# Capstone Project: Insight into the Impact of Marketing Activities on Sales Using Marketing Mix Modeling

University of Chicago, Master of Science in Analytics

Oct 1st, 2015

# Project Tracker

## Capstone Project Tracker: Implementation and Research Paper

**Planned Timing for Completion**

| | Task | Status | 29-Jun | 6-Jul | 13-Jul | 20-Jul | 27-Jul | 3-Aug | 10-Aug | 17-Aug | 24-Aug | 31-Aug | 7-Sep | 14-Sep | 21-Sep | 28-Sep |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Meetings | HAVI: On Site | Completed | | | | | | | | | | | | | | |
| | HAVI: Client Check-In | In Progress | | | | | | | | | | | | | | |
| | Anil: Check-In | In Progress | | | | | | | | | | | | | | |
| | Weekly team meeting | | | | | | | | | | | | | | | |
| Modeling | Clean and prepare data | Completed | | | | | | | | | | | | | | |
| | Descriptive Statistics | Completed | | | | | | | | | | | | | | |
| | Log linear models | Completed | | | | | | | | | | | | | | |
| | Hierarchical Linear Models | Completed | | | | | | | | | | | | | | |
| | MCMC Models | Completed | | | | | | | | | | | | | | |
| | Model Validation | In Progress | | | | | | | | | | | | | | |
| | Model Interpetation | Not Completed | | | | | | | | | | | | | | |

# Our Client

## HAVI Global Solutions

**Company Overview**

HAVI Global Solutions LLC provides supply chain management services for the packaging and supply chain industry in the United States and internationally. It develops, sources, markets, and sets up strategic supply chain and packaging services solutions. The company offers packaging, promotions management, and analytics and supply chain services. HAVI Global Solutions LLC was formerly known as Perseco and changed its name to HAVI Global Solutions LLC in 2006. The company was founded in 1975 and is based in Downers Grove, Illinois. HAVI Global Solutions LLC operates as a subsidiary of The HAVI Group, L.P.

3075 Highland Parkway
Downers Grove, IL 60515
United States

**Phone:** 630-493-7400
**Fax:** 630-493-7599
**www.havigs.com**

Founded in 1975

## Consulting & Analytics Services


## Key Clients

Abbott, Microsoft, Coca-Cola, Cargill, ConAgra Foods, McDonald's, Henkel, Darden, Nissan, General Mills, Diageo, Heinz, MillerCoors, bp, Del Monte, JanSport, Tetley, Panda Express, IKEA

# Client’s Initial Project Requirements

THE UNIVERSITY OF CHICAGO
GRAHAM SCHOOL
MScA Community Site

User account My Resume Log out

Home News Search Capstone Project Search Resumes Job search

| | |
|---|---|
| HAVI Global Solutions | HAVI Global Solutions |
| Company Profile | **HAVI Global Solutions (HGS) mission is to be your expert partner in managed services and consulting solutions across the global business supply chain, delivering value and growth from source to consumer. We reach beyond categories, industries, technologies and suppliers to leverage our expertise in an unbiased and impartial approach that is focused solely on your needs. Today some of the world's most respected brands rely on our packaging expertise, marketing and business analytics, supply chain services, promotions management know-how, and recycling and waste solutions to deliver a competitive advantage for their business and an unshakeable respect for their corporate and environmental program impacts.**<br>**Please visit us at** http://www.havigs.com/ |
| Project Title | Promo Mix Modeling |
| Problem Statement | HGS is in the process of building a scalable, promotion (marketing) mix model. Such a set of models will help a key client to be able to plan appropriate product promotions from 30-50 potential products at each of its 170+ markets. There are two key components to make such a set of models operational – a series of statistical exercises to ensure that accuracy of the underlying statistical models, and an user friendly simulator that can fit the latest data into the validated statistical models and allow users to simulate different promotion effects. Potential school project can focus on the first component of model building.<br>While HGS does possess all the POS data for this client that can be used to build such models, the realistic expectation for a Capstone project is for the students to dive into a subset of data that involves a few market. Students will be encouraged to provide analytical solutions to solve the following key issues during statistical model building:<br>• Cannibalization effect across different products<br>• Minimizing data collinearity risks in statistical models<br>• Developing an effective model validation process that efficiently validate models related large number of products<br>• Developing a scalable solution to quickly expand the solution to 170+ markets<br>Please contact Pan Chen at pchne@havigs.com for questions and more details. |

# Executive Summary

**Business Objectives**

- Identify the effects of different marketing and promotion activities of products on the total dollar sales as well as the total volume sales of other products
- Identify opportunities to further optimize promotion strategies across products

