Nov 7th, 9:00 AM - 5:00 PM

Statistical Issues in Predictive Modelling using Bootstrap Method

Elmira Kushta
University Ismail Qemali Vlora, kushtamira@gmail.com

Miranda Pajo
University Ismail Qemali Vlora, pajo.miranda@gmail.com

Orgeta Gjermeni
University Ismail Qemali Vlora, o.gjermeni@gmail.com

Statistical Issues in Predictive Modelling using Bootstrap Method

Elmira Kushta[1], Miranda Pajo[2], Orgeta Gjermeni[3]

[1,2,3]University "Ismail Qemali", Faculty of Technical Sciences, Vlora, Albania
kushtamira@gmail.com[1], pajo.miranda@gmail.com2
o.gjermeni@gmail.com[3]

Abstract. A model is a statement of reality or its approximation. Most phenomena in the social sciences are extremely complex. With a model we simplify the reality and focus on a manageable number of factors. It is impossible to completely understand why consumers default and identify all the factors influencing customer's default behavior. The bank manager sets up a statistical model that relates customer's default behavior to only two important factors, the income and the education. There surely are thousands of other variables that may influence customer's default behavior. This article describes the fundamentals of a statistical model building. We begin our discussion on the managerial justification for building a statistical model. Then we discuss three important statistical issues that are of prime importance to database marketers: model/variable selection, treatment of missing data, and evaluation of the model using Bootstrap Method.

Keywords: Predictive modelling, Customer, Evaluation, Bootstrap Method.

1. Introduction

This article describes the fundamentals of a statistical model building. We focus on issues that are important to database marketers. We begin our discussion on the managerial justification for building a statistical model. Then we discuss three important statistical issues that are of prime importance to database marketers: model/variable selection, treatment of missing data, and evaluation of the model.

2. Selection of Variables and Models

2.1 Variable Selection

Most predictive models (e.g., regression, logistic regression, neural nets) can be stated in the following regression-type format:

$$Y = f(X_1, X_2, X_3, \ldots X_K) + \varepsilon \quad (2.1.1)$$

where Y is the variable being predicted (customer response, customer value, etc.), the X's are the potential predictor variables, and ε are other (random) variables that have not been observed by researchers. Note that "K" is the number of potential predictor variables. In real- world applications, the value for K can be very high, easily in the hundreds if not in the thousands.

2.2 All-Possible Subset Regression

All possible subset regression is frequently used to determine the optimal set of independent variables. This procedure first requires the fitting of all possible combinations among the available independent variables. For example, if three independent variables are available, we need to fit eight regression equations, $\emptyset$, $\{X_1\}$, $\{X_2\}$, $\{X_3\}$, $\{X_1, X_2\}$, $\{X_1, X_3\}$, $\{X_2, X_3\}$, and $\{X_1, X_2, X_3\}$. Next, select the best regression equation using some statistical criteria such as adjusted R^2 AIC (Akaike Information Criteria) or BIC (Bayesian Information Criteria):

$$Adjusted\ R^2 = 1 - \left[\frac{n-1}{n-k}\right](1 - R^2) \quad (2.2.1)$$

$$AIC = -2\log\hat{L} + 2k \quad (2.2.2)$$

$$BIC = -2\log\hat{L} + k\log n \quad (2.2.3)$$

where n is the number of observations, k is the number of predictors including the intercept and $\hat{L}$ is the value of the likelihood function achieved by the model. Different from R^2, these criteria penalize more complex models so that a simple model may often be chosen if the increase in fit by including additional variables is not large enough. We select the best model with the largest adjusted R^2 or the lowest AIC or BIC. The adjusted R^2 is used for linear regression models while AIC and BI C can be used for both linear and non-linear models.

2.3 Principal Components Regression

Principal components analysis is a technique for combining a large number of variables into a smaller number of variables, while retaining as much information as possible in the original variables. Suppose we have an n × k matrix of X of n observations on k variables, and Σ is its variance–covariance matrix. The objective of principal components analysis is to find a linear transformation of X into a new set of data de- noted by P, where P is n×p and p ≤ k. The p variables in P are called "factors" and the n observations for each factor are called factor scores. Hence, the principal components transformation is a rotation from the original X coordinate system to the system defined by the principal axes of this ellipsoid. Specifically, the transformation to principal components is given by

$$P = M' X \quad (2.3.1)$$

To see how M (p × n) is determined, post-multiply P'. Then, P P' = M'X X'M. X X' is simply the variance–covariance matrix Σ. The variance–covariance matrix for principal components P P' = Λ should be diagonal by virtue of requirement (i) above. Hence, we have:

$$\Lambda = M' \Sigma M \quad (2.3.2)$$

Equation (2.3.2) is an orthogonal similarity transformation diagonalizing the symmetric matrix Σ. The transformation matrix M has an orthonormal set of eigenvectors of Σ as its columns, and P P'= Λ has the eigenvalues of Σ as its diagonal elements. If the columns of M are ordered so that the first diagonal element of Λ contains the largest eigenvalue of Σ, the second the next largest, etc., the principal components will be ordered as specified in requirement (ii). Instead of fitting a linear regression Y = Xβ + ε, principal components regression fits the following regression:

$$Y = P\gamma + \varepsilon \quad (2.3.3)$$

where P is factor scores from Equation (2.3.1). Once the values of P are determined from principal components analysis, the parameters γ can be estimated by ordinary regression.

