The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0959-0552.htm

Optimal pricing in retail: a Cox regression approach

Rudi Meijer
De Bijenkorf, Amsterdam, The Netherlands, and
Sandjai Bhulai
Department of Mathematics, VU University Amsterdam, Amsterdam, The Netherlands

Downloaded by Vrije Universiteit Amsterdam At 12:41 01 July 2016 (PT)

Abstract

Purpose – The purpose of this paper is to study the optimal pricing problem that retailers are challenged with when dealing with seasonal products. The friction between expected demand and realized demand creates a risk that supply during the season is not cleared, thus forcing the retailer to markdown overstocked supply.

Design/methodology/approach – The authors propose a framework based on a Cox regression analysis to determine optimal markdown paths. They illustrate this framework by a case study on a large department store.

Findings – The framework allows one to determine when and how much to markdown in order to optimize expected total profit given the available supply. When the law of demand holds at a disaggregated level, i.e. the individual retailer, it is also possible to optimize the markdown path.

Originality/value – This paper provides a framework for the complex dynamic pricing problem in retail using transactional data. The case study shows that significant revenues can be generated when applying this framework.

Keywords Cox regression, Markdown strategies, Optimal pricing, Retailers, Pricing policy, Survival analysis, Discounts

Paper type Research paper

1. Introduction

In much biomedical research the use of survival analysis to model the time to a certain event is common (Fleming and Harrington, 2005; Kalbfleisch and Prentice, 2002). In a survival context Cox regression is used to measure the influence of covariates on the survival rate. The same approach can be used in economic situations where we have supply and demand. In that case, we study the time it takes for an article to be sold given the price of the article. We characterise the distribution of the time to event for a given set of articles, and we compare different marketing policies for the articles. For example, a commercial price discount for a particular set of articles for a limited set of time versus a control set of articles with no discount at all.

Price is a covariant and one would like to find out whether the law of demand holds for the price and the time to event (Theil, 1976). If this law holds, then low-priced articles have short event times and high-priced articles have long event times. In our study we use T-shirts as the fashion article category that stands for most retail-fashion goods. Given the fact that in retail the selling price is in principle the cost price plus a markup, we can only optimise the profit when the law holds (Afriat, 1987).

In survival analysis the data are collected over a finite period of time, for retail this is a season, and consequently the time to event cannot be always observed. Fashion

International Journal of Retail & Distribution Management
Vol. 41 No. 4, 2013
pp. 311-320
© Emerald Group Publishing Limited
0959-0552
DOI 10.1108/09590551311330564

Downloaded by Vrije Universiteit Amsterdam At 12:41 01 July 2016 (PT)

mostly sustains for just one season. If the product is not sold at the end of the season we have so-called censored data. Because of censored data the normal summary statistics do not have the desired statistical properties such as unbiasedness (Zhang, 2005).

Right censoring is a simple example of censoring. When there is shortage, the potential demand is higher than the available supply; we are confronted with more complicated censoring. Calendar and event times are not necessarily the same. With different time entries for the supplies to stores we have also left truncation. The event time is the difference between birth (supply) and death (selling) time. The supply of the inventory is in bulk and consequently we have to deal with tied entries. The price is time covariant and we have to address this in our estimates. With time-dependent covariance the hazards are no longer proportional. As in biomedical applications there are good reasons in retail to use the apparatus of survival analysis (Fleming and Harrington, 2005; Zhang, 2005), due to the fact that:

- calendar and event times are not the same;
- possibly censored data;
- tied entries; and
- time covariance.

We study the optimal pricing problem and propose a framework based on Cox regression analysis to determine optimal markdown paths. The framework allows one to determine when and how much to markdown in order to optimise expected total profit given the available supply. When the law of demand holds at a disaggregated level, i.e., the individual retailer, we can also optimise the markdown path. We illustrate this framework by a case study on a large department store. The case study shows that significant revenues can be generated when applying this framework.

