sufficiency principle (Picard and Cook, 1984). We have found a way for our strategy not to suffer due to that violation - at each split we fit the model to part of the data and we evaluate it on the remaining data. After we have performed all the necessary splits, and if the repeated splits prove that the model validates, then we compute its coefficients using all the data, according to the sufficiency principle.

Here follows a brief conceptual presentation of the procedure used to evaluate a regression model.

i) Data-splitting

The original sample is randomly split into the fitting and validation samples.

ii) Model fitting

The model is fitted on the fitting sample using the SAS logistic procedure. The *SAS logistic procedure* computes the model's coefficients, the overall significance of the model and the partial significance of each variable included in the model.

iii) Event probability estimation

We use the fitted model to estimate the probability of a positive outcome for each of the subjects in both the fitting and the validation samples.

As a consequence, for each subject we know:

- the configuration of the covariates $\mathbf{X}$;
- the observed outcome Y;
- the predicted probability of a positive outcome $P(Y = 1 | \mathbf{X})$.

We want to stress the fact that the model is fitted on the fitting sample (i.e., its coefficients are computed taking into account only the observations in the fitting sample), and is then used to estimate the event probability for the subjects both in the fitting and in the validation samples.

iv) Computation of performance measures on both samples

For both the fitting and the validation samples we compute the following statistics:

- C statistic (measure of discrimination);
- Hosmer and Lemeshow chi-squared test (measure of calibration);
- bias (measure of calibration).

v) Iterations and full model

In order for a model to be validated, the above described procedure is repeated 100 times. After that we also fit the model on the full original sample.

vi) Results

Each time the procedure is repeated the sample is split into two random portions, the model is fitted on one of the two portions, and its performance is evaluated on both portions. Since each iteration is based on a different split of the original data, it results in different model coefficients, significance levels, and performance values.

At the end of the 100 repetitions we obtain:

- the model fitted on the whole original sample;
- the distribution of the model's coefficients over the 100 fitted models;
- the distribution of the model's overall significance over the 100 fitted models;
- the distribution of the partial significance of all the covariates included in the model over the 100 fitted models;
- the distribution of each of the performance measures (c statistic, Hosmer and Lemeshow chi-squared test, and bias) over the 100 fitting samples;
- the distribution of each of the performance measures over the 100 validation samples.

Before proceeding, we define two terms which we shall use in the following paragraphs - the *fitting distribution* of a variable is the distribution of that variable computed on the fitting samples, while the *validation distribution* of a variable is computed on the validation samples.

vii) Presentation of the results

Results are presented using both tables and box-plots.

viii) Interpretation of the results

As for the quality of fit, we are concerned with the variability of the estimates of the model's parameters over the 100 repetitions. If the variability is large, then the model's coefficients highly depend on the particular portion of the original data used to fit the model. This is a clear symptom of instability of the model and, what is worse, of *overfitting*. We say that a model is overfitted when there are not enough data to compute a reliable estimation of the model's parameters. The whole available set of data is used to compute the model's parameters and there are no data left to assess the goodness of the estimation.

Going back to the assessment of the model's quality of fit, we would also like the distributions of the estimates to be averaged around the same values as the estimates computed on the whole original sample. If this does not happen, the model cannot be validated because of its internal instability.

In order for the model to be validated, we need to assess its performance outside the fitting sample. This can be done by comparing the fitting to the validation distributions of the measures of discrimination and calibration. Because of the mathematical optimization nature of the regression modelling technique, we expect the model to perform better on the fitting sample, i.e. we expect lower levels for both discrimination and calibration when shifting from the fitting to the validation distribution. Thus, a reduction in the magnitude of the performance measures is to be expected. If the drop in value of the measures is too large, the model does not validate outside the fitting sample.

