A Frequentist Approach to Bayesian Regression

a Case Study on the Weekly Sales of Orange Juice

FEB23100: Bachelor's Thesis Econometrics and Operational Research

Anne de Weerdt, 412029

Erasmus University Rotterdam
Supervisor: R. Paap
Co-reader: W.Wang

July 2nd

Abstract

In this paper frequentist approaches to Bayesian regression are tested and compared to methods using ordinary least squares, pooling and Ridge estimation. We investigate how the prior parameters of the Bayesian regression could be established by means of a frequentist approach. The resulting estimates are evaluated based on their predictive performance. This paper focuses on the prediction of weekly sales of different types of orange juice incorporating explanatory variables of competing products including price and lagged sales. The data originates from the Dominick's Finer Food chain in the greater Chicago area and is provided by The Kilts Center for Marketing at the Universtiy of Chicago's Marketing Department. We introduce two Bayesian inspired shrinkage methods of which one outperforms all the other methods showing great promise to modeling cross-effects between competing products using limited data.

1 Introduction

Accurately predicting product sales is crucial to the optimization of profit in the field of retail. In order to minimize stock and waste, while simultaneously maximizing sales, it is necessary to have a good knowledge on the performance of individual products. When companies predict the sales of a product accurately this prevents under and over stocking. Especially products that are expensive to keep in stock by taking up a lot of space and/or losing their value over time, are of interest for this study.

The sales of products can depend on a variety of factors. Besides commonly used explanatory variables such as price and promotions, also individual product characteristics and brand recognition can be of great influence. When combining many of such variables, the amount of coefficients to be estimated quickly grows because many cross-effects need to be computed. When data is scarce, this can lead to inaccurate results of the estimated coefficients due to the reliance on asymptotic distribution theory.

An easy way to reduce the number of parameters would be to pool certain explanatory variables by assuming equal effects across individual products. This might however be inappropriate as not all product's sales act in the same way leading to less than optimal predictions. For example a product like fresh orange juice has quite a short expiration date, which disincentives people to buy many packages at once. Shelf-stable orange juice on the other hand can be kept easily for months and is therefore more likely to respond to sales.

Since regular ordinary least squares is likely to fail when incorporating many cross-effects with competing products, and pooling might not be justified, we look into several other methods that allow for heterogeneity in both intercept and slope parameters.

The methods described above belong to the branch of frequentist regressions. In order to find appropriate techniques to incorporate many regressors into the analysis it is beneficial to look into a different spectrum: Bayesian. In Bayesian regression uncertainty is modeled into the model parameters. This leads to correct estimates even when the number of explanatory variables exceeds the number of observations. The main disadvantage of a full Bayesian regression however is that estimation is complicated (Rossi and Allenby, 1993). Therefore we implement several methods closely related to Bayesian regression.

First, shrinkage methods such as the Ridge regression can be considered. The Ridge estimator adds a small constant to the diagonal of the $X'X$ term of the ordinary least squares estimator. This solves possible problems with computing the inverse and has proven to give reliable results. In essence, what the Ridge regression does, is that it shrinks the estimate towards zero. The main advantage of Ridge is that it controls for the instability associated with the ordinary least squares(Arthur E. Hoerl, 1970). This way allowing for more regressors.

Shrinking the estimates towards zero might intuitively not make sense, we there-

fore introduce a slight alteration of the Ridge estimate that pools towards a common effect instead, this will be referred to as *Pooled Ridge*.

Lastly we introduce a similar method called Frequentist Bayes which additionally incorporates the correlations of the common effect.

In this paper we investigate the predictive performance of Pooled Ridge and Frequentist Bayes. The methods are compared to individual ordinary least squares, pooling and a base-case which simply predicts the mean sales. To do this, a dataset on sales of orange juice is used. This dataset is described in Section 2. For the prediction of weekly sales the Frequentist Bayes method turns out to be highly instable and performs worst overall. Pooled Ridge on the other hand gives very promising results and seems to slightly improve the traditional Ridge estimator.

After the data description the aforementioned methods and their evaluation are thoroughly explained in Section 3. Then the results are represented in Section 4 after which they are further interpreted and discussed in Section 5.

2 Data

The data[1] used in this paper is on weekly sales of orange juice from a Dominicks Finer Food Chain store over a period of almost two years (Wedel and Zhang, 2004).

We use a sample consisting of the first 66 observations. This sample is equally divided into an estimation sample and hold-out sample.

In total seven different brands are included. The juice can be one of three categories: frozen, refrigerated or shelf stable. Besides that the size can vary from 6 ounces to 128. This information is summarized in Table 1. The unique combination of brand, category and size constitute to a total of 21 different products. The set containing these products is denoted by $\boldsymbol{P}$.

Table 1: Overview of Unique Products

Nr. of	Brand	Content[a](oz) and Category[b]							
Products	Name	6	10	12	16	46	64	96	128
2	FloridaGold	-	-	F	-	-	R	-	-
6	Minute Maid	F	F	F	F	-	R	R	-
5	Tropicana	-	-	F	F	S	R	R	-
5	Dominick's	F	-	F	F	-	R	-	R
1	Florida's Natural	-	-	-	-	-	R	-	-
1	Hi C	-	-	-	-	-	S	-	-
1	Gatorade	-	-	-	-	-	S	-	-
Total 21		(F)	(F)	(F)	(F)	(S)	(R,S)	(R)	(R)

[a] For frozen Orange juice it is recommended to add three times as much water therefore the comparable size is four times what is reported(Rossi and Allenby, 1993).
[b] Available categories: F = Frozen, R = Refrigerated, S = Shelf Stable

Lastly information on the price and the presence of in-store non-price promotions of each product is provided. These variables, their possible lags and the sales lags can be part of the set of explanatory variables $\boldsymbol{K}$. Summary statistics on the sales and prices of all products for the different samples is given in Appendices 9 and 10.

A depiction of the sales and price over time for three products is given in Figures 1,2,3. Figure 1 shows a clear interaction between price and sales; when prices are lowered sales increase. Figure 2 does not, due to constant prices for the largest part of the sample. This is unique for this particular product. In the last figure (3) sales also seem to have a one-to-one correspondence with price. Figures for the remaining 18 products can be found in Appendix B.

[1] The data is originally provided by The Kilts Center for Marketing at the Universtiy of Chicago's Marketing Department