2. Cox regression

We give three formulas: the hazard, the survival function, and a derived relation between the markdown and the price elasticity estimate. The hazard and the survival estimates are common in most commercial software and directly available for operational use (SPPS Inc., 2010). Cox regression estimates the proportional hazard. The hazard rate is the instantaneous rate of failure at time t given that an individual is alive (at risk) at time t. In a retail context, it is the instantaneous rate of selling the product at time t given that the product is available (at risk) at time t, i.e.:

$$h(t|x) = h_0(t)e^{\sum x_i \beta_i}.$$

In Cox regression the hazard function estimates the relative risk of failure. The hazard function is a rate, it is not a probability. The Cox regression is used to determine the influence of predictor variables on a dependent variable. The x is the covariate, in our context the realised retail price. The β is the regression coefficient. Common in most software is to use the partial likelihood to estimate the regression coefficients. The proportional hazard model implies that the effect of the covariate on the relative hazard is constant over time. If there is non-proportionality over time we have to create a time-dependent covariate. The time-dependent covariate can then be used to fit a non-proportional hazard model in which time is included as a predictor $h_0(t)$.

The base rate is to be estimated independently from the proportional hazard:

$$S(t|x) = e^{-h(t|x)}.$$

The survival function is an estimate of the probability of surviving longer than a specified time. In the retail context, this is the probability that the article has not yet been sold. As we can see, the survival function and the hazards are related. The definition of sell-through is given by: $1 -$ "survival rate". Given the sell-through, we can derive at time t the required price difference to aspire to a desirable sell-through at time t, given by:

$$\Delta p = -\frac{\ln\left[\frac{\ln S_1(t)}{\ln S_0(t)}\right]}{bp_0}. \tag{1}$$

3. Data preparation

The necessary input data for the regression analysis must be carefully prepared. First, we had to derive for every article the event time. An article arrives in an outlet in multiple amounts. In our data preparation we use the first in – first out principle. This means that the item first arrived is considered to be sold first. Because of the arrival in bulk we have tied data. In the literature there are several suggestions to address the problem of tied data. Tied data means that the different items of the same article have the same arrival time and cannot be individually followed (Kalbfleisch and Prentice, 2002; Zhang, 2005). We index the different items and interpolate the event times. When during the season the article is not sold, then the event time is censored. We distinguish two types of markdowns. When the markdown is temporary, we have a commercial discount. When the markdown is permanent, we have a seasonal markdown. This means that in the data, the retail price after a calendar time does not jump back to the full initial retail price. Notice that the retail price still can be reduced additionally during the calendar time. Shrinkage and customer returns of the article are considered censored. Before estimating, we create a time-dependent covariant. For all non-censored events we multiplied the full retail price with a descending index of event time. To put the time covariant on the same scale as the price we divided by 1,000. The Cox regression now takes place on the realised (turnover) price and the time-related covariant.

4. Dataset

We gathered a dataset from a department store in The Netherlands for the first half year of 2011. We examined 206,714 items for women and 175,430 items for men. In total there were 101 different brands. We measured in days, so a season is about 185 days. One can find more detailed information in Table I.

The distribution of the price points are given in Figure 1. The average price for men was 32.01 and for women this was 40.86.

	Total	Brands	Sell-through without markdown	Sell-through total
Women	206,174	62	0.50	0.95
Men	1754,30	39	0.68	0.94

Table I.
Factsheet

Downloaded by Vrije Universiteit Amsterdam At 12:41 01 July 2016 (PT)