5. EXAMPLE FROM A MANAGEMENT STUDY

The procedure has been developed in order to measure the performance of a model developed within a management study. In the specificity of the models, the subjects of the sample are companies, the response variable is a dichotomous variable measuring time performance in the New Product Development (NPD) Process of the company (where the occurrence of the event signifies that the company is among the best in the sample as for that dimension of performance), and the independent variables are drivers related to the NPD process, to the NPD practices, or to the NP Strategic guide and internal environment. The sample size consists of 85 subjects.

We present the results of the analysis of a logistic regression model including three variables: X_1 X_2 X_3 and all interactions up to the third-order interaction.

We remind the reader that the mathematical model is:

$$P(Y=1|\mathbf{X})=\frac{e^{b_0+b_1x_1+b_2x_2+b_3x_3+b_{12}x_{12}+b_{13}x_{13}+b_{23}x_{23}+b_{123}x_{123}}}{1+e^{b_0+b_1x_1+b_2x_2+b_3x_3+b_{12}x_{12}+b_{13}x_{13}+b_{23}x_{23}+b_{123}x_{123}}} \qquad (6)$$

5.1 *The model's quality of fit*

i) Stability of the parameters' estimates

Table 2 presents the parameters' estimates computed on the full sample of 85 units, and the information describing the fitting distribution of their estimates.

TABLE 2

Quality of the model's fit: stability of the parameters' estimates

Stability of the parameters' estimate						
	FIT=100%	FIT=75% - 100 iterations				
	FULL	MEDIANE	QRNG/2*	~ C Var**	MIN	MAX
b_0	5.64	5.31	2.03	38%	-5.26	16.4
b_1	-0.70	-0.58	0.61	105%	-4.44	3.87
b_2	-1.57	-1.55	0.61	39%	-5.26	0.99
b_3	-1.03	-0.88	0.76	86%	-4.92	2.37
b_{12}	-0.14	-0.19	0.20	106%	-1.47	0.79
b_{13}	0.10	0.06	0.23	379%	-0.91	1.28
b_{23}	0.22	0.19	0.18	95%	-0.64	1.60
b_{123}	0.07	0.08	0.06	72%	-0.27	0.35

* half of the interquartile range; ** ratio between QRNG/2 and the median

The variability is large, indicating the presence of some degree of overfitting. The estimates average around the values computed on the full sample.

We remind the reader that, when we find that the model validates outside the available sample, we use the parameter's estimates computed on the full sample. In this case, the model would be the following:

$$P(Y=1|\mathbf{X})=\frac{e^{5.64-0.70x_1-1.57x_2-1.03x_3-0.14x_{12}+0.10x_{13}+0.22x_{23}+0.07x_{123}}}{1+e^{5.64-0.70x_1-1.57x_2-1.03x_3-0.14x_{12}+0.10x_{13}+0.22x_{23}+0.07x_{123}}} \tag{7}$$

ii) Model's significance

Table 3 presents the overall significance of the model and the partial significance of each of the covariates. Again, we present both the information related to the full model and the information describing the fitting distribution.

The overall model has a very high significance both on the full sample (overall *p*-value = 0,001) and on the 100 fitting samples (median of the overall *p*-value fitting distribution = 0,006; interquartile range = 0,012).

All the covariates in the model are partially non significant. It is likely that the number of covariates included in the model (7) is too large if compared to the sample size (85).

As for the quality of the model's fit, we can conclude that the model is highly significant. At the same time, we can reasonably believe that there are too many covariates in the model compared to the sample size of 85 observations. This results in some degree of overfitting. As a matter of fact, there is too much variability in the estimates of the model's parameters over the 100 fitting samples, and, furthermore, the partial significance of any single covariate is too small.

TABLE 3

Quality of the model's fit: significance

Overall and partial significance of the model						
	FIT=100%	FIT=75% - 100 iterations				
	FULL	MEDIANE	QRNG/2	~ C Var	MIN	MAX
Overall_p	0.001	0.006	0.006	94%	0.000	0.099
p_0	0.387	0.489	0.160	33%	0.116	0.967
p_1	0.700	0.715	0.175	25%	0.146	0.994
p_2	0.421	0.514	0.147	29%	0.133	0.968
p_3	0.607	0.671	0.180	27%	0.135	0.978
p_{12}	0.789	0.715	0.165	23%	0.135	0.996
p_{13}	0.854	0.733	0.168	23%	0.172	0.995
p_{23}	0.683	0.708	0.152	21%	0.134	0.994
p_{123}	0.626	0.623	0.163	26%	0.132	0.998

5.2 *Generalizability of the model: validation*

i) Discrimination: C statistic

Table 4 and Figure 4 present the information concerning discrimination.