3. Bootstrap

Given a dataset of size n, the principle of the bootstrap is to select samples of size n with replacement from the original sample. Since the bootstrap samples are selected with replacement, some cases are typically sampled more than once. Originally introduced by Efron (1983), bootstrapping has been shown to work better than other cross-validation techniques, especially in small samples. There are various bootstrap methods that can be used for estimating prediction error and confidence bounds. One of the simplest is the 0.632 bootstrap in which a dataset of n observations is selected (with replacement) from an original sample of size n. Since some cases are sampled more than once, there are cases that are not picked. Those observations not included in the bootstrap sample are used as validation samples.

4. Application of Bootstrap Method in R Programming

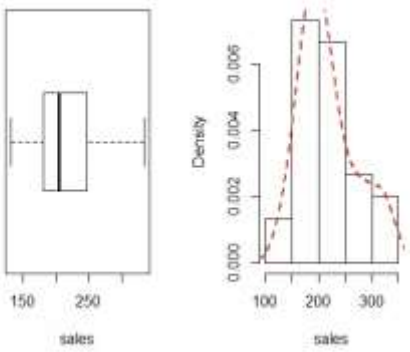
A fabricate company, Big Market store chain, is selling a new type of grape juice in some of its stores for pilot selling. The marketing team of ABC wants to analyze:
Which type of in-store advertisement is more effective? They have placed two types of ads in stores for testing, one theme is natural production of the juice, the other theme is family health caring;

The Price Elasticity – the reactions of sales quantity of the grape juice to its price change;
The Cross-price Elasticity – the reactions of sales quantity of the grape juice to the price changes of other products such as apple juice and cookies in the same store;
How to find the best unit price of the grape juice which can maximize the profit and the forecast of sales with that price.
The marketing team has randomly sampled 30 observations and constructed the following dataset for the analysis. There are 5 variables (data columns) in the dataset.
Data Exploration

```
#load the libraries needed in the following codes
> library(s20x)
> library(car)
> #read the dataset from an existing .csv file
> df <- read.csv(file.choose(),header=T)
> #list the name of each variable (data column) and the first six rows of the dataset
> head(df)
sales price ad_type price_apple price_cookies
1   222  9.83       0        7.36         8.80
2   201  9.72       1        7.43         9.62
3   247 10.15       1        7.66         8.90
4   169 10.04       0        7.57        10.26
5   317  8.38       1        7.33         9.54
6   227  9.74       0        7.51         9.49
> # basic statistics of the variables
> summary(df)
sales          price          ad_type     price_apple    price_cookies
Min.   :131.0   Min.   : 8.200    Min.   :0.0   Min.   :7.300   Min.   : 8.790
1st Qu.:182.5 1st Qu.: 9.585  1st Qu.:0.0  1st Qu.:7.438    1st Qu.: 9.190
Median :204.5 Median : 9.855 Median :0.5  Median :7.580   Median : 9.515
Mean:   216.7  Mean:   9.738   Mean   :0.5   Mean:   7.659  Mean   : 9.622
3rd Qu.:244.2  3rd Qu.:10.268 3rd Qu.:1.0   3rd Qu.:7.805    3rd Qu.:10.140
Max. :   335.0   Max. :   10.490  Max. :   1.0   Max. :   8.290   Max. :   10.580
```

From the above summary table, we can roughly know the basic statistics of each numeric variable. For example, the mean value of sales is 216.7 units, the min value is 131, and the max value is 335. Please ignore the statistics of the "ad_type" there since it is a categorical variable.
We can further explore the distribution of the data of sales by visualizing the data in graphical form as follows.

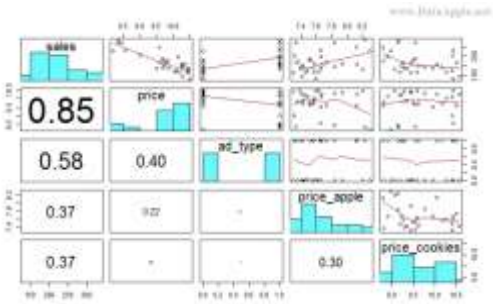
```
> #set the 1 by 2 layout plot window
> par(mfrow = c(1,2))
> # boxplot to check if there are outliers
> boxplot(df$sales,horizontal = TRUE, xlab="sales")
> # histogram to explore the data distribution shape
> hist(df$sales,main="",xlab="sales",prob=T)
> lines(density(df$sales),lty="dashed",lwd=2.5,col="red")
```



We don't find outliers in the above box plot graph and the sales data distribution is roughly normal. It is not necessary to apply further data cleaning and treatment to the data set.