Figure 1: Sales FloridaGold Refrigerated 64 ounces

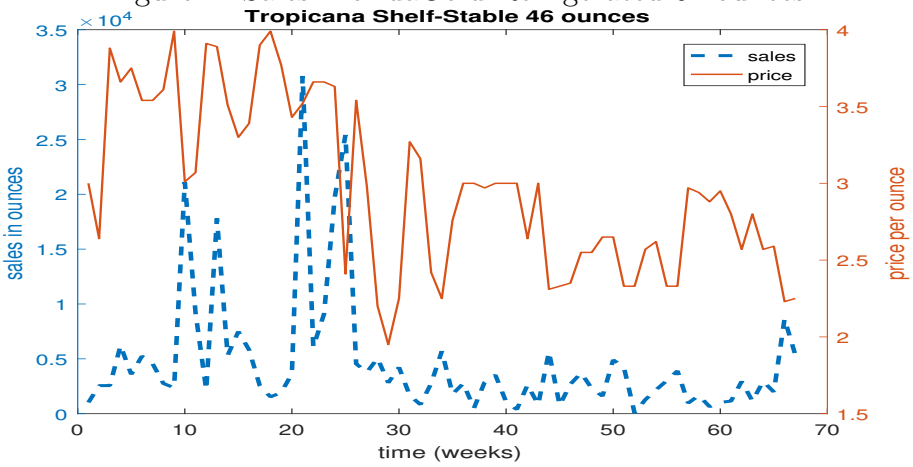


Figure 2: Sales Minute Maid Frozen 6 ounces

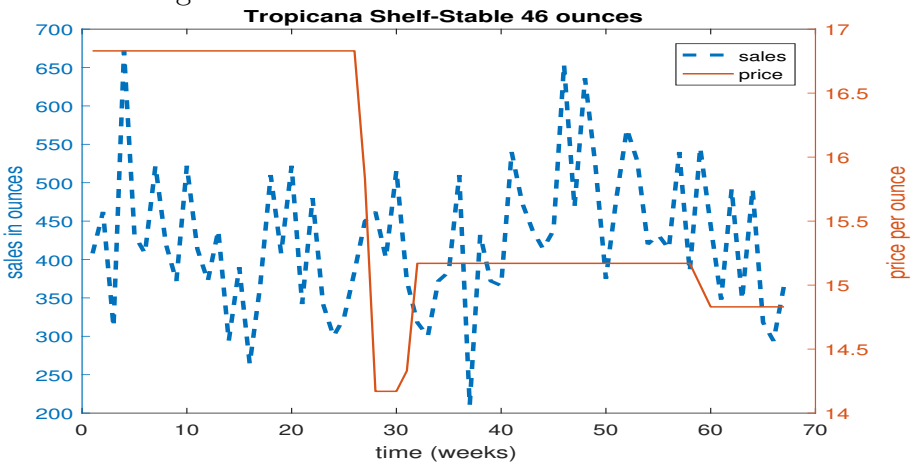
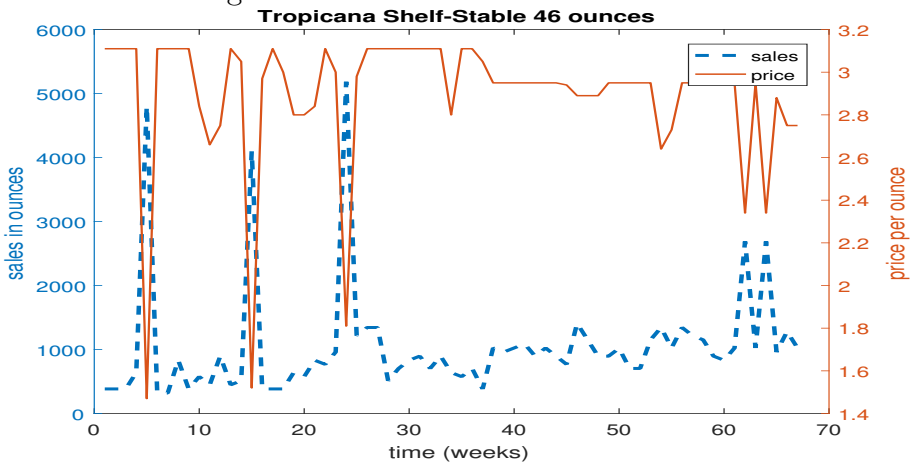


Figure 3: Hi C Shelf-Stable 64 ounces



3 Methods

In this section first the pure frequentist methods are introduced (Section 3.1 and 3.2) which will be the starting point of this analysis. Then a short introduction to Bayesian regression is given after which the shrinkage methods are introduce, but first some general notation and phrases are explained.

In this paper the term products refers to the individual products defined by the unique combination of brand, category and content as described in Section 2.

Vectors are indicated by bold symbols, matrices by capital letters and sets by bold capital letters. Estimates are indicated with a hat on top.

3.1 Individual Ordinary Least Squares

For explanatory purposes it is useful to distinguish between Frequentist and Bayesian regression. To do so it is necessary to briefly explain the basis first: in any regression analysis you start with a set of observations belonging to an individual i for a variable y (represented in a vector $\boldsymbol{y}_i$) that need to be modeled. To do so you have multiple observation sets for other variables (represented as column vectors stacked alongside each other in a matrix X). These are used to explain/model the observations of variable y. For this research y represents sales, X includes a composition of own and competitor's price, sales and promotion variables, and individuals represent the different types of orange juice products.

It is assumed that the data $\boldsymbol{y}$ comes from a specific data generating process (DGP). This is often described using a linear functional form:

$$\boldsymbol{y}_i = X_i\boldsymbol{\beta_i} + \boldsymbol{\varepsilon}_i \qquad \forall i \in \boldsymbol{P} \tag{1}$$

here $\boldsymbol{\beta}_i$ represents the vector of coefficients/parameters characterizing the DGP of product i, and $\boldsymbol{\varepsilon}_i$ the corresponding disturbance term generally assumed to be normally distributed with mean zero and covariance matrix $\sigma_i^2 I$.

Ordinary least squares (OLS) provides an estimate $\hat{\boldsymbol{\beta}}_i$ for each individual i in $\boldsymbol{P}$, which ensures that $\boldsymbol{y}_i$ is modeled with the smallest residual sum of squares. Residuals ($\boldsymbol{e}_i$) are defined by subtracting the fitted values of $\boldsymbol{y}_i$, $\hat{\boldsymbol{y}}_i = X_i\hat{\boldsymbol{\beta}}_i$, from the observed $\boldsymbol{y}_i$. The resulting residual sum of squares is then defined as in (2).

$$\boldsymbol{e}_i'\boldsymbol{e}_i = (\boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}}_i)'(\boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}}_i) \qquad \forall i \in \boldsymbol{P} \tag{2}$$

The corresponding OLS estimate is shown in (3), derivations can be found in the book by Heij et al (2004).

$$\hat{\boldsymbol{\beta}}_i = (X_i'X_i)^{-1}X_i'\boldsymbol{y}_i \qquad \forall i \in \boldsymbol{P} \tag{3}$$

Problems with the OLS estimator often manifest around taking the inverse of $X_i'X_i$.

Firstly when X_i contains highly correlated regressors the inverse of $X_i'X_i$ is likely to become nearly singular resulting in less accurate estimates.

Secondly when (due to this high correlation for example) $X_i'X_i$ is not of full rank the inverse even gets infeasible according to the Invertible Matrix Theorem (Poole, 2014). This often occurs when the amount of explanatory variables (the number of columns in X_i) is large with respect to the amount of observations (the number of rows in X_i).

3.2 Pooling

One possible solution to overcome the inverse problems encountered with the OLS estimate is to use more data. This can be done by estimating one model for all different products at the same time resulting in a pooled estimate. In order to perform such a pooled regression it is necessary to properly structure the explanatory matrix, therefore first the structure of the explanatory matrix X is described. The construction of the explanatory matrix will be consistent across all methods.

Cross-effects between competing products are included in the explanatory matrix in the following way:

$$\begin{aligned} X &= [X_1 \quad ... \quad X_{|\boldsymbol{K}|}] \\ X_j &= [\boldsymbol{x}_{j,1} \quad ... \quad \boldsymbol{x}_{j,p}] \qquad \forall j \in \boldsymbol{K} \end{aligned} \tag{4}$$

where p is the total number of unique products (which equals 21 in this analysis), and $\boldsymbol{x}_{j,i}$ is the vector containing T observations of explanatory variable j for product i. This means that for each product the own-effect is located differently, but therefore all products have identical explanatory matrices which allows us to substitute X_i by X in (1), (2) and (3).

In order to perform a pooled regression that incorporates cross-effects between different competing products, it is necessary to pool over two dimensions: products and categories. Pooling over categories is necessary to insure compatibility of the coefficients. Here compatibility involves that the meaning of the coefficients is equal for all products. This is achieved by declaring a new coefficient vector $\boldsymbol{\delta}$. This vector contains the effects between categories instead of between all products and is therefore smaller than the beta vector. The structure of $\boldsymbol{\delta}$ is as follows:

$$\begin{aligned} \boldsymbol{\delta} &= [\boldsymbol{\delta}_1' \quad ... \quad \boldsymbol{\delta}_{|\boldsymbol{K}|}'] \\ \boldsymbol{\delta}_j &= [\delta_{j,0} \quad \delta_{j,R,R} \quad \delta_{j,R,F} \quad \delta_{j,R,S} \quad \delta_{j,F,R} \quad \delta_{j,F,F} \quad \delta_{j,F,S} \quad \delta_{j,S,R} \quad \delta_{j,S,R} \quad \delta_{j,S,S}] \\ & \forall j \in \boldsymbol{K} \end{aligned} \tag{5}$$

where R,F,S refer to categories refrigerated, frozen and shelf-stable respectively. Here $\delta_{j,0}$ is the estimator for the own-effect of explanatory variable j, $\delta_{j,A,B}$ the cross-effect between category A and B for explanatory variable j. Table 2 gives a schematic description of the entries of $\boldsymbol{\delta}$.

Table 2: Schematic Description of Cross-Effects between Categories

	Explanatory variables (X)		
	Frozen	Refrigerated	Shelf-Stable
Dependent variable (y)			
Frozen	$\delta_{j,F,F}$	$\delta_{j,F,R}$	$\delta_{j,F,S}$
Refrigerated	$\delta_{j,R,F}$	$\delta_{j,R,R}$	$\delta_{j,R,S}$
Shelf-Stable	$\delta_{j,S,F}$	$\delta_{j,S,R}$	$\delta_{j,S,S}$

This table should be read as follows: $\delta_{j,A,B}$ is the cross-effect of explanatory variables j (e.g. price) of a product from category B on a product of category A

The resulting pooled regression is as follows:

$$\boldsymbol{y}_i = XW_i\boldsymbol{\delta} + \boldsymbol{\varepsilon} \qquad \forall i \in \boldsymbol{P} \tag{6}$$

here W_i is the indicator matrix transforming $\boldsymbol{\delta}$ such that it matches the dimensions of X. Each column in W_i is linked to a $\boldsymbol{\delta}_{j,A,B}$ and each row is linked to a product i.