**Research Methods**

- Linear Regressions
- Bayesian Hierarchical Linear Models (grouping factor: store)

**Key Findings**

- Adding *log(Discount)* and *Week of year* as new predictors significantly improve the performance of the models
- The strategy of having a model to predict the daily total volume sales of a product in a store has the best predicting performance (GLM6)
- Bayesian hierarchical linear model (MCMC2) using store as the grouping factor has comparative predictive power as the strategy of multiple small models

**Remaining Issues**

- Residuals of both models (GLM6 and MCM2) show seasonal patterns (i.e., autocorrelated)
- Some incorrect coefficients for the variables of product price

# Capstone Project Purpose and Questions/Hypotheses

## Research Purpose

The purpose of the study is to identify the key vehicles of marketing and promotion activities that drive the boost of revenue and profit. By providing new insights into how different marketing and promotion activities across products affect the total revenue, the study can help HGS's client make more informed promotional decisions for driving volume, revenue, and profit. To achieve this, the solutions should be able to:

- Correctly identify and incorporate the cannibalization and halo effect of promotions across different products.
- Minimize data collinearity risks in statistical models.
- Contain a model validation process that efficiently validates models related to large numbers of products.
- Allow for scalability into more markets.

## Research Questions

- What is the effect of different marketing and promotion activities on the total sales and the sales of other items?
- Are there opportunities to further optimize promotion strategies across products?

## Research Hypothesis

- Changes in marketing and promotion activities within one Co-op impact sales differently across products.

# Understanding Data and Exploratory Data Analysis

# Description of Data

| Dependent Variables | Independent Variables | |
|---|---|---|
| ▪ **Total (4 variables)**<br>Daily total sales ($)<br>Daily total transections<br>DrivThru daily total sales ($)<br>DrivThru daily total transections<br><br>▪ **Items (28 variables)**<br>units_total_BF1, units_total_BF2, units_combo_BF2, units_total_BF3 units_total_BF4, units_combo_BF4, units_total_BF5, units_combo_BF5 units_total_BK1, units_combo_BK1, units_total_CK1, units_combo_CK1 units_total_CK2, units_combo_CK2, units_total_DS1, units_total_HB1 units_total_CB1, units_total_CB2, units_total_CB3, units_total_CB4 units_total_DS2, units_total_CB5, units_total_DS3, units_total_CK3 units_combo_CK3, units_total_SI1, units_total_SI2, units_total_SI3 | ▪ **Weather (52 variables)**<br>Weather 07-JAN-2010, Weather 29-JAN-2010, Weather 30-JAN-2010,Weather 06-FEB-2010,Weather 08-FEB-2010 … | Dummy Coding |
| | ▪ **Day of week and holidays (114 variables)**<br>• Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday | Dummy Coding |
| | • New Years Day(Sunday), New Years Day(Tuesday), New Years Day(Wednesday), New Years Day(Friday), New Years Day(Saturday) …… | Dummy Coding |
| | ▪ **National promotions (24 variables)** | Dummy Coding |
| | ▪ **Local promotions (40 variables)** | Dummy Coding |
| | ▪ **Price reduction promotions (15 variables)**<br>• Price Discount for different item<br>• Calculated as log(promoted price / regular price) | Numeric |
| | ▪ **Regular Price of items at store level (15 variables)** | Numeric |
| | ▪ **Promotion Price of items at store level (15 variables)** | Numeric |