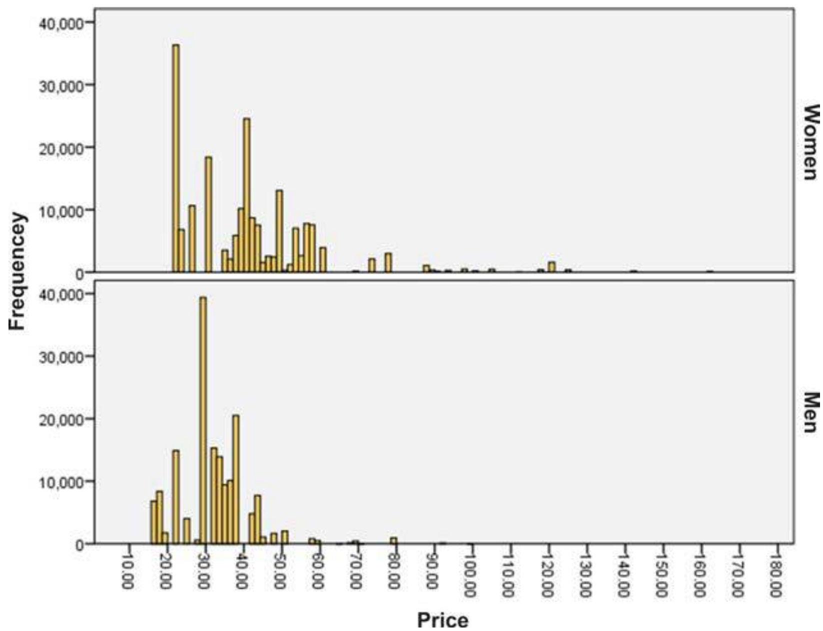

Figure 1.
Price points

Downloaded by Vrije Universiteit Amsterdam At 12:41 01 July 2016 (PT)

5. Results from estimation

In Table II one can see a sample of the available data after preparation. A naive approach would be to get rid of the censored data and then to estimate the survival probabilities. We entered 235,997 times, the number of terminal events was 185,405. A naive estimate gives us $185,405/235,997 = 0.67$. The survival rate would be $1 - 0.67 = 0.33$. The right estimate that corrects for censored data gives 0.29. The naive approach is over-estimating the true survival rate and is too pessimistic for the sell-through. Correcting for censored data is important for getting unbiased results. Here we analysed 5,962 different T-shirts items for both men and women in different colours and sizes during one season. A first estimate without dealing with time-related covariance would be a regression on the realised retail prices. We do not intent to use the different characteristics like colour or size as regression factors, while the influences are implicit in the event times. Our main goal is to find the price-elasticities and use them in optimal pricing.

Statistics for T-shirts status

Our first estimate for just one brand shows how important it is to deal with time-covariance, see Table III. In Table III, *B* is the estimated coefficient. It is

Table II.
Sample data after preparation

Time	Enter	Withdrawal	Terminal
28	136,580	181	2,639
29	133,430	385	2,533
30	130,512	0	1,800
31	128,712	142	1,087

interpreted as the predicted change in the log hazard for a unit increase in the predictor (price). The variable SE is the standard error of the estimated coefficient *B*, and Wald is the Wald statistic. If df = 1, the Wald statistic can be calculated as $(B/\mathrm{SE})^2$. It is used to test whether the estimated coefficients are significantly different from 0. In this test, the used distribution is chi-square, and df is the degree of freedom. We do not use category variables so the degree of freedom is 1. Sig is the significance level for the Wald statistic, and Exp(*B*) is e^{xb}, it is the relative risk (i.e., the ratio of the risk for different price levels).

The positive value of the regression coefficient suggests that a markdown is contrary to our beliefs. It would mean that by lowering the price, the survival probability increases and the sell-through is decreasing. In our example the estimate is, however, significant, since the Wald statistic is large. Since Exp(*B*) is greater than 1, it indicates that there is an increased relative risk when the retail price is increased. It is not what we expect from the law of demand. We have to incorporate time-covariance to test whether the law of demand holds.

The estimates with the time covariant were done for the entire population of T-shirts, see Table IV. Introduction of the time-dependent covariant gives us the confirmation of the law of demand. Both estimates are highly significant with high Wald statistics. The value of Exp(*B*) for the price is less than 1, which indicates that there is an increased relative risk when the price is decreasing. This is what we expect. The survival rate is decreasing and the sell-through is increasing with a markdown of the price. As already mentioned we do not use the characteristics of the product, like colour and size.