The pooled regression results in the OLS estimate of $\boldsymbol{\delta}$ as follows:

$$\hat{\boldsymbol{\delta}} = \left(\sum_{i\in\boldsymbol{P}} W_i X' X W_i'\right)^{-1} \left(\sum_{i\in\boldsymbol{P}} W_i' X' \boldsymbol{y}_i\right) \tag{7}$$

3.3 Bayesian and Shrinkage Methods

In order to introduce Bayesian regression it is important to highlight the main difference with frequentist regression. Frequentist approaches are characterized by the assumption of data generating process (DGP) as for example in (1). Here it is assumed that the betas are the true parameters, and the disturbance terms account for uncertainty in the data. The aim of frequentist approaches is to get estimates for these betas. The most common methods of frequentist estimation are ordinary least squares (OLS) and maximum likelihood of which the first is described in Section 3.1.

Bayesian regression also uses the DGP, however now you do not assume there exist unique true values for beta. Instead you treat the betas as random variables with a specific distribution. General explanation of how this works (for a single product) is given in the following paragraphs.

Bayesian regression first generates a *prior*: $\pi(\boldsymbol{\beta})$. The prior is the assumed distribution of the betas before obtaining the data. Often a normal distribution is chosen as appropriate representation.

Then the data is used to update the prior distribution and obtain the so called *posterior* (Greenberg, 2008).

$$posterior(\boldsymbol{\beta}|\boldsymbol{y}) = \frac{p(\boldsymbol{\beta}, \boldsymbol{y})}{p(\boldsymbol{y})} \tag{8}$$

The posterior can be rewritten by using Bayes Rule (Greenberg, 2008) as follows:

$$posterior(\boldsymbol{\beta}|\boldsymbol{y}) = \frac{\pi(\boldsymbol{\beta})l(\boldsymbol{y}|\boldsymbol{\beta})}{p(\boldsymbol{y})} \quad \propto \quad \pi(\boldsymbol{\beta})l(\boldsymbol{y}|\boldsymbol{\beta}) \tag{9}$$

here $l(\boldsymbol{y}|\boldsymbol{\beta})$ equals the likelihood function of the data $\boldsymbol{y}$, and $p(\boldsymbol{y})$ can be disregarded from the analysis as a constant.

When the prior is of the same form (yet possibly different parameters) as the posterior it is called a conjugate prior (Greenberg, 2008). If the DGP has a linear functional form as described in (1), and assumes a normal distribution for the disturbances then in order to get a conjugate prior for $\boldsymbol{\beta}$ you need a normal distribution with mean $\boldsymbol{b}$ and covariance $\sigma_i^2 B$.

A common estimator to update $\hat{\boldsymbol{\beta}}$ is the mean of the posterior. For the case as discussed above the formula is given by (10). Derivations of this can be found in the book by Greenberg (2008).

$$\hat{\boldsymbol{\beta}} = (X'X + B^{-1})^{-1}(X'\boldsymbol{y} + B^{-1}b) \tag{10}$$

When fully implemented the Bayesian approach gives reliable estimates for the betas even in small samples. This is because inference is done on the betas conditional on a single observation, and therefore no asymptotic distribution theory is required.

One of the main difficulties in Bayesian regression however, is that the parameters of the prior are difficult to obtain. Next, three methods are discussed, which all choose these parameters differently.

3.3.1 Ridge

As mentioned before the OLS estimator can encounter problems when computing the inverse of $X'X$. One possible solution for this is to add a small constant λ to the diagonal of $X'X$. This results in the definition of the Ridge estimate (Arthur E. Hoerl, 1970) as shown here:

$$\hat{\boldsymbol{\beta}}_i = (X'X + \lambda I)^{-1}X'\boldsymbol{y_i} \qquad \forall i \in \boldsymbol{P} \tag{11}$$

This formula can be derived from the Bayesian estimate (10) by setting the parameters B equal to the identity matrix multiplied by λ, and $\boldsymbol{b}$ by zero. A common interpretation of Ridge is to say that it shrinks the estimates towards zero. This becomes clear when looking at the corresponding residual sum of squares (Gruber, 1998) as shown here:

$$\boldsymbol{e}_i'\boldsymbol{e}_i = (\boldsymbol{y}_i - X\hat{\boldsymbol{\beta}}_i)'(\boldsymbol{y}_i - X\hat{\boldsymbol{\beta}}_i) + \lambda\hat{\boldsymbol{\beta}}_i'\hat{\boldsymbol{\beta}}_i \qquad \forall i \in \boldsymbol{P} \tag{12}$$

When minimizing $\boldsymbol{e}_i'\boldsymbol{e}_i$ for some product i two components are involved. Firstly the fitted value of $\boldsymbol{y}_i$ is driven towards the real value of $\boldsymbol{y}_i$, just as in the OLS estimator. In ridge however $\lambda\boldsymbol{\beta}_i'\boldsymbol{\beta}_i$ is minimized at the same time. Since λ is a predefined constant this results into a shrinkage of $\boldsymbol{\beta}_i$ towards zero.

3.3.2 Pooled Ridge

Since the shrinkage to zero used in Ridge might not give the best intuitive explanation, we propose to shrink the estimate towards a common effect instead. This is done by setting the parameter $\boldsymbol{b}$ of the Bayesian estimator equal to the pooled estimate. For each product i in $\boldsymbol{P}$ the pooled estimate is obtained by multiplying the indicator matrix with the estimate of delta: $W_i\hat{\boldsymbol{\delta}}$. This results in the following:

$$\hat{\boldsymbol{\beta}}_i = (X'X + \lambda I)^{-1}\left(X'\boldsymbol{y_i} + \lambda W_i\hat{\boldsymbol{\delta}}\right) \qquad \forall i \in \boldsymbol{P} \tag{13}$$

In order to show how this method forces the estimates towards the pooled value the residual sum of squares is presented:

$$\boldsymbol{e}_i'\boldsymbol{e}_i = (\boldsymbol{y}_i - X\hat{\boldsymbol{\beta}}_i)'(\boldsymbol{y}_i - X\hat{\boldsymbol{\beta}}_i) + \lambda(\hat{\boldsymbol{\beta}}_i - W_i\boldsymbol{\delta}_i)'(\hat{\boldsymbol{\beta}}_i - W_i\boldsymbol{\delta}_i) \qquad \forall i \in \boldsymbol{P} \tag{14}$$

In order to minimize the λ-term, $\hat{\boldsymbol{\beta}}_i$ needs to be as close as possible to $W_i\boldsymbol{\delta}$ resulting in a shrinkage towards the pooled estimates.

3.3.3 Frequentist Bayes

The last shrinkage method this paper introduces is *Frequentist Bayes*, which is inspired by the research of Rossi and Allenby (1993), who propose to construct the Bayesian prior from a pooled estimation. Besides determining the mean of the prior ($\boldsymbol{b}$) by pooling (as in Pooled Ridge), also the covariance matrix ($\boldsymbol{B}$) is chosen based on the pooled estimation. So again the estimate of Bayes (10) is used, where $\boldsymbol{b}$ is set equal to the extended delta estimator. However, now $\boldsymbol{B}^{-1}$ is not merely a diagonal matrix. Instead we set $\boldsymbol{B}^{-1}$ equal to the extended inverted covariance matrix of the delta estimator multiplied by λ. This is defined as follows:

$$\boldsymbol{B}^{-1} = \lambda W_i Cov(\hat{\boldsymbol{\delta}})^{-1}W_i' = \lambda W_i'\left(\hat{\sigma}^2\sum_{j\in\boldsymbol{P}} W_j'X'XW_j\right)^{-1}W_i \tag{15}$$

where $\hat{\sigma}^2$ is the estimate of the variance of the disturbance term (Heij et al., 2004):

$$\hat{\sigma}^2 = \frac{1}{n^2 - |\boldsymbol{\delta}|}\sum_{i\in\boldsymbol{P}}\left(\boldsymbol{y}_i - XW_i\hat{\boldsymbol{\delta}}\right) \tag{16}$$

here n is the total number of observations for all individuals together.

The final estimate for Frequentist Bayes is now defined as follows:

$$\hat{\boldsymbol{\beta}}_i = (X'X + \lambda W_i Cov(\hat{\boldsymbol{\delta}})^{-1}W_i')^{-1}(X'\boldsymbol{y} + \lambda W_i Cov(\hat{\boldsymbol{\delta}})^{-1}W_i'W_i\hat{\boldsymbol{\delta}}) \qquad \forall i \in \boldsymbol{P} \tag{17}$$

3.4 Evaluation

In this section the evaluation of the different methods is explained in detail. For all methods predictive performance is measured using the root mean squared prediction error for each product (RMSPE). The RMSPE for each product i in $\boldsymbol{P}$ is defined as follows:

$$RMSPE_i = \frac{1}{|\boldsymbol{H}|} \sum_{t \in \boldsymbol{H}} (y_{i,t} - \hat{y}_{i,t})^2 \tag{18}$$

here $y_{i,t}$ corresponds to observation t of product i and $\boldsymbol{H}$ refers to the set of observations in the hold-out-sample.

The different methods are evaluated based on one store. The summary statistic for this store are displayed in Section 2 and the Appendix.