# ~ 12.3% of observations were removed from the data set due to containing missing values

# rows before omitting NA = 184156
# rows after omitting NA = 161477

**Removal % ~ 12.3%**

1000-1939
500-999
0-499

| storeID | before omit NA | after omit NA | storeID | before omit NA | after omit NA | storeID | before omit NA | after omit NA | storeID | before omit NA | after omit NA |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1939 | 1914 | 26 | 1939 | 1806 | 51 | 1939 | 1889 | 76 | 1939 | 1867 |
| 2 | 1939 | 1579 | 27 | 1939 | 1897 | 52 | 1939 | 1811 | 77 | 1939 | 1343 |
| 3 | 1939 | 1752 | 28 | 1939 | 1902 | 53 | 1939 | 1763 | 78 | 1939 | 1007 |
| 4 | 1939 | 1550 | 29 | 1939 | 1875 | 54 | 1939 | 1917 | 79 | 1939 | 1794 |
| 5 | 1939 | 1760 | 30 | 1939 | 1840 | 55 | 1939 | 1905 | 80 | 1939 | 1811 |
| 6 | 1939 | 1081 | 31 | 1939 | 1831 | 56 | 1939 | 1889 | 81 | 1939 | 1875 |
| 7 | 1939 | 1698 | 32 | 1939 | 1592 | 57 | 1939 | 1900 | 82 | 1939 | 1832 |
| 8 | 1939 | 1864 | 33 | 1939 | 1599 | 58 | 1939 | 984 | 83 | 1939 | 1837 |
| 9 | 1939 | 1589 | 34 | 1939 | 1681 | 59 | 1939 | 1870 | 84 | 1939 | 1896 |
| 10 | 1939 | 1863 | 35 | 1939 | 1550 | 60 | 1939 | 1877 | 85 | 1939 | 1615 |
| 11 | 1939 | 1880 | 36 | 1939 | 1825 | 61 | 1939 | 1887 | 86 | 1939 | **0** |
| 12 | 1939 | 1894 | 37 | 1939 | 1708 | 62 | 1939 | 1893 | 87 | 1939 | 1912 |
| 13 | 1939 | 1839 | 38 | 1939 | 1721 | 63 | 1939 | 1850 | 88 | 1939 | 1655 |
| 14 | 1939 | 1798 | 39 | 1939 | 1779 | 64 | 1939 | 1909 | 89 | 1939 | 1885 |
| 15 | 1939 | 1628 | 40 | 1939 | 1850 | 65 | 1939 | 1764 | 90 | 1939 | 1873 |
| 16 | 1939 | 1864 | 41 | 1939 | 1679 | 66 | 1939 | 1756 | 91 | 1939 | 1781 |
| 17 | 1939 | 1882 | 42 | 1939 | 1074 | 67 | 1939 | 1674 | 92 | 1939 | 1899 |
| 18 | 1939 | 1869 | 43 | 1939 | 1775 | 68 | 1939 | 1007 | 93 | 1939 | 1822 |
| 19 | 1939 | 1869 | 44 | 1939 | 1906 | 69 | 1939 | **0** | 94 | 1439 | 1162 |
| 20 | 1939 | 998 | 45 | 1939 | 1335 | 70 | 1939 | 1909 | 95 | 1374 | 1352 |
| 21 | 1939 | 1350 | 46 | 1939 | 1817 | 71 | 1939 | 1864 | 96 | 325 | **316** |
| 22 | 1939 | 1810 | 47 | 1939 | 1908 | 72 | 1939 | 1130 | 97 | 367 | **356** |
| 23 | 1939 | 1805 | 48 | 1939 | 1907 | 73 | 1939 | 1869 | 98 | 324 | **319** |
| 24 | 1939 | 1880 | 49 | 1939 | 1884 | 74 | 1939 | 1837 | | | |
| 25 | 1939 | 1594 | 50 | 1939 | 1890 | 75 | 1939 | 1873 | | | |

# 14 (among 28) products have complete price information (regular and promotion price)

| product_name | regular_price | | promo_price | |
|---|---|---|---|---|
| | Min. | Max. | Min. | Max. |
| units_total_BF1 | | | | |
| units_total_BF2 | 0.79 | 1.09 | 0.08 | 1.09 |
| units_combo_BF2 | | | | |
| units_total_BF3 | 1.00 | 1.79 | 0.47 | 1.59 |
| units_total_BF4 | 2.40 | 5.39 | 0.00 | 3.99 |
| units_combo_BF4 | | | | |
| units_total_BF5 | 2.40 | 3.99 | 0.00 | 3.99 |
| units_combo_BF5 | | | | |
| units_total_BK1 | 1.90 | 3.29 | 0.00 | 3.29 |
| units_combo_BK1 | | | | |
| units_total_CK1 | 2.1 | 3.5 | 0.0 | 3.5 |
| units_combo_CK1 | | | | |
| units_total_CK2 | 4.70 | 6.99 | 0.00 | 6.99 |
| units_combo_CK2 | | | | |
| units_total_DS1 | 0.25 | 1.49 | 0.00 | 1.27 |
| units_total_HB1 | 0.95 | 1.50 | 0.15 | 1.50 |
| units_total_CB1 | | | | |
| units_total_CB2 | | | | |
| units_total_CB3 | | | | |
| units_total_CB4 | 1.00 | 1.65 | 0.09 | 1.59 |
| units_total_DS2 | 0.50 | **20.01** | 0.04 | 1.03 |
| units_total_CB5 | 1.59 | 3.19 | 0.00 | 2.94 |
| units_total_DS3 | 1.99 | 2.89 | 0.00 | 2.89 |
| units_total_CK3 | 3.99 | 6.19 | 0.00 | 6.19 |
| units_combo_CK3 | | | | |
| units_total_SI1 | | | | |
| units_total_SI2 | | | | |
| units_total_SI3 | | | | |

