This assignment contains 9
numbered pages.

Erasmus University Rotterdam

Econometric Institute

Code FEB22009(X,Q,S)

Introductory Seminar Case Studies Econometrics and Operations Research

Academic year 2024–2025

Marketing Case

Eoghan O'Neill, Stan Koobs, Bernhard van der Sluis & Hans Ligtenberg

1 Introduction

Due to the introduction of barcodes on grocery products and scanner checkouts it is nowadays straightforward for supermarkets to register the total sales of products and brands on a daily or weekly basis. At the same time the prices and available promotional activities of products and brands are stored. These enriched sales data allow store managers to get insight in the (dynamic) effects of promotional activities. It also gives store managers the opportunity to predict the future sales of products. Accurate forecasts of future sales are necessary to determine the optimal stock of supermarkets to prevent empty shelves.

Constructing accurate forecasts of sales of SKUs (Store Keeping Units, i.e. distinct types of items for sale) is a challenge as there may be many predictors available. Not only prices and promotional activities of the SKU itself may be relevant but also the prices and promotional activities of competitive SKUs may influence the sales. Furthermore, due to post promotion dips lagged sales, lagged prices and lagged promotional activities may play a role as well. If the number of SKUs in a product category is large, one may have many potential predictors of which some may be irrelevant.

Selecting the optimal number of predictors is not a straightforward task. Statistical tests to select the relevant predictors are often not appropriate. First of all, it is difficult to control the size of the test in case you do many statistical tests in a row. Furthermore, the selection of the regressors often depends on the order of the tests (e.g. general-to-specific versus specific-to-general approach) and it is not always clear which order is best. Finally, the number of predictors may be larger than the number of available observations such that it is even not possible to estimate the parameters of a sales model with all predictors. The last decade alternative variable selection methods have become popular in the statistical learning literature. These methods are based on shrinking the parameters measuring the influence of the predictors towards zero, see James et al. (2013, Section 6.2).

In this case study we will consider forecasting the sales of refrigerated orange juice in a particular store. Each team will get 102 weeks of sales data from a sample of 10 stores of the same chain. The main output for this case will be forecasting models for the unit sales of 11 brands of refrigerated orange juice for the stores under consideration. The focus is on the selection of the forecasting model. You may consider statistical testing approaches for predictor selection and machine learning approaches, including shrinkage methods.

2 Data

For the case we use the Dominick's Finer Foods chain data collected in the greater Chicago area. The data contain weekly unit sales, shelf prices, and non-price promotions for 11 brands of orange juice in the refrigerated subcategory over 102 weeks from September 1989 to August 1991 for a sample of 10 stores.

For this assignment, every group of students receives data from 10 stores in an Excel file. Your group number indicates which file to use. The file contains

- week_t: week number
- $\text{sales}_{it,\ell}$: unit sales of SKU i in week t for store ℓ.
- $\text{price}_{it,\ell}$: price per ounce of SKU i in week t in cents for store ℓ.
- $\text{feat}_{it,\ell}$: 1 in case there is feature promotional activity for SKU i in week t and 0 otherwise for store ℓ.
- $\text{deal}_{it,\ell}$: 1 in case there is in-store coupon activity for SKU i in week t and 0 otherwise for store ℓ.

for SKU $i = 1, \ldots, 11$, week $t = 1, \ldots, 102$, and store $\ell = 1, \ldots, 10$. More information about the source of the data can be found in, for example, Wedel and Zhang (2004). The 11 SKUs are defined as a combination of brand name and size of the package

1.	Tropicana Premium 64 oz	7.	Citrus Hill 64 oz	
2.	Tropicana Premium 96 oz	8.	Tree Fresh 64 oz	
3.	Florida's Natural 64 oz	9.	Florida Gold 64 oz	quick
4.	Tropicana 64 oz	10.	Dominick's 64 oz	
5.	Minute Maid 64 oz	11.	Dominick's 128 oz	
6.	Minute Maid 96 oz			

Note that Dominick's is a store brand (private label). The 11×10 sales series are the series which you have to predict.