All shrinkage methods depend on the λ-term, which indicates to which extend the pooled data contributes to the estimate. Since the pooled data contains 21 times as many observations as the individual data, a reasonable value for λ would be $\frac{1}{21}$. This theoretically ensures that the individual and the pooled data count with the same weight. We also test the methods for higher and lower values of λ since it is not necessarily optimal to weigh the pooled and individual data equally.

As a base-case prediction we predict all sales by the average sales of the estimation-sample.

4 Results

The results of the before discussed regression methods are displayed in Figure 3 and 4. Here only current price of all products are used as explanatory variables. Including more explanatory variables led to instable results for all shrinkage methods.

4.1 Out-of-Sample Performance

OLS

The ordinary least squares estimate performs worst than the base-case prediction for 18 out of 21 products.

Pooled

The RMSPEs of the pooled estimates are lower for 14 out of 21 products with respect to the base-case prediction. On average the pooled estimates improve the RMSPEs by 1,280 per product.

Ridge

For most products the best results for Ridge are obtained for λ equal to $\frac{1}{21}$. For this value of λ, 18 out of 21 products are estimated more accurately than the base-case, however only for 11 products Ridge outperformed the simple pooling method.

Pooled Ridge
On average Pooled Ridge obtained the lowest RMSPEs for λ equal to 1. Here for 18 out of 21 products Pooled Ridge outperforms the base-case prediction and for 17 products it also predicted better than Ridge.

Frequentist Bayes
After OLS the Frequentist Bayes method performs the worst leading to RMSPEs higher than the base-case prediction for (more than) half of the products for all values of λ.

Table 3: Out-of-Sample Results OLS and Pooled

Product Number[a]	Mean	OLS	Pooled
1	0.29	4.16	0.55
2	4.58	4.77	4.39
3	0.20	0.36	0.23
4	7.09	8.29	5.72
5	1.24	1.44	1.31
6	2.83	9.14	2.36
7	0.67	1.96	0.67
8	0.40	0.55	0.38
9	0.40	0.82	0.33
10	0.01	0.02	0.01
11	0.04	0.04	0.02
12	1.44	1.66	1.06
13	0.04	0.05	0.05
14	1.17	0.72	0.97
15	0.06	3.48	0.04
16	0.14	0.64	0.07
17	0.63	1.17E+02	0.42
18	0.12	0.16	0.13
19	0.06	0.05	0.05
20	0.09	0.09	0.06
21	0.14	0.20	0.12

This table shows the RMSPE(1E+4) in sample S2
[a] Product numbering can be found in Appendix A

Compared to OLS the basic pooling method and Ridge turn out to give good predictive results. However Pooled Ridge seems to improve the predictive quality even more. Frequentist Bayes turns out to perform worst of all shrinkage methods.

Table 4: Out-of-Sample Results Shrinkage Methods

	Ridge			Pooled Ridge			Frequentist Bayes		
λ	1E-4	$\frac{1}{21}$	1	1E-4	$\frac{1}{21}$	1	1E-4	$\frac{1}{21}$	1
Product Number[a]									
1	1.38	0.16	0.17	1.38	0.16	0.17	0.36	0.21	0.21
2	4.73	4.55	4.63	4.73	4.44	4.36	4.51	5.21	5.24
3	0.30	0.19	0.20	0.30	0.21	0.21	0.29	0.39	0.39
4	10.00	3.38	6.33	10.07	5.76	4.12	10.49	7.71	7.74
5	1.41	1.49	1.42	1.40	1.29	1.27	1.24	1.31	1.31
6	9.07	2.13	2.60	9.07	2.10	2.37	5.22	11.14	10.87
7	1.70	0.72	0.65	1.70	0.83	0.67	1.59	0.97	0.96
8	0.48	0.38	0.39	0.48	0.30	0.32	0.27	0.26	0.26
9	0.86	0.32	0.32	0.86	0.33	0.28	0.64	0.56	0.58
10	0.02	0.01	0.01	0.02	0.01	0.01	0.02	0.01	0.01
11	0.04	0.03	0.03	0.04	0.03	0.03	0.04	0.04	0.04
12	1.53	1.17	1.19	1.53	1.16	1.09	1.33	1.39	1.39
13	0.04	0.03	0.03	0.04	0.03	0.03	0.03	0.03	0.03
14	0.64	0.55	0.89	0.64	0.53	0.72	0.64	1.13	1.23
15	1.22	0.05	0.04	1.22	0.05	0.04	0.24	0.15	0.15
16	0.38	0.05	0.11	0.38	0.05	0.07	0.06	0.05	0.05
17	8.78	0.43	0.50	8.79	0.43	0.41	1.40	1.44	1.47
18	0.14	0.10	0.12	0.14	0.10	0.10	0.13	0.12	0.12
19	0.05	0.06	0.06	0.05	0.05	0.05	0.05	0.06	0.06
20	0.08	0.08	0.06	0.08	0.08	0.07	0.08	0.09	0.09
21	0.15	0.13	0.12	0.15	0.12	0.10	0.14	0.14	0.14

This table shows the RMSPE(1E+4) in sample S2
[a] Product numbering can be found in Appendix A

4.2 In-Sample Fit

Insight into the relation between the in- and out-of-sample performance of the methods is informative because good in-sample fit combined with bad out-of-sample performance can indicate that a method has a problem with over-fitting. This means that the model is tailored too specifically to the values of the explanatory variables in the estimation sample. Consequently, this results in bad predictions for slightly different values of the explanatory variables.

The results of the fit in the estimation sample for all methods are given in Tables 5 and 6.
Here we see that the OLS estimates give a better in-sample fit than the base-case prediction. On average it decreases the MSE's by 1,699. Also for 18 out of 21 products it has a lower RMSE than the pooled estimates.

Frequentist Bayes performs slightly better for products 6 and 16 for λ equal to $\frac{1}{21}$ and 1. However this does not compare to the decrease in fit for all other products.

None of the other shrinkage methods have a better in-sample fit than OLS.

The results of the in-sample fit of the different methods show a completely different picture than the out-of-sample behavior. OLS was the worst method for predicting, but turned out to be the best method for in-sample fitting. This indicates that OLS struggles with over-fitting.

Table 5: In-of-Sample Results OLS and Pooled

Product Number[a]	Mean	OLS	Pooled
1	0.76	0.45	0.81
2	3.80	1.39	3.21
3	0.14	0.06	0.14
4	4.27	2.24	3.93
5	0.51	0.16	0.33
6	4.34	1.60	3.34
7	0.41	0.12	0.28
8	0.26	0.11	0.24
9	0.70	0.11	0.54
10	0.01	0.00	0.01
11	0.03	0.01	0.01
12	0.99	0.12	0.49
13	0.03	0.01	0.04
14	0.72	0.20	0.52
15	0.05	0.01	0.03
16	0.33	0.03	0.17
17	0.75	0.25	0.55
18	0.06	0.02	0.06
19	0.12	0.03	0.06
20	0.17	0.04	0.14
21	0.13	0.04	0.14

This table show the RMSPEs(1E+4)
[a] Product numbering can be found in Appendix A

Table 6: In-Sample Results Shrinkage Methods

	Ridge			Pooled Ridge			Frequentist Bayes		
λ	1E-4	$\frac{1}{21}$	1	1E-4	$\frac{1}{21}$	1	1E-4	$\frac{1}{21}$	1
Product Number[a]									
1	0.46	0.53	0.66	0.46	0.53	0.62	0.48	0.50	0.50
2	1.41	2.00	3.48	1.41	1.81	2.91	1.60	2.38	2.40
3	0.06	0.07	0.10	0.06	0.07	0.11	0.07	0.08	0.08
4	2.25	2.54	3.67	2.25	2.41	3.19	2.31	2.57	2.58
5	0.16	0.24	0.35	0.16	0.21	0.28	0.17	0.21	0.22
6	1.60	2.29	3.40	1.60	2.27	2.96	1.67	1.44	1.45
7	0.12	0.19	0.34	0.12	0.17	0.23	0.12	0.17	0.18
8	0.11	0.13	0.21	0.11	0.13	0.18	0.11	0.12	0.12
9	0.12	0.24	0.54	0.12	0.22	0.41	0.22	0.41	0.43
10	0.00	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.01
11	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.01	0.01
12	0.15	0.26	0.56	0.15	0.25	0.38	0.21	0.23	0.23
13	0.01	0.01	0.02	0.01	0.01	0.02	0.01	0.01	0.01
14	0.21	0.20	0.53	0.21	0.19	0.37	0.22	0.28	0.30
15	0.02	0.02	0.03	0.02	0.02	0.02	0.02	0.02	0.02
16	0.03	0.10	0.24	0.03	0.08	0.13	0.02	0.03	0.03
17	0.30	0.43	0.52	0.30	0.44	0.51	0.32	0.32	0.33
18	0.02	0.03	0.05	0.02	0.03	0.04	0.02	0.03	0.03
19	0.03	0.05	0.09	0.03	0.05	0.05	0.04	0.04	0.04
20	0.04	0.09	0.15	0.04	0.08	0.12	0.05	0.07	0.07
21	0.04	0.06	0.08	0.04	0.06	0.09	0.04	0.06	0.06

This table show the RMSPEs(1E+4)

[a] Product numbering can be found in Appendix A

4.3 Coefficients

In order to investigate the difference between the in- and out-of-sample performances of the methods it is useful to look at the coefficients. Table 7 shows the coefficients of price effects for Product 1 (Appendix A). Here values for λ were chosen based on the lowest RMSPE in the hold-out-sample.