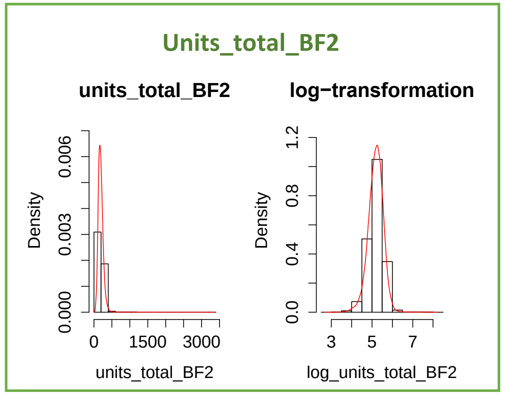
Pie (store ID=7)

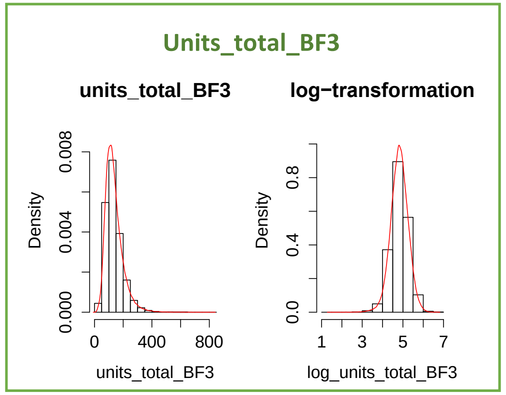
- Among 28 products, 14 products have complete price information in the data set: units_total_BF2, units_total_BF3, units_total_BF4, units_total_BF5, units_total_BK1, units_total_CK1, units_total_CK2, units_total_DS1, units_total_HB1, units_total_CB4, units_total_DS2, units_total_CB5, units_total_DS3, units_total_CK3

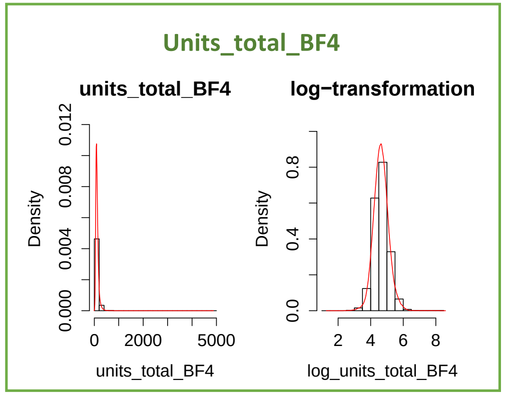
  **We will model the daily unit sales of these 14 products first**

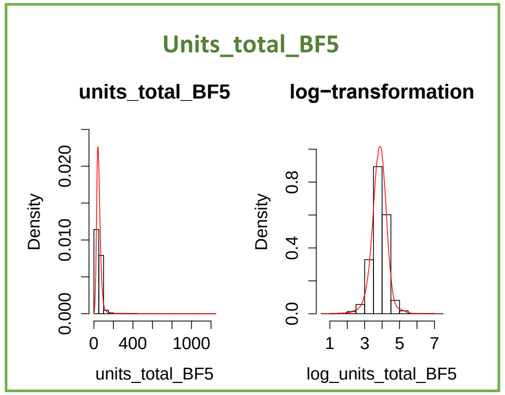
- The range of the regular price of units_total_DS2 (which is product of pie) was $0.50-$20.0. This abnormal observation was mainly from storeID=7.

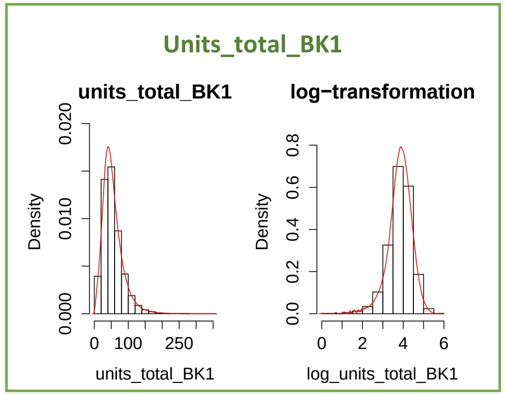
# Log-transformation was applied to the 14 dependent variables for the predictions