The model fitted on the full sample has an excellent discrimination (C = 0,803).

The model fitted on the full sample has an excellent discrimination (C = 0,803).

The fitting distribution of the C statistic is also very good. The median is 0,805 and the distribution is concentrated around the median (QRNG/2 = 0,019; ~C Var = 2%). Furthermore, the minimum value of the distribution is 0,741, which is still an acceptable value for discrimination. Thus the model discriminates very well between the two classes of outcome in the 100 fitting samples.

TABLE 4

Discrimination

Discrimination: C statistic						
	FIT=100%	FIT=75% - 100 iterations				
	FULL	MEDIANE	QRNG/2	~ C Var	MIN	MAX
C_fit	0.803	0.805	0.019	2%	0.741	0.874
C_val		0.725	0.062	9%	0.445	0.931

The validation distribution of the C statistic is not as good as the previous one, but still proves to be an acceptable discrimination of the model on the validation samples. The median value of the distribution is 0,725 and nearly 25% of the distribution is larger in value than 0,8.

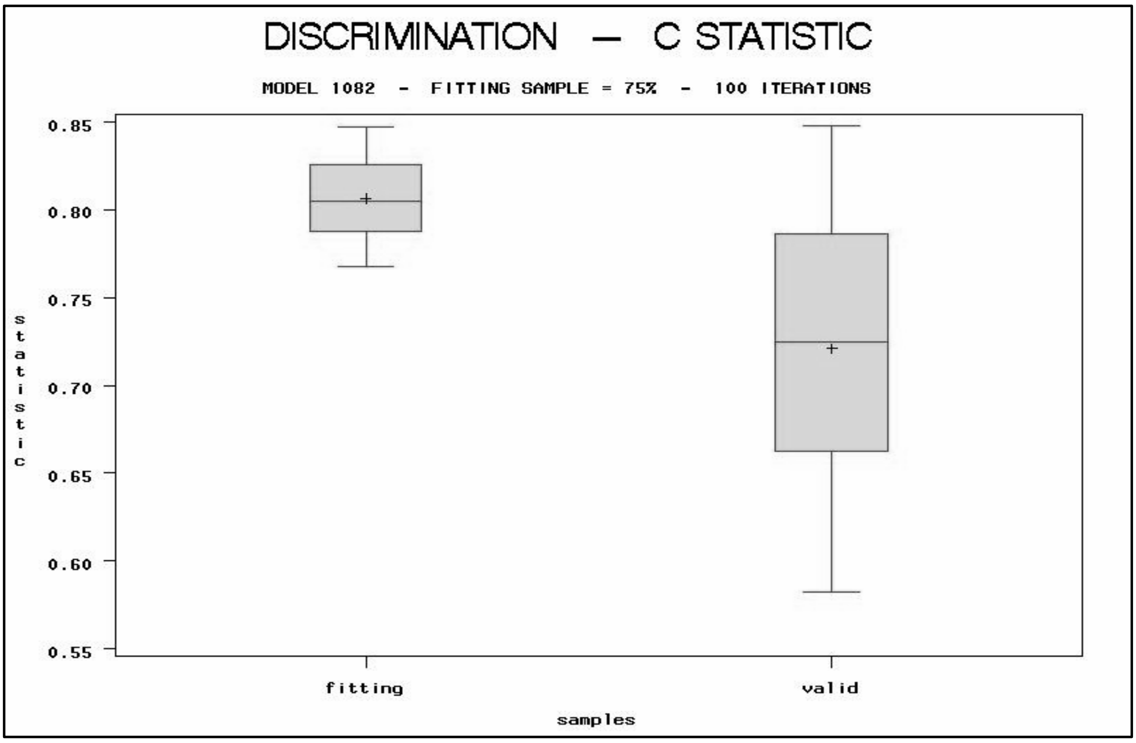


Figure 4 – Discrimination: fitting versus validation distributions.