3 Predicting Sales

Let $S_{it,\ell}$ denote the sales of SKU i in week t for store ℓ and let $x_{jt,\ell}$ be a potential predictor for $j = 1, \ldots, K$. Although unit sales $S_{it,\ell}$ is discrete we often treat these sales figures as a continuous variable and use a linear regression model to predict sales

$$\ln S_{it,\ell} = \beta_0 + \sum_{j=1}^{K} x_{jt,\ell}\beta_j + \varepsilon_{it,\ell} \tag{1}$$

for $t = 1, \ldots, T$ and $i = 1, \ldots, 11$, where β_0 is the intercept parameter and β_j describes the effect of the predictors $x_{jt,\ell}$ on log sales. The term ε_{it} denotes the error term. [In this example, a separate model is specified for each store $\ell = 1, \ldots, 10$, but you may consider other modeling approaches]

The goal of this case is to predict the future sales of SKU i, that is, you want to predict $S_{i,T+1}$. The question is which potential predictors $x_{jt,\ell}$ to include in the prediction model.

[Note that if you decide to transform the target variable to the log of sales or apply some other transformation, you will have to decide on the appropriate re-transformation to obtain forecasts of the original variable (the level of sales), see, for example `https://en.wikipedia.org/wiki/Smearing_retransformation`]

Variable Selection

In practice the number of potential predictors K may be large. Forgetting to include a relevant predictor may lead to a bias in your forecast. At the same time including an irrelevant predictor may increase the variance of your forecast due to efficiency loss. For optimal forecasting, it is necessary to balance between these two situations.

A strategy to determine the relevant predictors is to use statistical tests. One can, for example, conduct individual t-tests on the significance of the β_j parameters. A famous approach in this context is the general-to-specific approach, see among others, Heij (2004, p. 281). Such approaches have drawback as already mentioned in the introduction.

Another strategy is to iterate over all possible selections of regressors and using information criteria like BIC (Heij, 2004, p. 279) to select the optimal number of regressors. This approach is usually not feasible. In case of K predictors, the number of different sets of predictors is 2^K and hence in case $K = 20$, we already have more than 1 million possible different selections of regressors, see also James et al. (2013, Section 6.1) for a discussion about this approach.

To circumvent the shortcomings of the above mentioned methods, other variable selection methods introduced in the machine learning literature are nowadays often used. To select the appropriate set of regressors, one puts a penalty term on the size of the parameters which shrinks the parameters towards 0. The parameters $\beta = (\beta_0, \beta_1, \ldots, \beta_K)$ can be estimated using OLS (Ordinary Least Squares). To impose shrinkage a penalty term is directly put on the parameter values. For a simple introduction to these shrinkage methods, one can look at Section 6.2 of James et al. (2013). To treat every predictor the same, we first normalize all variables in (1). Hence, we subtract the mean from $\ln S_{it}$ and divide by its standard deviation. We do the same for every predictor x_{jt} for $j = 1, \ldots, K$. Hence, we subtract the mean and divide by the standard deviation of the predictor. The normalized data denoted by $\widetilde{\ln S_{it}}$ and $\tilde{x}_{jt}$. All variables now have zero mean and in case of no functional misspecification we can remove the intercept from the model. The parameters of the model with the normalized variables are now indicated by $\tilde{\beta}_j$ for $j = 1, \ldots, K$.

Ridge regression

The first regularization one may use is a quadratic penalty on the value of the parameters. The estimator is given by

$$\min_{\tilde{\beta}} \left(\sum_{t=1}^{T} (\widetilde{\ln S_{it}} - \sum_{j=1}^{K} \tilde{x}_{jt}\tilde{\beta}_j)^2 + \lambda_R \sum_{j=1}^{K} \tilde{\beta}_j^2 \right), \tag{2}$$

The resulting estimator for $\tilde{\beta}$ is called the Ridge estimator (Hoerl and Kennard, 1970b,a). The parameter λ_R has to be set by the researcher. The estimator can even be computed if $K > T$ (if $\lambda_R > 0$) . There exists a closed form expression for this Ridge estimator given by

$$\hat{\beta}_R = \left(\sum_{t=1}^{T} \tilde{X}_t \tilde{X}_t' + \lambda_R I_K \right)^{-1} \left(\sum_{t=1}^{T} \tilde{X}_t \widetilde{\ln S_{it}} \right) \tag{3}$$

where $\tilde{X}_t = (\tilde{x}_{1t}, \ldots, \tilde{x}_{Kt})'$ and I_K a K-dimensional identity matrix. Note for $\lambda_R = 0$ the Ridge estimator is the regular OLS estimator. This approach however does not select regressors but limits the influence of regressors, see (James et al., 2013, Section 6.2). It shrinks the parameters towards zero and the larger λ_R the smaller the values of the Ridge estimator (in absolute terms). This estimator usually works well in situations where one has many weak predictors. In case one really wants to select regressors one has to put a different penalty term on the $\tilde{\beta}$ parameters.