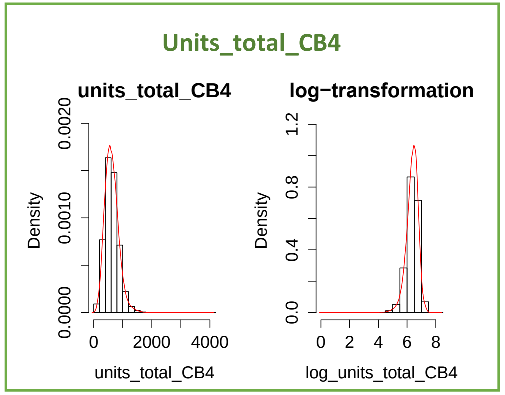


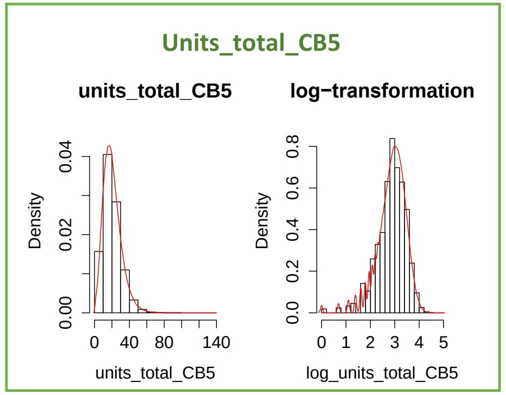



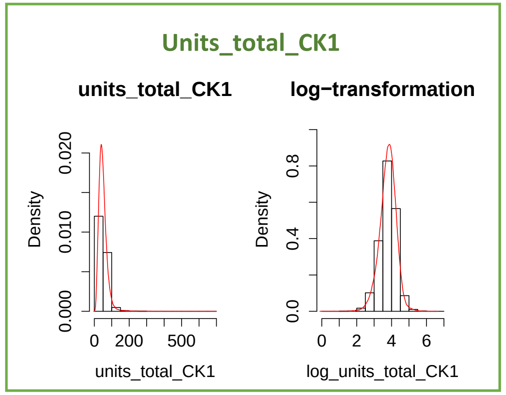



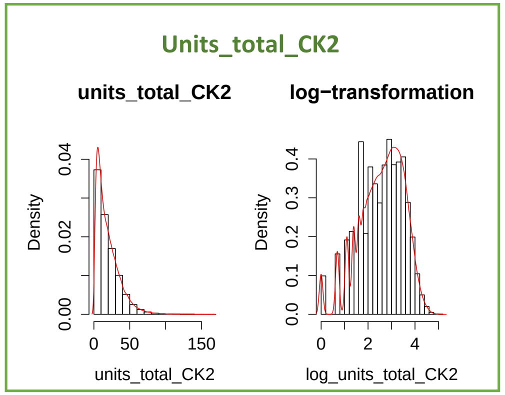



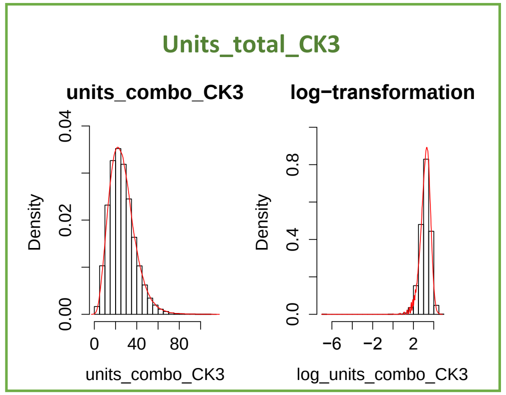



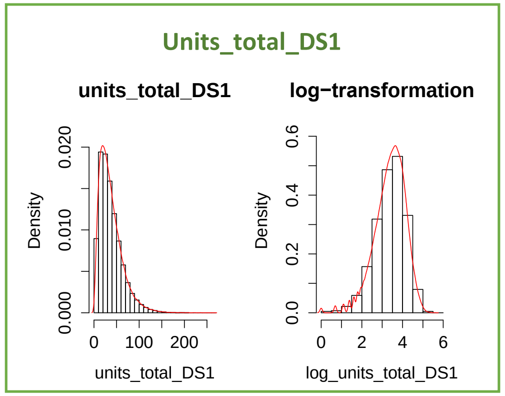



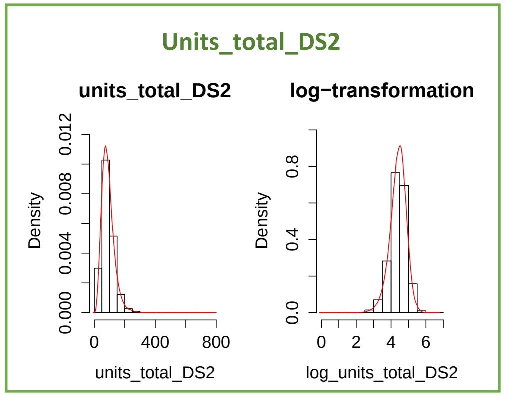



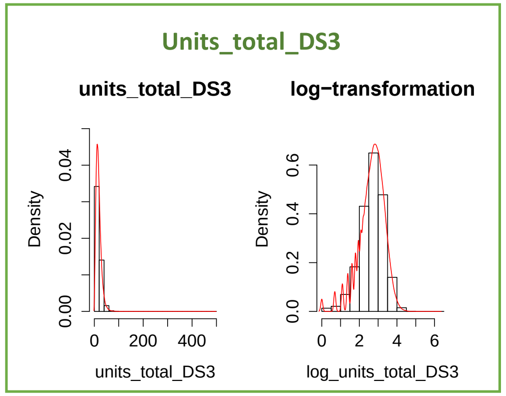



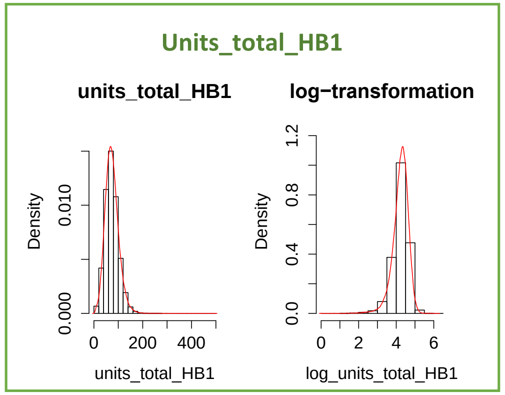