We can conclude saying that the model perfectly discriminates on the original sample and discriminates well outside it. Thus we believe that the regression equation underlying this model has a discrimination property which is not specific to the available sample, but is extendable to the whole population. In other words, it is very likely that the model discriminates well on data which is new and independent from the sample used to fit the model.

ii) Calibration: Hosmer and Lemeshow test and bias

Table 5 and Figure 5 present the information regarding calibration.

We have introduced two measures of calibration:

- bias is a very simple but still desirable measure.
- the Hosmer and Lemeshow test is a very reliable measure of calibration, which results in a statistical test.

Note that we compute the statistic using only four groups in order to have enough observations in each group both for the fitting and for the validation samples. Since there are only four groups, we are not testing for an accurate calibration. We only require the model to rank each observation out of four possible classes of probability of the outcome as being positive.

We remind the reader that, under the null hypothesis that the logistic regression model is the correct model, the Hosmer and Lemeshow test has approximately an asymptotic chi-squared distribution with Q-2 degrees of freedom, where Q is the number of groups. For this model then, under the null hypothesis, the test has a chi-squared distribution with two degrees of freedom.

TABLE 5

Calibration

Calibration: Hosmer and Lemeshow test and bias

	FIT=100%	FIT=75% - 100 iterations				
	FULL	MEDIANE	QRNG/2	-CVar	MIN	MAX
HL_fit	1.312	1.15	0.60	53%	0.02	5.03
HL_val		5.53	3.41	62%	0.62	100.80
Bias_fit	0.00	-3E-08	1.42E-07	456	-4E-06	3.66E-06
Bias_val		-0.010	0.081	808	-0.233	0.201

Chi-square Distribuztion: 2DF
Alpha = 0.10: critical value =4.6052
Alpha = 0.05: critical value =5.9915
Alpha = 0.01: critical value =9.2103

Table 5 presents the critical values for the chi-squared distribution with 2 degrees of freedom for three different choices of the alpha-level. In figure 5, we only draw the critical value corresponding to the 0,5 alpha-level, i.e. 5,9915. That is, a value of the test smaller than 5,9915 indicates good calibration for the model in the example.

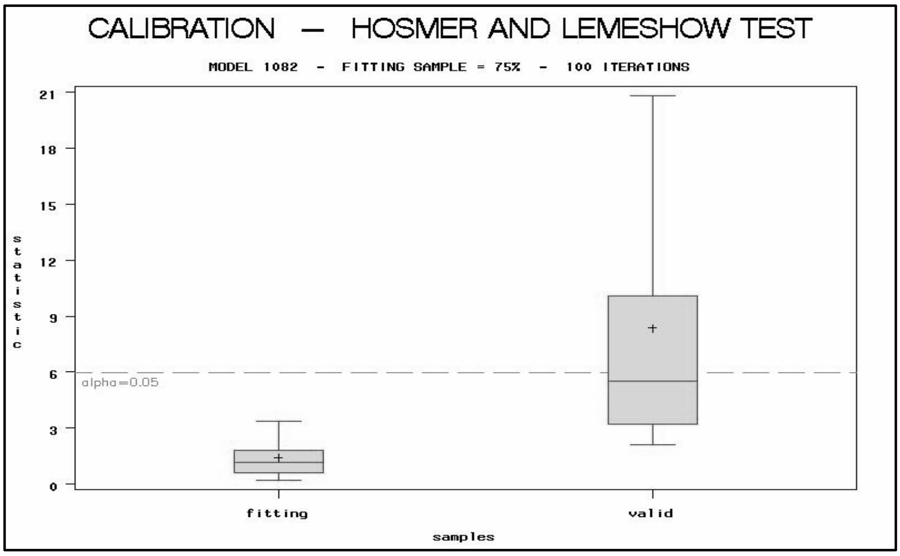


Figure 5 – Calibration: fitting versus validation distributions.