LASSO

Instead of putting a quadratic penalty term on the regression parameters one can also put a penalty on the absolute value of the parameters and consider

$$\hat{\beta}_L = \min_{\tilde{\beta}} \left(\sum_{t=1}^{T} (\widetilde{\ln S_{it}} - \sum_{j=1}^{K} \tilde{x}_{jt}\tilde{\beta}_j)^2 + \lambda_L \sum_{j=1}^{K} |\tilde{\beta}_j| \right), \tag{4}$$

where λ_L is again a tuning parameter to be set by the researcher. The resulting estimator is called a LASSO estimator (Tibshirani, 1996). There is no closed form expression for the estimator but there are excellent algorithms to compute its value. A property of this estimator is that it puts zero weights to some regressors and hence that it does variable selection and parameter estimation simultaneously, see again James et al. (2013) for an introduction.

Potential Predictors

Finally, we have to decide which predictors we are allowed to use to forecast the 11 refrigerated orange juice series. To make the prediction exercise realistic you are only allowed to use variables for predicting sales of SKU i in week $T+1$ which are known in week T for $i = 1, \ldots, 8$. There is however one exception. Store managers know at least 1 week ahead what the prices and promotional activities of their SKUs will be. Therefore, when predicting the sales for week $T+1$ you are allowed to use the prices and promotional activities of all SKUs (and subcategories prices and promotions) in week $T+1$.

Forecast Evaluation

The data sets covers 102 weeks. You should use a rolling window approach to evaluate one-week ahead forecasts. You must choose the window length and the number of windows. If the first window begins at time t_1 and ends at time t_L, then the window length is $t_L - t_1$, and the model for the first window is trained with all data from t_1 to t_L. This model is used to forecast sales in period $t_L + 1$. You can use all data from t_1 to t_L and also prices, $feat$ and $deal$ in $t_L + 1$ to produce the sales forecasts for $t_L + 1$. The second window begins at time $t_1 + 1$ and ends at time $t_L + 1$, and the model trained using the second window is used to forecast sales at time $t_L + 2$, and so on. You can use the one week ahead forecast errors produced across all windows to evaluate the models.

A variable selection procedure can be re-implemented separately for each window, or it can be applied only to the first window or a subset or windows. You cannot use data that occurs after the end of a window to select or train the model used for that window's forecast.

You can decide exactly how to implement the model evaluation procedure. There are no strict rules for window size or number of windows. However, models trained on larger windows are more accurate, so a window size of at least 40 observations might be a reasonable minimum. On the other hand, a larger number of windows will improve the precision of the estimation od expected mean forecast error (or other criteria), and improve the power of statistical tests (for example, tests comparing expected forecast losses of methods). Therefore, a minimum of 20 rolling windows would be expected.

For a review of forecast evaluation methods, including rolling windows, see Tashman (2000). For some forecasting tasks, standard cross-validation can be applied for predictor selection, parameter tuning, or forecast comparison. See Bergmeir and Benítez (2012) and Bergmeir et al. (2018).

To pass this case, you should at least compare a statistical testing procedure to select/shrink the predictors with a LASSO and Ridge approach to make forecasts for the 11 SKUs. Extensions, e.g. machine learning methods or tests for differences in expected forecast losses, may improve your grade when well motivated, described and executed.

4 Econometric Software

You are free to choose any software package you want to solve the case.

Take care that you provide enough information in your report to make it possible for a reader to reproduce your results when the data are available. So you should be clear about estimation sample, which estimator you used, which variables you include and so forth. Do not provide information which is irrelevant. It is, for example, important to mention that you take the logarithm of a variable but it is of course not important to mention that you have done this in Eviews with the command *genr lnx=log(x)*. In case you use commercial or free software written by someone else, you have to indicate the name of the package and the version number.

5 Questions

For questions about the case, please use the discussion board of CANVAS. We will check the discussion boards at the end of the day. Do not post answers on CANVAS!