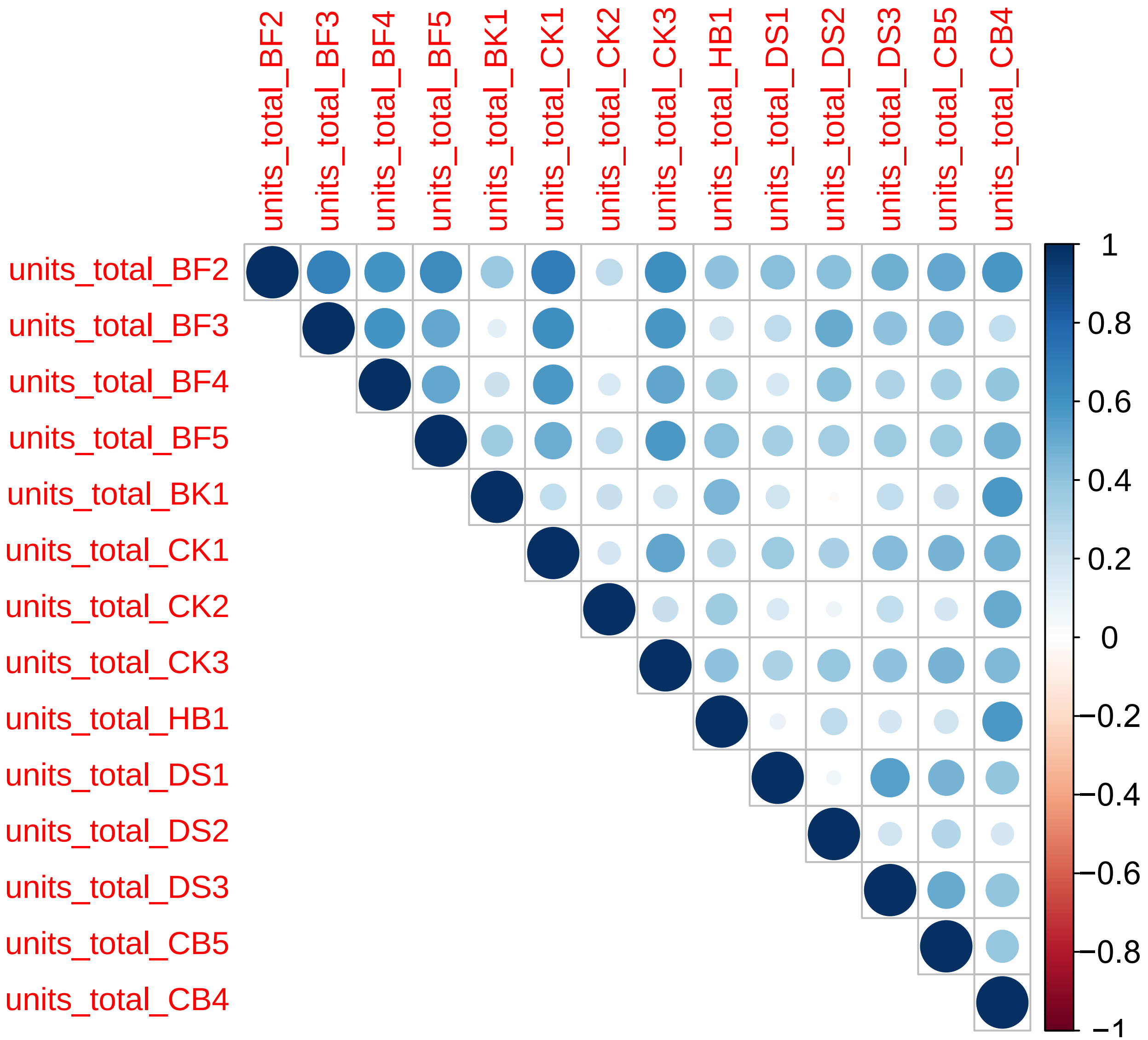



# Correlation matrix of 14 Dependent Variables for modeling



# Data Manipulation and Transformation

# Data Manipulation and Transformation

- For all y variables as well as the promotion prices of items

$$x = \begin{cases} 0.001 & while\ x = 0 \\ x & while\ x \neq 0 \end{cases}$$

- Create new variables: *logDiscount*, *logRegular_price*, *week_of_year*