It is not recommended, for example, to use a stepwise approach. The residuals in the survival do not have the same properties as in normal linear regression. The characteristics would be proved in a stepwise approach to be significant while the price dependency drops out (Therneau and Grambsch, 2000). The Cox regression uses the partial likelihood estimation to find the parameters. The base rate is estimated independently of the Cox regression. The partial likelihood is rather heuristic, and is nowadays motivated by martingale theory. The theory gives interesting results especially when we are investigating the residuals in Cox regression (Aalen *et al.*, 2009; Gill, 1984; Gjessing *et al.*, 2008). One way to find the parameters is to simulate the martingale residuals and use an MCMC approach to derive parameters for every desired characteristic level of the product. If we picture the derived different hazard curves we find a multitude of hazards. We see the importance of the price for the explanation of the variance in the sell-through of articles.

	B	SE	Wald	df	Sig.	Exp(*B*)
Price	0.312	0.028	121.1562	1	0.000	1.366

Table III. Estimates for one brand without time-covariates

	B	SE	Wald	df	Sig.	Exp(*B*)
Price	−0.015	0.000	8,294.783	1	0.000	0.985
Tcov	0.006	0.000	122,498.036	1	0.000	1.006

Table IV. Estimates for all brands with time-covariates

Downloaded by Vrije Universiteit Amsterdam At 12:41 01 July 2016 (PT)

The base survival curve is depicted in Figure 2(a). On the horizontal axis we have the days in the season, on the vertical axis we find the survival rate. As an illustration of a multitude of survival curves we pictured for one brand and one colour the graphs in Figure 2(b). In the table with the hazards per colour, Table V, we see that there are clear differences in colours. Yellow colours have high hazard while orange colours have low hazard. Differences in taste are expressed in the hazard for the product. Systematic measuring of the hazard can be significant for an optimal allocation of the supply. In Table VI, we see that larger sizes have higher hazards. We can expect that markdowns probably will appear in the small sizes. Between men and women differences exist in hazard mainly because of the multiple use of T-shirts by men, see Table VII. Men use T-shirts as a fashion item but also as underwear. The different uses are expressed in the different hazards. Clearly, the commercial discount had no real increasing effect on

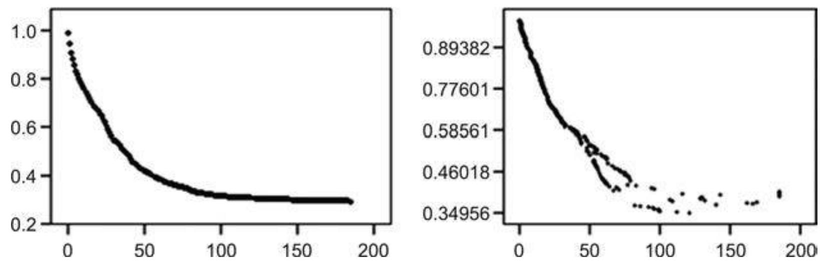

Figure 2.
Hazard curves

Table V.
Hazard values per colour

Colour group	Hazard
Blue	0.73
Brown	0.66
Yellow	0.82
Grey black	0.68
Green	0.69
White	0.66
Orange	0.47
Purple	0.73
Red	0.73

Table VI.
Hazard values per size

Size	Hazard
Small	0.66
Medium	0.67
Large	0.70
X-large	0.72

Table VII.
Hazard values per gender

Gender	Hazard
Men	0.72
Women	0.66

Downloaded by Vrije Universiteit Amsterdam At 12:41 01 July 2016 (PT)

the hazard of the articles, see Table VIII. The seasonal markdown lags the product with no discount. This is what can be expected, given the problematic sell-through of these articles. In Figure 3(a) we can see the relation between the initial retail price and the derived hazard.

Conclusion. Survival analysis can be put in practice in retail questions like finding the optimal path for markdowns. There are clear differences in characteristics of the articles. Preference for different brands is reflected in the realised retail price.