We can interpret the two distributions saying that the calibration on the fitting samples is excellent but it degenerates somewhat on the validation samples.

The validation distribution of the test has, in fact, a median value which is slightly smaller than the critical value: 5,53. That is, for almost half of the iterations, the model does not properly classify the subjects in the validation sample out of the four probability groups. As for bias, Figure 6 describes the fitting and the validation distributions.

Logistic regression is, as it should be, unbiased for the fitting samples. As for the validation distribution, the first quartile is –0,075, and the third is 0,087.

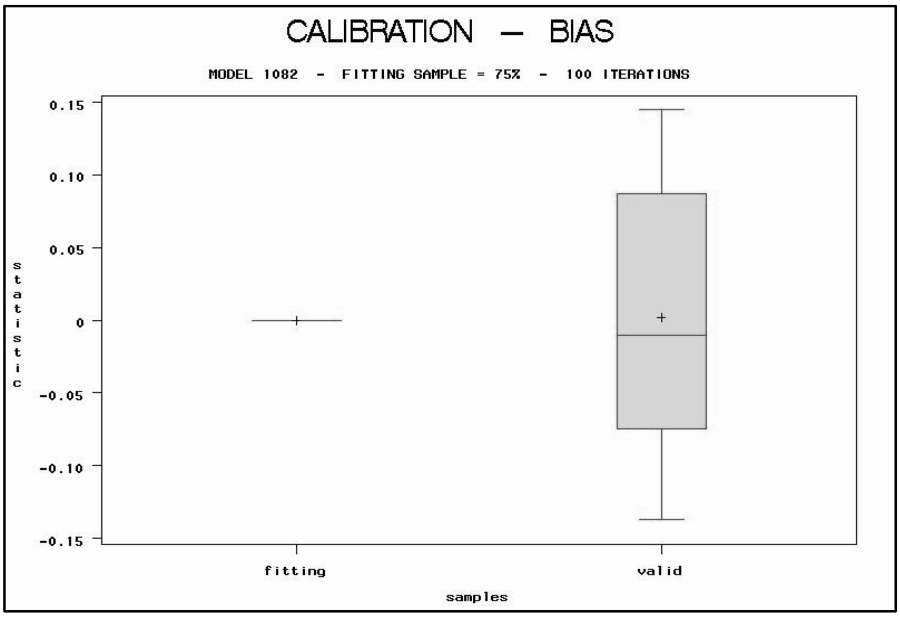


Figure 6 – Calibration: bias.

We have defined bias as the mean value of the predicted probabilities minus the event frequency in the sample. Each of the 100 validation samples has a size of 21 and about 10 positive subjects. This means that the observed event frequency in the sample is about 50%. A bias value of for instance 0,08 means that on average the predicted frequency is 58%. This results in a relative error of 16%. This confirms the indications of the Hosmer and Lemeshow test.

5.3 *Generic information about the samples*

Table 6 illustrates summary information about the different samples involved in the procedure:

TABLE 6

Generic information about the samples

Samples information						
	FIT=100%	FIT=75% - 100 iterations				
	FULL	MEDIANE	QRNG/2	-CVar	MIN	MAX
Fitting site	85	64	0			
n1_fit	41	31.0	1.0	3%	26.0	35.0
Validation site		21	0			
n1_val		10.0	1.0	10%	6.0	15.0

The original sample has a size of 85 observations and each split divides it into a fitting sample containing 64 observations, and validation sample containing the remaining 21 observations. Out of the 85 observations 41 are positive. Out of the 64 observations in the fitting samples about 31 are positive. Out of the 21 observations in the fitting samples about 10 are positive.

After having illustrated the procedure, and commented on its results, we want to assess the point of how many observations should be reserved to the fitting sample.