$$logDiscount = \begin{cases} ln(0.001) & \text{while} \dfrac{promo\ price}{regular\ price} = 1 \\ ln\left(1 - \dfrac{promo\ price}{regular\ price}\right) & \text{while} \dfrac{promo\ price}{regular\ price} < 1 \end{cases}$$

$$logRegular\_price = ln(regular\_price)$$

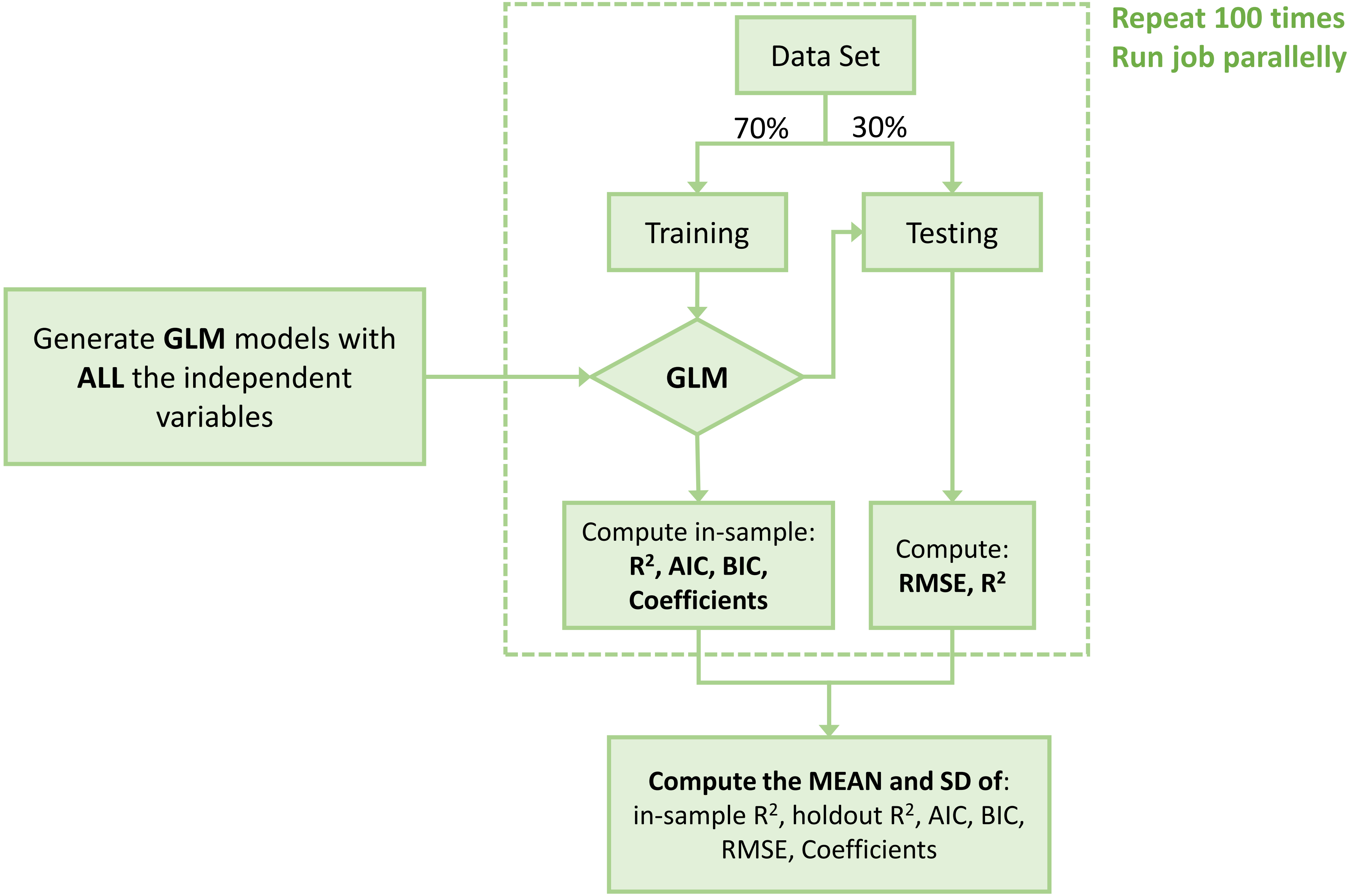
$$week\ of\ year: 1\ to\ 53\ (dummy\ coding)$$