6. Optimal markdowns

We can use equation (1) of Section 2 to optimise the expected sell-through related to the needed intensity of the markdown. Finding the optimal path can be done with Markov Decision Processes (MDPs) but is beyond the scope of this article. In Figure 2(a) we notice that the basic survival curve also could be estimated parametrically. The Weibull distribution is flexible enough and can be used as a reference norm illustrating the possibilities. Because residuals in survival analysis as a counting process are martingales we can use the Doob-Meyer decomposition. The decomposition counting process = compensator + martingale is analogous to the statistical decomposition: data = model + noise (Therneau and Grambsch, 2000), and is given by:

$$N_i(t) = M_i(t) + \int_0^t Y_i(x)\lambda_i(x)dx.$$

The $N_i(t)$ is the counting of the event at time t, $M_i(t)$ is the residual, and $Y_i(x)$ $\lambda_i(x)$ is the model.

We can use the analogue to model the hazard rate in relation to the price variation. The residuals as the difference between the true survival rate and the estimated base survival curve capture still the not yet identified influence of the covariates (prices). We find this depicted in Figure 3(b). We used a smoother to identify the relation captured

Table VIII. Hazard values per status

Status	Mean hazard
Censored	0.81
No markdown	0.72
Commercial discount	0.23
Seasonal markdown	0.53

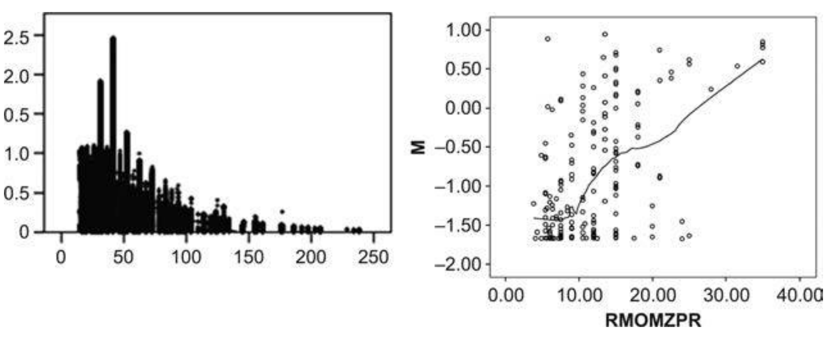

Figure 3. Price curves

Downloaded by Vrije Universiteit Amsterdam At 12:41 01 July 2016 (PT)

in the residuals and price. See for example the remarks in Therneau and Grambsch (2000) for the importance of smoothing and identifying relations.

Price differences give us an explanation for the variation in survival (sell-through). So we have a relation for the change in sell-through for a change in price which is just the definition for the price elasticity for an article.

We use the Weibull distribution (Zhang, 2005) at the start of a new season to have a norm sell-through curve. If a particular article lags behind this norm, then we can derive, given the norm, how much markdown is needed to fill the gap between planned and realised sell-through via:

$$S(t) = e^{-0.115t^{0.484}}.$$

As an example, let us process the articles for which at calendar time $t = 100$ days the sell-through lags behind the Weibull norm. The mean norm at calendar day 100 for our articles was 0.48 (see Table IX). The mean sell-through for the articles at calendar day 100 was 0.43. We found 782 articles items for which the sell-through lags behind the expected norm. It confirms our results, low-priced articles have a higher sell-through risk compared with high prices articles.

When we apply equation (1) we get the results in Table X. The Δp in equation (1) is related to the norm and the existing survival and existing retail price. Our estimated price-elasticity is now a universal constant. If the norm $S_0 \geq S_1$, then Δp is negative. So the necessary markdown is dependent on the difference (gap) between the norm and the actual survival weighted by price and price-elasticity. Contrarily, it can also be argued that when the norm stays behind the actual survival, there is a risk for shortage and price increase could theoretically be an option.

Equation (1) opens the way for optimal pricing where we already noticed that finding the optimal path at the beginning of the season in principle could be done in a Markov Decision Problem context. The Weibull curve in our approach is used as a

Table IX. Parameters for illustrative example

	n	Min.	Max.	Mean
Norm	782	0.76	0.11	0.48
Sell-through(t)	782	0.70	0.06	0.43
Price	782	16.64	161.20	33.01

Table X. Markdowns to achieve a desirable sell-through

Price	Markdown
20	−0.50
30	−0.34
40	−0.25
50	−0.20
60	−0.17
70	−0.14
80	−0.13
90	−0.11
100	−0.10
160	−0.06

Downloaded by Vrije Universiteit Amsterdam At 12:41 01 July 2016 (PT)

prediction so that we can use it as our expected value that illustrates the potential for a more rational approach in the retail price strategy. While we do not explore the potential of MDPs, we can still illustrate the considerations to be done when a markdown is suggested. We can find the opportunity costs, i.e., the profit lost when we have markdowns and the profit lost when at the end of the season we write-off the articles not sold.