Table 7 shows the betas of the different methods side by side.

The pooled betas are closest to zero. This is likely due to the alternating signs of the effects of the different products. The pooled estimate averages all these effects resulting in very conservative betas.

The betas for Pooled Ridge are most similar to the pooled betas. The mean absolute difference equals 0.61. One of the reasons for this are the values for lambda. For more than half of the products a value of λ larger than $\frac{1}{21}$ is optimal. For those products the pooled data counts more heavily than the individual product data. However the estimates still show differences. For example five out of 21 betas have a different sign compared to the pooled betas.

The absolute values of the OLS coefficients are very large. This is likely due to over-fitting, since the out-of-sample fit of OLS was very low compared to the in-sample fit. Reason for this is that the number of observations (which equaled 33 in all samples) was very small with respect to the number of explanatory variables (of which there were 63), making the inverse of $X'X$ very unstable. The own price effect of the OLS estimate is positive, this is intuitively difficult to explain and probably contributes to the high RMSPE for the hold-out-sample

Lastly the coefficient estimates which are high in absolute terms for OLS are also high for Frequentist Bayes. For many products the optimal value of lambda is lower than $\frac{1}{21}$ causing the estimates of Frequentist Bayes to shrink more towards the OLS estimates.

Table 7: Betas for FloridaGold Refrigerated 64 Ounces

	OLS	Pooled	FB[a]		Ridge		PR[b]	
	$\boldsymbol{\beta}_i$	$\boldsymbol{\beta}_i$	$\boldsymbol{\beta}_i$	λ	$\boldsymbol{\beta}_i$	λ	$\boldsymbol{\beta}_i$	λ
OPE[c]Product 1[d]	1.71	-3.49	-3.5	1	-1.93	1/21	-2.63	1/21
CPE[e]Product 1-2	-6.18	0.13	-4.1	1E-4	-3.34	1/21	-0.51	1
CPE Product 1-3	1.51	0.13	-0.8	1E-4	-0.84	1/21	-1.18	1/21
CPE Product 1-4	-7.09	0.13	-4.1	1/21	-0.23	1/21	0.43	1
CPE Product 1-5	18.23	0.13	7.2	1E-4	1.93	1/21	0.30	1
CPE Product 1-6	-7.81	0.13	-4.2	1/21	-1.86	1/21	-1.78	1/21
CPE Product 1-7	-3.39	0.13	-1.2	1	-0.46	1	-0.13	1
CPE Product 1-8	1.48	0.13	2.9	1E-4	0.86	1/21	0.79	1/21
CPE Product 1-9	-0.69	0.07	0.5	1E-4	0.16	1	0.30	1
CPE Product 1-10	-139.53	0.07	-7.6	1	-0.04	1	0.04	1
CPE Product 1-11	4.39	0.07	-1.7	1	-0.16	1/21	0.17	1
CPE Product 1-12	-5.85	0.07	-3.1	1E-4	0.82	1/21	0.26	1
CPE Product 1-13	-7.82	0.07	4.9	1	0.00	1	1.04	1/21
CPE Product 1-14	3.92	0.07	4.0	1E-4	1.38	1/21	1.52	1/21
CPE Product 1-15	-18.27	0.07	0.4	1	0.16	1	0.08	1
CPE Product 1-16	5.34	0.07	4.0	1/21	0.53	1/21	0.61	1/21
CPE Product 1-17	3.63	0.07	2.2	1E-4	-0.29	1/21	-0.09	1
CPE Product 1-18	2.29	0.07	2.2	1	0.43	1/21	0.09	1
CPE Product 1-19	-10.16	-0.46	-8.0	1E-4	-3.20	1E-4	-3.12	1/21
CPE Product 1-20	-18.90	-0.46	-12.5	1E-4	-0.18	1	-0.54	1
CPE Product 1-21	-7.63	-0.46	-9.8	1E-4	-0.22	1	-0.64	1

[a] FB = Frequentist Bayes, [b] PR = Pooled Ridge
[c] OPE = Own Price Effect
[d] Product numbering can be found in Appendix A
[e] CPE Product i-j = Cross Price Effect between product i and j

5 Discussion and Conclusion

In order to predict weekly sales of products like orange juice it is useful to model cross-effects between competing products of variables such as price, sales and promotion. Incorporating these cross-effects results in many explanatory variables which leads to over-fitting when using individual ordinary least squares. The corresponding estimates are of bad predictive quality, caused by instability of the estimates. Therefore we introduced Bayesian inspired shrinkage methods which use pooled (frequentist) regression to establish the parameters of the Bayesian prior distribution.

We started with a simple pooled regression which resolved the instability problems and improved the predictive performance with respect to individual ordinary least squares. The estimated coefficients turn out very close to zero due to the averaging over all different products. This makes it a safe method for handling data with outliers. It generally performed good, overall improving the base-case prediction. Besides that it also improved Ridge for half of the products.

Our newly introduced method Pooled Ridge works similar to Ridge, however instead of shrinking towards zero the estimates are shrunk towards a common effect: the pooled coefficients. Overall this method performed better than Ridge, however not for every single product. The optimal shrinkage term turned out to be higher than $\frac{1}{21}$, leading to more shrinkage towards the value of the pooled regression. The in-sample fit and the predictive performance of Pooled Ridge were both better than Ridge, leading to the conclusion that the individual product data positively influenced the estimates.

Lastly we introduced a method called Frequentist Bayes. This method is similar to Pooled Ridge however, besides incorporating the pooled cross-effects, it also incorporates the corresponding correlations. This is done by choosing a normal distribution for the Bayesian prior and using the pooled estimates and the corresponding covariance matrix as the location and shape parameters respectively. This method turned out to perform very poorly.

The problem encountered in this method is the size of the covariance matrix. In order to model the cross-effects between all different products separately we extended the covariance matrix of the pooled estimates. This resulted in a lot of duplicate data causing the covariance matrix to not reach full-rank.

Theoretically the Frequentist Bayes estimate works as follows: Firstly it shrinks the estimates towards the pooled values, just like Pooled Ridge. Secondly it adds values to the matrix $X'X$ in order to restore the rank, such that you can compute a stable inverse. The main difference between Frequentist Bayes and Pooled ridge however is that, instead of adding a meaningless constant to the diagonal of $X'X$, it adds the pooled covariance matrix to $X'X$. The shrinkage term λ can be interpreted as the extend to which $X'X$ borrows information from the covariance matrix.

So when the covariance matrix is not full rank, using it to complete the rank of $X'X$ does not work. Therefore the question still remains as to how (co)variances should be incorporated such that they improve the predictive quality of the estimate.

Some more general limitations of the introduced methods are now discussed.

Firstly, the choice of λ is very influential on the performance of the shrinkage methods. However a good method to determine λ, ensuring stability of the estimates is difficult to create. Difficulties arise with prediction of new data, because λ is dependent on the size and scale of the data.

Secondly in pure frequentist or Bayesian regression approaches, there is the possibility to subject the estimates to statistical tests. One of the setbacks of combining these two approaches is that it removes this function. The significance of the explanatory variables cannot be formally tested, resulting in a loss of explanatory power. The proposed methods would be merely of practical use.

Lastly, the methods should be tested on different data sets, using different compositions of explanatory variables and varying sample sizes.

Since new products are introduced all the time and available products are continuously renewed, sales data is often scarce. It is of great importance for corporations such as supermarkets to predict these sales accurately. Therefore it is necessary to investigate methods designed to handle small data samples. Overall we can conclude that using frequentist methods to establish parameters of the Bayesian prior can be of great practical value. This paper shows that when modeling the sales of orange juice, cross-effects can be easily incorporated, while limiting the amount of observations necessary. This led to very promising predictive results.

References

Arthur E. Hoerl, R. W. K. (1970). Ridge regression: Applications to nonorthogonal problems. *Technometrics*, 12(1):69–82.

Greenberg, E. (2008). *Introduction to Bayesian Econometrics*. Cambridge University Press.

Gruber, M. (1998). *Improving Efficiency by Shrinkage: The James–Stein and Ridge Regression Estimators*, volume 156. CRC Press.

Heij, C., De Boer, P., Franses, P. H., Kloek, T., Van Dijk, H. K., et al. (2004). *Econometric methods with applications in business and economics*. OUP Oxford.

Poole, D. (2014). *Linear algebra: A modern introduction*. Cengage Learning.