6. FINAL REMARKS: DETERMINATION OF THE OPTIMAL NUMBER OF OBSERVATIONS FOR THE FITTING AND VALIDATION SAMPLES

As a general rule, the ability to predict future observations suffers when too small a portion of the available data is reserved for model fitting. Similarly, the ability to assess the fitted model suffers when too small a portion of the available data is reserved for validation (Picard and Berk, 1990). Properly splitting the data involves a trade-off between these considerations.

In practice, the portion of observations reserved for validation is always less than ½ and is usually in the ¼ to $^{1}/_{3}$ range (Harrell *et al*, 1996).

We now want to justify the choice of reserving 75% of the available data to the fitting of the model, and the remaining 25% to validation. In particular, we want to demonstrate that the chosen cut-off point constitutes the best trade-off between the reliability of the model fitting phase and the reliability of the model validation phase. We do this by testing different cut-off points for the above described model and by comparing the results of the validation procedure.

Table 7 reports the model fitted on the full available sample and the results of the procedure for five different cut-off points:

- 1st column: the model is fitted on the whole available data;
- 2nd column: the fitting sample includes 80% of the available data;
- 3rd column: the fitting sample includes 75% of the available data;
- 4th column: the fitting sample includes 70% of the available data;
- 5th column: the fitting sample includes 66% of the available data;
- 6th column: the fitting sample includes 50% of the available data

From table 7 we can clearly see that the larger the portion of the available data reserved for the validation samples, the smaller the variability of the estimates of the model, and, on the contrary, the worse the calibration of the model on the validation samples.

TABLE 7

Determination of the optimal size of the fitting and validation samples

Data splitting: optimum cut-off point --- 100 iterations

Fitting size	100%	80%		75%		70%		66%		50%	
	All sample	Median	- CVar	Median	-CVar	Median	- Var	Median	- CVar	Median	-CVar
b_0	5.64	5.7	32	5.3	38	5.6	48	5.6	56	6.3	87
b_1	-0.70	-0.6	-96	-0.6	-105	-0.6	-147	-0.6	-175	-0.6	-343
b_2	-1.57	-1.6	-32	-1.6	-39	-1.7	-52	-1.7	-51	-1.8	-77
b_3	-1.03	-1.0	-47	-0.9	-86	-0.7	-113	-1.0	-94	-1.4	-133
b_{12}	-0.14	-0.2	-91	-0.2	-106	-0.1	-244	-0.1	-199	-0.2	-207
b_{13}	0.10	0.1	282	0.1	379	0.0	-4654	0.0	1328	0.1	781
b_{23}	0.22	0.2	69	0.2	95	0.2	115	0.2	101	0.3	134
b_{123}	0.07	0.1	53	0.1	72	0.1	113	0.1	86	0.1	164
Overall_p	0.001	0.005	84	0.006	94	0.006	241	0.011	121	0.022	186
p_0	0.387	0.465	29	0.489	33	0.514	35	0.479	39	0.529	40
p_1	0.700	0.717	26	0.715	25	0.662	27	0.690	26	0.547	43
p_2	0.421	0.481	27	0.514	29	0.520	39	0.517	36	0.530	39
p_3	0.607	0.664	19	0.671	27	0.706	27	0.641	28	0.614	38
p_{12}	0.789	0.719	18	0.715	23	0.674	34	0.652	31	0.636	32
p_{13}	0.854	0.776	22	0.733	23	0.701	24	0.701	27	0.560	36
p_{23}	0.683	0.705	19	10708	21	0.714	26	0.674	21	0.633	29
p_{123}	0.626	0.639	22	0.623	26	0.666	33	0.642	31	0.631	36
C_fit	0.803	0.807	3	0.805	2	0.815	4	0.815	3	0.833	4
C_val		0.736	11	0.725	9	0.716	9	0.726	7	0.693	7
HL_fit	1.31	1.08	68	1.15	53	1.15	88	1.26	74	1.47	58
HL_val		5.20	100	5.53	62	5.74	105	7.71	84	12	130
Bias_fit	0.00	-7E-08		-3E-08		-1E-08		-7E-09		-3E-09	
Bias_val		0.037		0.010		0.034		0.003		0.007	
Sample size_fit	85	68		64		59		56		42	
Sample size_val		17		21		26		29		43	

Also note that if the validation samples are too small, we do not have enough observations in each of the four groups used to reliably compute the Hosmer and Lemeshow test.