Notice that the Weibull curve in this context gives us a prediction of the sell-through without markdown. The prediction of the sell-through with markdown is just to apply the Cox-regression model derived with the change in price at calendar time t. At calendar time $t = 100$ we calculated for 3,348 articles the opportunity return while their actual sell-through lagged behind the Weibull-norm. The return is defined as the "realised retail price" − "cost price". The cost price is given and constant and the realised retail price changes as a consequence of the planned markdowns.

We compare the weighted return with markdown and without markdown. The weighting is based on the expectations of the sell-through. We examined a -0.30 markdown. For only 4.3 per cent of the articles the expected return with markdown was not profitable compared with a write-off of the articles not sold at the end of the season. The average return with markdown compared without markdown and write-off increased with 3 per cent. Given these preliminary results further investigation to derive an optimal allocation of prices is warranted.

References

Aalen, O.O., Andersen, P.K., Borgan, Ø., Gill, R.D. and Keiding, N. (2009), "History of applications of martingales in survival analysis", *Journal Électronique d'Histoire des Probabilités et de la Statistique*, Vol. 5 No. 1, pp. 1-28.

Afriat, S.N. (1987), *Logic of Choice and Economic Theory*, Clarendon Press, Oxford.

Fleming, T.R. and Harrington, D.P. (2005), *Counting Processes and Survival Analysis*, Wiley, Chichester.

Gill, R.D. (1984), "Understanding Cox's regression model: a martingale approach", *Journal of the American Statistical Association*, Vol. 79 No. 386, pp. 441-447.

Gjessing, H.K., Aalen, O.O. and Borgan, Ø. (2008), *Survival and Event History Analysis: A Process Point of View*, Springer Verlag, New York, NY.

Kalbfleisch, J.D. and Prentice, R.L. (2002), *The Statistical Analysis of Failure Time Data*, Wiley-Interscience, New York, NY.

SPPS Inc. (2010), *IBM SPSS Statistics 19 Algorithms*, SPSS, Chicago, IL.

Theil, H. (1976), *Theory and Measurement of Consumer Demand*, North Holland Publishing, Amsterdam.

Therneau, T.M. and Grambsch, P.M. (2000), *Modelling Survival Data: Extending the Cox Model*, Springer Verlag, New York, NY.

Zhang, D. (2005), *Analysis of Survival Data*, Lecture Notes, North Carolina State University, Raleigh, NC.

About the authors

Rudi Meijer has been an active research member for more than 25 years in retail. His fields of research are related to consumer analytics and logistic challenges. He has been actively

Downloaded by Vrije Universiteit Amsterdam At 12:41 01 July 2016 (PT)

developing working systems for markdown policies and optimal replenishment. Currently, he is carrying out research on retail to obtain his PhD degree at VU University Amsterdam.

Sandjai Bhulai is an Associate Professor in Applied Probability and Operations Research at VU University Amsterdam. His primary research interests are in the general area of stochastic modelling and optimisation, and in particular, the theory and applications of Markov decision processes. His favourite application areas include telecommunication networks, call centres, health care, and logistics. He is currently involved in the control of time-varying systems, partial information models, dynamic programming value functions, and reinforcement learning. Sandjai Bhulai is the corresponding author and can be contacted at: s.bhulai@vu.nl

To purchase reprints of this article please e-mail: **reprints@emeraldinsight.com**
Or visit our web site for further details: **www.emeraldinsight.com/reprints**

Downloaded by Vrije Universiteit Amsterdam At 12:41 01 July 2016 (PT)