Rossi, P. E. and Allenby, G. M. (1993). A bayesian approach to estimating household parameters. *Journal of Marketing Research*, 30(2):171–182.

Wedel, M. and Zhang, J. (2004). Analyzing brand competition across subcategories. *Journal of Marketing Research*, 41(4):448–456.

A Appendix - Data Statistics

Table 8: Product Information

Product Number	Category[a]	Brand	Content[b](oz)
1	F	FloridaGold	64
2	F	Minute Maid	64
3	F	Minute Maid	96
4	F	Tropicana	64
5	F	Tropicana	96
6	F	Dominick	64
7	F	Dominick	128
8	F	Floridas Natural	64
9	R	FloridaGold	12
10	R	Minute Maid	6
11	R	Minute Maid	10
12	R	Minute Maid	12
13	R	Minute Maid	16
14	R	Tropicana	12
15	R	Tropicana	16
16	R	Dominick	6
17	R	Dominick	12
18	R	Dominick	16
19	S	Hi C	64
20	S	Gatorade	64
21	S	Tropicana	46

[a] Category is as specified in Table 1

[b] For frozen Orange juice it is recommended to add three times as much water therefore the comparable size is four times what is reported(Rossi and Allenby, 1993).

Table 9: Overview of Sales of Orange Juice

Product	Whole Sample		Estimation Sample		Hold-Out-Sample	
Number[a]	mean[b]	SD[c]	mean	SD	mean	SD
1	4.9	5.7	7.0	7.4	2.5	1.8
2	38.8	42.6	36.5	37.6	43.5	43.7
3	8.6	1.6	8.8	1.5	8.6	2.0
4	81.7	52.0	68.1	42.2	93.0	62.9
5	29.3	15.1	25.3	5.1	30.4	11.3
6	29.0	33.0	38.4	40.8	24.4	28.6
7	9.6	5.2	10.1	4.1	10.9	6.6
8	10.3	8.5	9.2	2.6	10.9	3.5
9	3.8	5.6	4.4	6.8	3.5	3.9
10	0.4	0.1	0.4	0.1	0.4	0.1
11	0.5	0.3	0.5	0.3	0.5	0.4
12	12.1	11.7	10.6	9.7	13.2	13.7
13	1.3	0.3	1.5	0.3	1.3	0.3
14	5.8	8.3	5.1	6.9	8.2	10.6
15	1.1	0.4	0.9	0.5	1.2	0.4
16	1.2	2.9	1.2	3.3	0.7	1.4
17	6.1	6.3	6.9	7.3	6.1	6.2
18	1.7	1.0	1.6	0.6	1.7	1.2
19	1.0	0.8	1.0	1.2	1.1	0.5
20	1.3	1.2	1.8	1.7	0.8	0.4
21	3.6	1.5	4.4	1.3	3.5	1.2

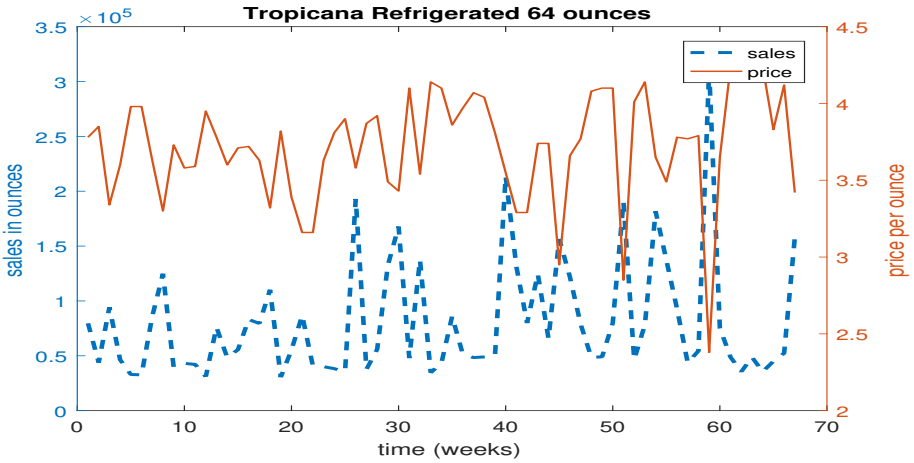
[a] Product Number as specified in A
[b] SD = Standard Deviation
[c] Mean and SD are reported in thousands

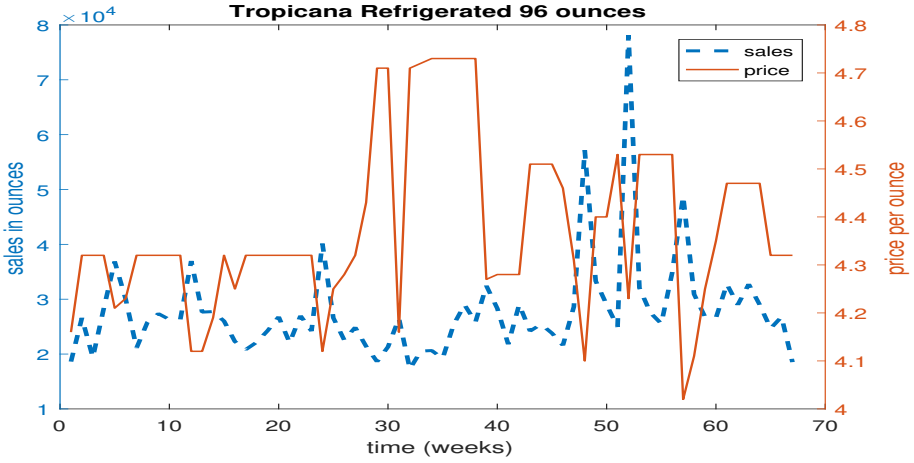
Table 10: Overview of Prices of Orange Juice

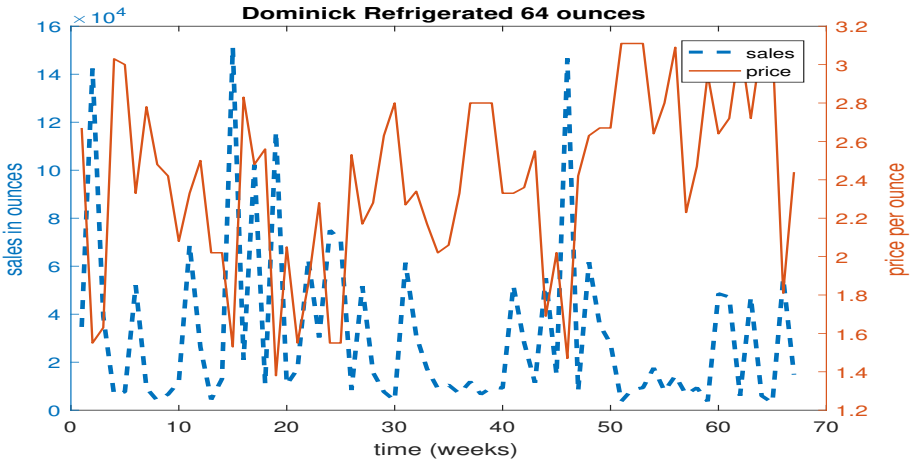
Product	Whole Sample		Estimation Sample		Hold-Out Sample	
Number[a]	mean[a]	SD[b]	mean	SD	mean	SD
1	2.8	0.5	3.3	0.6	2.7	0.3
2	3.7	0.4	3.6	0.2	3.8	0.3
3	3.9	0.3	3.9	0.3	4.1	0.3
4	3.7	0.4	3.7	0.3	3.8	0.4
5	4.2	0.3	4.3	0.2	4.4	0.2
6	2.4	0.5	2.2	0.5	2.6	0.4
7	2.7	0.3	2.6	0.3	2.6	0.3
8	3.8	0.3	3.7	0.2	3.7	0.3
9	11.1	1.8	11.8	2.0	10.5	1.3
10	15.4	0.9	16.3	0.9	15.1	0.1
11	14.9	2.6	15.1	3.2	15.1	2.6
12	12.3	2.1	12.5	2.7	12.3	2.0
13	13.1	1.1	14.3	1.1	12.4	0.3
14	11.4	1.9	12.6	2.1	10.5	1.6
15	12.3	2.0	14.4	2.0	11.1	0.6
16	12.3	3.1	13.9	3.5	11.8	1.5
17	9.8	1.8	10.4	2.2	9.5	1.3
18	10.0	1.9	12.1	1.8	8.9	0.5
19	2.8	0.4	2.9	0.4	2.9	0.2
20	4.1	0.4	3.9	0.5	4.2	0.3
21	5.2	0.3	5.4	0.3	5.3	0.3