Among the five possible cut-off points, the best one seems to be the 75% one. It allows us to have enough data to fit the model and obtain good estimates of the parameters and, at the same time, enough data to validate it.

Centro per la modellistica, il calcolo e la statistica ROSA ARBORETTI GIANCRISTOFARO
Università di Ferrara LUIGI SALMASO

ACKNOWLEDGEMENT

Authors wish to thank Prof. Ralph B. D'Agostino for helpful suggestions and comments.

REFERENCES

R.B. D'AGOSTINO, J.L. GRIFFITH, C.H. SCHMIDT, N. TERRIN, (1998), *Measure for evaluating model performance*, in "Proceedings of the biometrics section", 1997, Alexandria, VA. American Statistical Association. Biometrics Section, pp. 253-258.

G.A. DIAMOND, (1992), *What price perfection? Calibration and discrimination of clinical prediction models*, "Journal of Clinical Epidemiology", 45, pp. 85-89.

B. EFRON, R. TIBSHIRANI, (1993), *An introduction to the bootstrap*, Chapman and Hall, New York.

J.A. HANLEY, B.J. MCNEIL, (1982), *The measure and use of the area under a receiver operating characteristic (ROC) curve*, "Radiology", 143, pp. 29-36.

F.E. HARRELL, K.L. LEE, D.B. MARK, (1996), *Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors*, "Statistics in Medicine", 15, pp. 361-387.

D.W. HOSMER, S. LEMESHOW, (1980), *A goodness of fit test for the multiple logistic regression model*, "Communications in Statistics", A10, pp. 1043-1069.

D.W. HOSMER, S. LEMESHOW, (2000), *Applied Logistic Regression*, 2nd edition. John Wiley & Sons, Inc., New York.

F. MOSTELLER, J.W. TUKEY, (1977), *Data analysis and regression: a second course in statistics*. Reading, MA; Addison-Wesley.

R.R. PICARD, K.N. BERK, (1990), *Data Splitting*. "American Statistician", 44, pp. 140-147.

R.R. PICARD, R.D. COOK, (1984), *Cross-validation of regression models*, "Journal of the American Statistical Association", 70, pp. 575-583.

SAS INSTITUTE INC., *SAS OnlineDoc®, Version 8*, Cary, NC: SAS Institute Inc., 1999.

J. STEVENS, (1996), *Applied multivariate statistics for the social sciences, third edition*, Lawrence Erlbaum Associates, Mahwan, New Jersey.

M. STONE, (1974), *Cross-validatory choice and assessment of statistical Predictions*, "Journal of the Royal Statistical Society", ser. B, 36, pp. 111-133.

RIASSUNTO

Analisi di adattamento e validazione di modelli di regressione logistica

In questo lavoro viene presentata una nuova procedura di validazione per modelli di regressione logistica. Inizialmente viene illustrata una breve sintesi delle diverse tecniche di validazione dei modelli. Vengono definite quindi le proprietà richieste ad un modello per essere considerato "buono" e un insieme di misure quantitative di adattamento. Infine è descritta una metodologia per la valutazione dell'adattamento di un modello utilizzato in uno studio di management aziendale.

SUMMARY

Model performance analysis and model validation in logistic regression

In this paper a new model validation procedure for a logistic regression model is presented. At first, we illustrate a brief review of different techniques of model validation. Next, we define a number of properties required for a model to be considered "good," and a number of quantitative performance measures. Lastly, we describe a methodology for the assessment of the performance of a given model by using an example taken from a management study.