Sales Driver Analysis and Price Elasticity Analysis

With the information given in the data set, we can explore how grape juice price, ad type, apple juice price, cookies price influence the sales of grape juice in a store by multiple linear regression analysis. Here, "sales" is the dependent variable and the others are independent variables.
Let's investigate the correlation between the sales and other variables by displaying the correlation coefficients in pairs.
> pairs(df,col="blue",pch=20)
> pairs20x(df)



The correlation coefficients between sales and price, ad_type, price_apple, and price_cookies are 0.85, 0.58, 0.37, and 0.37 respectively, that means they all might have some influences to the sales, so we can try to add all of the independent variables into the regression model as follows.
> sales.reg<-lm(sales~price+ad_type+price_apple+price_cookies,df)
> summary(sales.reg)
Call:
lm(formula = sales ~ price + ad_type + price_apple + price_cookies,
data = df)
Residuals:
Min 1Q Median 3Q Max
-36.290 -10.488 0.884 10.483 29.471
Coefficients:
Estimate Std. Error t value Pr(>|t|)

```
(Intercept)    774.813    145.349   5.331 1.59e-05 ***
price          -51.239      5.321  -9.630 6.83e-10 ***
ad_type         29.742      7.249   4.103 0.000380 ***
price_apple     22.089     12.512   1.765 0.089710 .
price_cookies  -25.277      6.296  -4.015 0.000477 ***
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Residual standard error: 18.2 on 25 degrees of freedom
Multiple R-squared: 0.8974, Adjusted R-squared: 0.881
F-statistic: 54.67 on 4 and 25 DF, p-value: 5.318e-12

The p-value for price, ad_type, and price_cookies in the last column of the above output is much less than 0.05. They are significant in explaining the sales. We are confident to include these three variables into the model.

The p-value of price_apple is a bit larger than 0.05, seems there are no strong evidence for apple juice price to explain the sales. However, according to our real-life experience, we know when apple juice price is lower, consumers likely to buy more apple juice, and then the sales of other fruit juice will decrease. So we can also add it into the model to explain the grape juice sales.

The Adjusted R-squared is 0.881, which indicates a reasonable goodness of fit and 88% of the variation in sales can be explained by the four variables. The remaining 12% can be attributed to other factors or inherent variability. Please note the R-squared is very high here because the dataset were made up rather than from real world data sources.

Conclusion

The scientific method for predicting the future is based on the assumption that the future repeats the past. For many applications, this assumption is reasonable. Suppose we try to predict monthly sales of color television. We may build forecasting models, (whether they are time-series models or regression models) based on historical sales of color television, and isolate pat- terns from random variations. The predicted sales of a color TV are based on the estimated model (or identified patterns). However, the future can be very different from the past especially the market conditions are changing. The model becomes useless. This is why we need to keep updating models. Sometimes it may be enough to re-estimate the model with additional data. Sometimes we need to change the model itself. Remember that the model cannot be static.

References

1. B. Efron, R. Tibshirani (1986). Bootstrap Methods for Standart Errors, Confidence Intervals, and Other Measures of Statistical Accuracy.
2. Shao, J. & Tu, D. (1996), The Jackknife and Bootstrap, Springer.
3. Abrevaya, J. and Huang, J. (2005). On the Bootstrap of the Maximum Score Esti-mator. Econometrica, 73 1175{1204.
4. Bickel, P. and Freedman, D. (1981). Some Asymptotic Theory for the Bootstrap. Ann. Statis., 9, 1196{1217.
5. Sen, B. and Banerjee, M. and Ëoodroofe, M. (2008). Inconsistency of Bootstrap: the Grenander estimator. Submitted.
6. Bodhisattva, S. (2008). A Study of Bootstrap and Likelihood Based Methods inNon-Standart Problems
7. LeBaron, B. (1997). An Evolutionary Bootstrap Approach to Neural Netëork Pruning and Generalization Philip M. Dixon (2001).
8. The Bootstrap Hesterberg,T. and Monaghan,Sh. and S. Moore,D. and Clipson,A. and Epstein, R. (2003). Bootstrap Method and Permutation Tests.
9. George Antonogeorgos, Demosthenos B. Panagiotakos, Kostas N. Priftis and Anastasia Tzonou, (2009).

10. Logistic Regression and Linear Discriminat Analysis in Evaluating Factors Associated ëith Asthma Prevalence among 10- to 12-Years –Old Children: Divergence and Similaritity of the Two Statistical Methods. Hindawi
11. Publishing Corporation, International Journal of Pediatrics.
12. Muça M., PUKA Ll., BANI K, SHAKAJ F. (2013) "Logistic Regression Analysis: A Model to Predict
13. Entrance Probability in Higher Education". "1st International Western Balkans Conference of Mathematical
14. Sciences – IWBCMS-2013", në Elbasan/ALBANIA PROCEEDINGS

Sadri Alija, Lazim Kamberi and Llukan Puka, (2011). Logistic regressions and an application in teaching