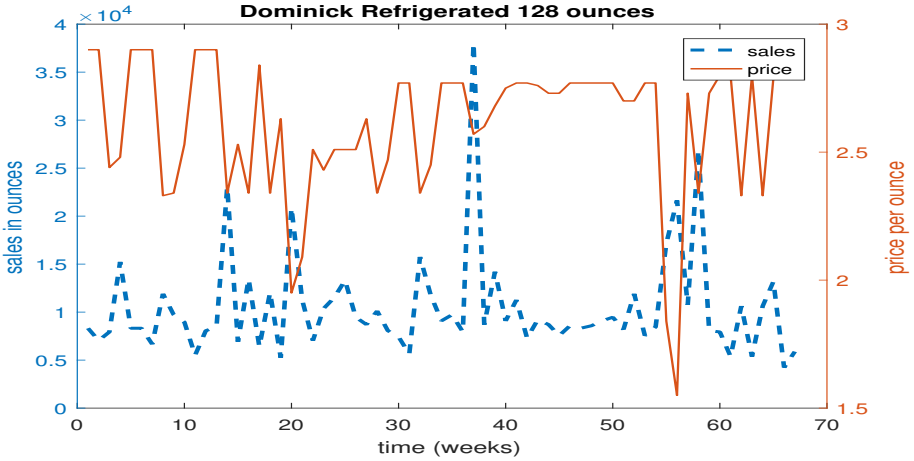
[a] Product Number as specified in A
[b] SD = Standard Deviation
[c] Mean and SD are reported in thousands

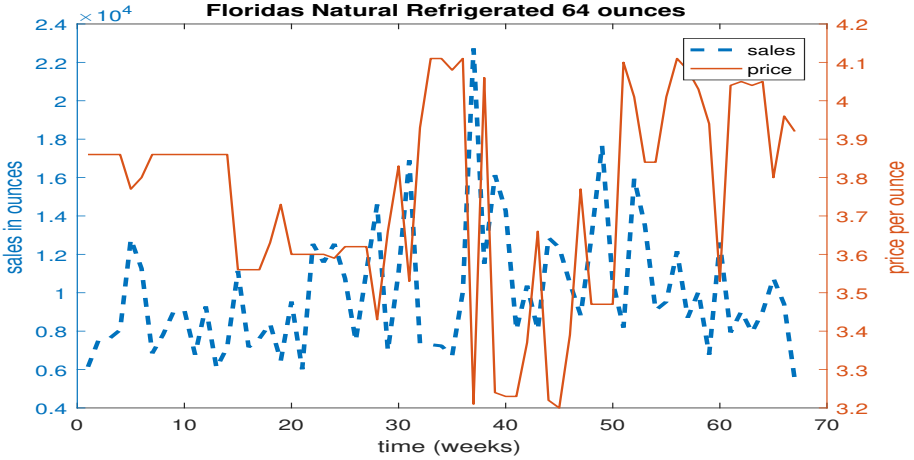
B Appendix - Extra Data Visualization

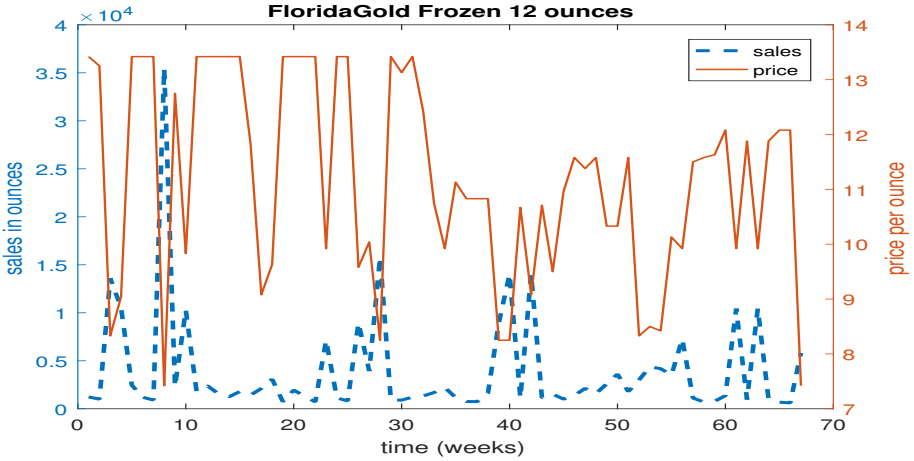


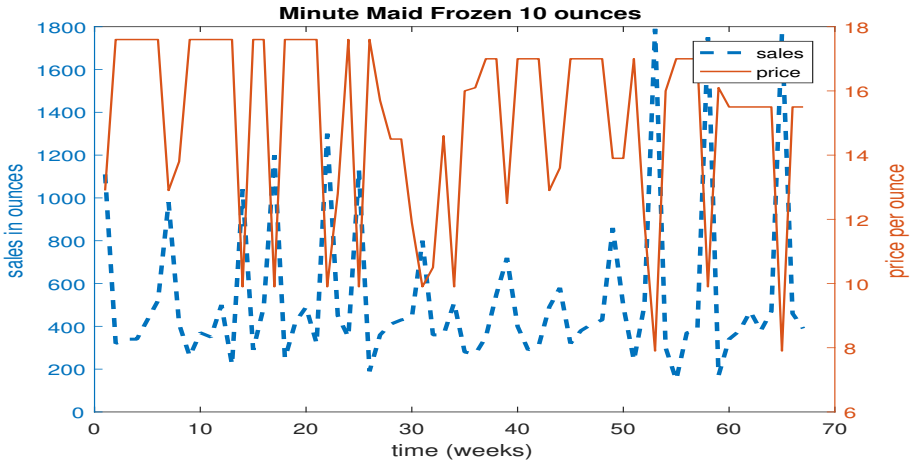


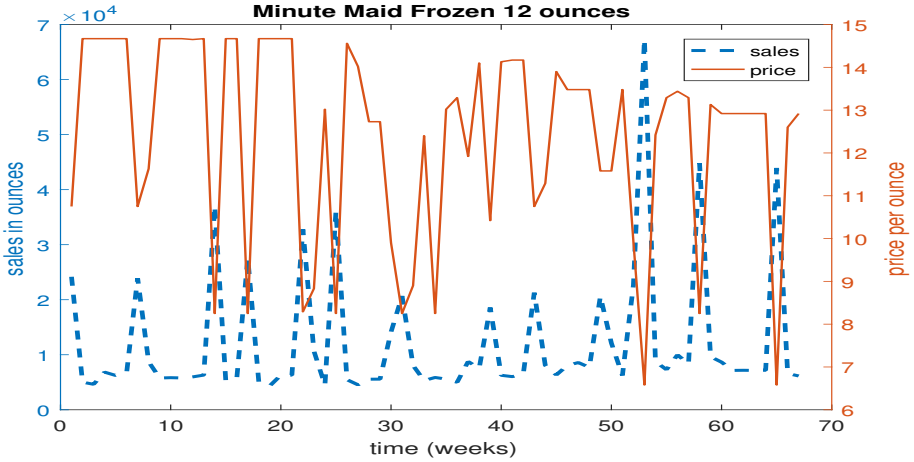


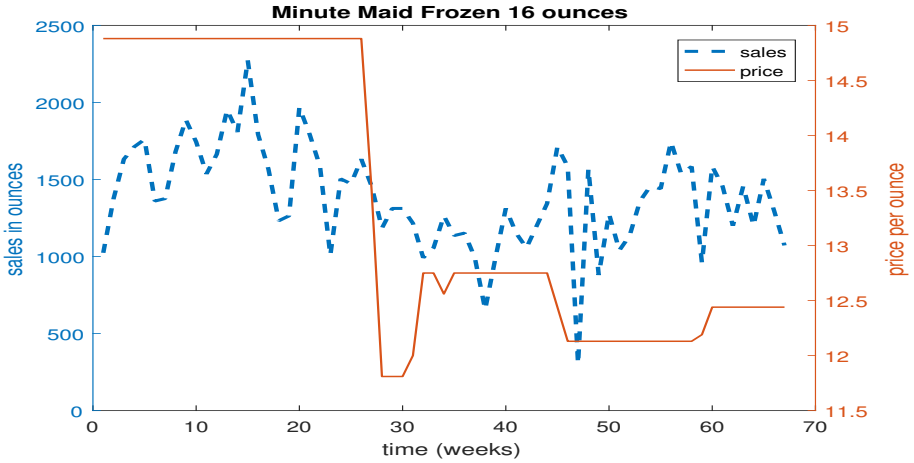

Dominick Refrigerated 128 ounces
×10^4
sales
price
sales in ounces
price per ounce
time (weeks)