6 Reports

6.1 Research Proposal

On June 5, before 12:00hrs, you must submit your proposal on CANVAS. It should contain a maximum of 2 pages with the statistical/econometric models and methods you are going to use to solve the research question including a discussion on explanatory variables in your models. This proposal will be discussed at the compulsory meetings on June 6. The content of the 2 pages determines the level of independence and creativity of your research. For a fair judgment, oral adjustments and extensions of the proposal will not be taken into account when judging independence (groups can learn from each other what is correct and wrong). The schedule of these meetings will be made available on CANVAS.

The research proposal should at least contain:

- The set of potential predictors you are going to use to predict sales at time t. Provide the right time index for the variables you are going to include to predict $\ln S_{i,t,\ell}$, for example, ($\text{price}_{1,t,\ell}$,...,$\text{price}_{11,t,\ell}$)

- How you are going to construct the predictions for the level of sales (not the log sales) given the parameter estimates and which criteria you are going to use to evaluate the predictions and why?.
- A discussion how you are going to determine the values of λ_R and λ_L

Be clear and use proper notation (correct indices)!

6.2 Interim/progress Report

For the second compulsory meeting you have to prepare at least 2 questions (maximum one page) and submit them before Wednesday June 11, 12:00hrs via CANVAS. Note you are not allowed to ask for solutions to the research project or direct questions whether your results are correct. The schedule of the second compulsory group meetings will be made available on CANVAS and takes place on June 12.

6.3 Final Report

The results of this assignment have to be reported in a scientific report, see also the general seminar rules for the requirement for this report. The report has to be scientific and hence it is not the idea that you provide a time line of the work you did. The targeted audience is a fellow econometrician with Bachelor 2 knowledge of econometrics with no knowledge of the case description. Think carefully about which results you want to report and which results are not of interest.

The econometric analyses have to be reproducible. Hence, mention not all details but **enough details to reproduce** your results. It is, for example, very important for this case to report how **exactly** you determine the tuning parameters in Ridge and LASSO. If we are unable to reproduce your results, the grade cannot be high. The proposed econometric methodology must be described in detail and well motivated and you have to indicate which data (including transformations and estimation sample) you use for parameter estimation. The report also has to contain a discussion about the results and it should contain a general conclusion in line with the results.

It is not appropriate if Eviews output (or the output of any other package) is directly included in the report, especially if the reader has to puzzle what the parameters and statistics in the Eviews table mean. Provide insightful names to variables in the report. Only include tables and graphs if they are relevant for your research and if you discuss them in your report.

The first version of the final report has to be submitted via CANVAS before Monday June 16, 12:00hrs. The peer review on this report has to be ready before Wednesday June 18, 12:00hrs. The final report has to be submitted before Friday June 20, 23:59hrs.

Note: Only the student who submits the first version of the final report, can submit the peer review and also receives feedback.

Summary Deadlines

1. June 2: Start Case 3
2. June 5, 12:00hrs : Submission Proposal
3. June 6: Discussion Proposal (Compulsory)
4. June 11, 12:00hrs : Submission Progress
5. June 12: Discussion Progress (Compulsory)
6. June 16, 12:00hrs : Draft Report
7. June 18, 12:00hrs : Peer Review Report
8. June 20, 23:59hrs : Final Report

See description above for details about the exact content and rules of the submissions

References

Bergmeir C, Benítez JM. 2012. On the use of cross-validation for time series predictor evaluation. *Information Sciences* **191**: 192–213.

Bergmeir C, Hyndman RJ, Koo B. 2018. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis* **120**: 70–83.

Heij C. 2004. *Econometric methods with applications in business and economics.* Oxford University Press.

Hoerl AE, Kennard RW. 1970a. Ridge regression: applications to nonorthogonal problems. *Technometrics* **12**: 69–82.

Hoerl AE, Kennard RW. 1970b. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**: 55–67.

James G, Witten D, Hastie T, Tibshirani R, et al. 2013. *An introduction to statistical learning*, volume 112. Springer.

Tashman LJ. 2000. Out-of-sample tests of forecasting accuracy: an analysis and review. *International journal of forecasting* **16**: 437–450.

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **58**: 267–288.

Wedel M, Zhang J. 2004. Analyzing brand competition across subcategories. *Journal of marketing research* **41**: 448–456.