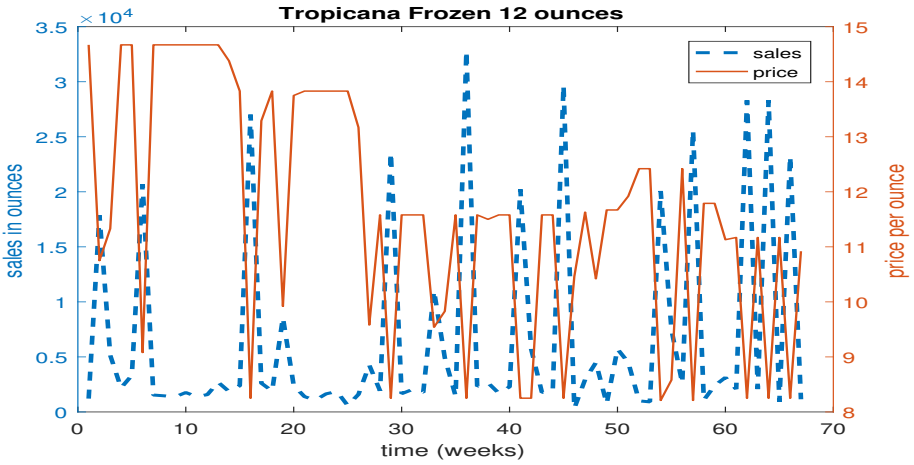

Floridas Natural Refrigerated 64 ounces
×10^4
sales
price
sales in ounces
price per ounce
time (weeks)

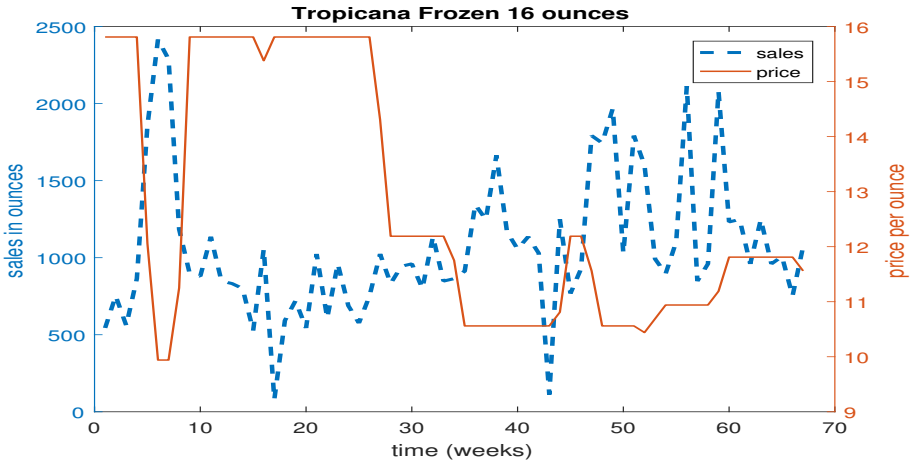

FloridaGold Frozen 12 ounces
×10^4
sales
price
sales in ounces
price per ounce
time (weeks)

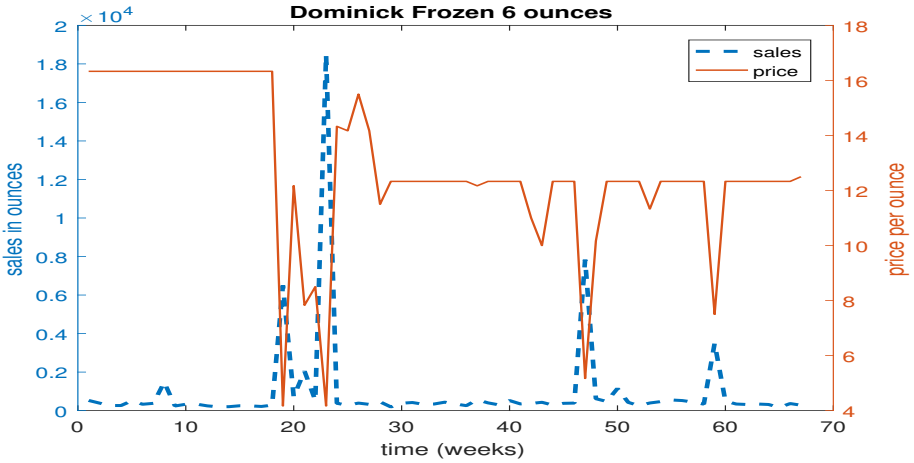

Minute Maid Frozen 10 ounces
sales
price
sales in ounces
price per ounce
time (weeks)

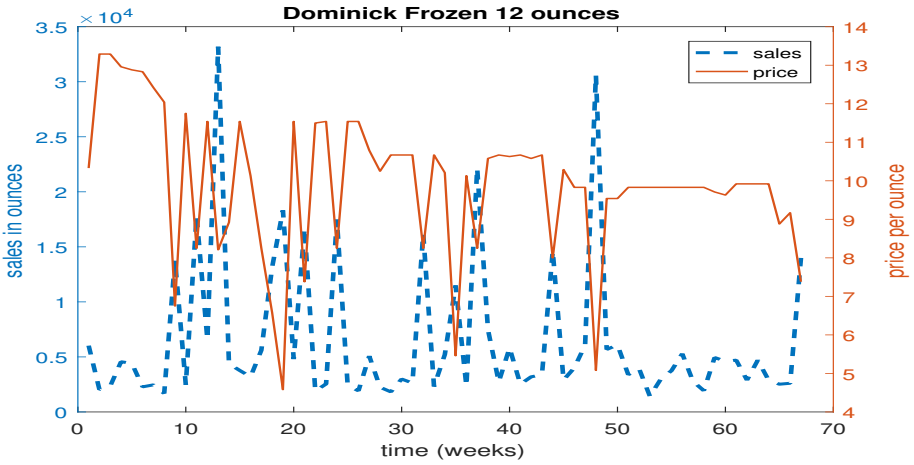

Minute Maid Frozen 12 ounces
×10^4
sales
price
sales in ounces
price per ounce
time (weeks)

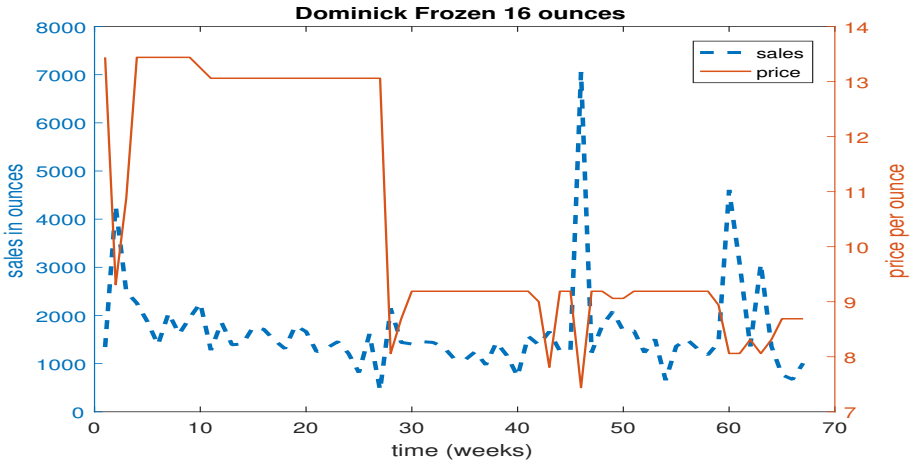

Minute Maid Frozen 16 ounces
sales
price
sales in ounces
price per ounce
time (weeks)

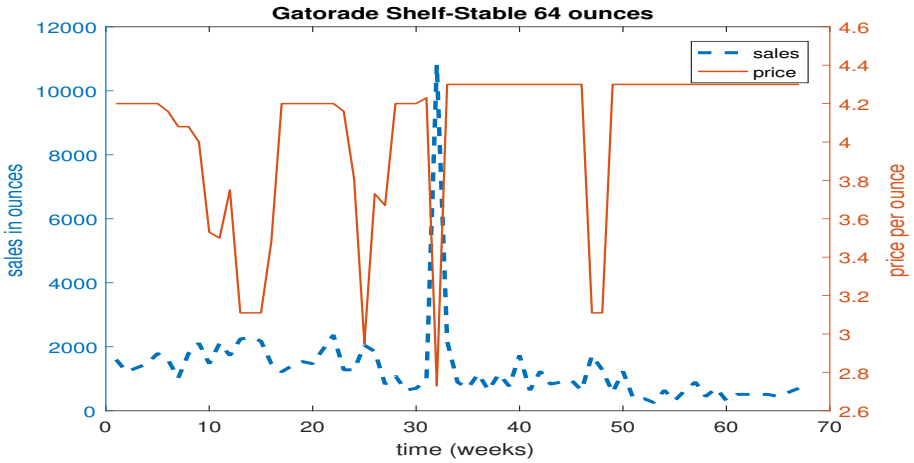

Tropicana Frozen 12 ounces
×10^4
sales
price
sales in ounces
price per ounce
time (weeks)


Tropicana Frozen 16 ounces
sales
price
sales in ounces
price per ounce
time (weeks)


Dominick Frozen 6 ounces
×10^4
sales
price
sales in ounces
price per ounce
time (weeks)


Dominick Frozen 12 ounces
×10^4
sales
price
sales in ounces
price per ounce
time (weeks)


Dominick Frozen 16 ounces
sales
price
sales in ounces
price per ounce
time (weeks)


Gatorade Shelf-Stable 64 ounces
sales
price
sales in ounces
price per ounce
time (weeks)