# Computational Details and Preliminary Results

## Part I. Prediction of Daily Sales Volume of Products

# Design of Spliting Dataset for 70-30 Validation

# Working Flow of Validation Process



# Full Model Comparison

| | |
|---|---|
| **GLM 1** | $\log(Sales\ Volume) = \beta_0 + \sum_i \beta_{1i}(day\ of\ week)_i + \sum_j \beta_{2j}(holiday)_j + \sum_k \beta_{3k}(weather)_k + \sum_l \beta_{4l}(national\ promotion)_l + \sum_m \beta_{5m}(tactic\ promotion)_m + \sum_n \beta_{6n} log\left(\frac{promo\_price_n}{regular\_price_n}\right)$ |
| **GLM 2** | $\log(Sales\ Volume) = \beta_0 + \sum_i \beta_{1i}(day\ of\ week)_i + \sum_j \beta_{2j}(holiday)_j + \sum_k \beta_{3k}(weather)_k + \sum_l \beta_{4l}(national\ promotion)_l + \sum_m \beta_{5m}(tactic\ promotion)_m + \sum_n \beta_{6n} logDiscount_n$ |
| | |
| | |

# Adopting log(Discount) rather than log(price ratio) significantly improve the prediction

| | in-sample $R^2$ | | RMSE | | AIC | | BIC | |
|---|---|---|---|---|---|---|---|---|
| Product Name | GLM 1 | GLM 2 | GLM 1 | GLM 2 | GLM 1 | GLM 2 | GLM 1 | GLM 2 |
| units_total_BF2 | 0.15* | 0.21* | 0.344 | 0.331 | 79207 | 70570 | 81442 | 72853 |
| units_total_BF3 | 0.26* | 0.30* | 0.374 | 0.362 | 98135 | 91236 | 100370 | 93520 |
| units_total_BF4 | 0.33* | 0.33* | 0.380 | 0.379 | 101856 | 101713 | 104091 | 103997 |
| units_total_BF5 | 0.23* | 0.26* | 0.396 | 0.388 | 111303 | 106557 | 113539 | 108841 |
| units_total_BK1 | 0.26* | 0.30* | 0.507 | 0.494 | 167376 | 161430 | 169611 | 163713 |
| units_total_CB4 | 0.24* | 0.29* | 0.368 | 0.355 | 94672 | 86648 | 96907 | 88931 |
| units_total_CB5 | 0.18* | 0.19* | 0.528 | 0.525 | 176464 | 175374 | 178699 | 177656 |
| units_total_CK1 | 0.25* | 0.31* | 0.412 | 0.394 | 120361 | 109915 | 122596 | 112198 |
| units_total_CK2 | 0.40* | 0.43* | 0.721 | 0.702 | 247038 | 240884 | 249273 | 243167 |
| units_total_CK3 | 0.24* | 0.26* | 0.393 | 0.388 | 109589 | 106520 | 111824 | 108803 |
| units_total_DS1 | 0.34* | 0.38* | 0.609 | 0.592 | 208952 | 202511 | 211187 | 204794 |
| units_total_DS2 | 0.19* | 0.25* | 0.431 | 0.415 | 130443 | 122057 | 132678 | 124340 |
| units_total_DS3 | 0.21* | 0.23* | 0.591 | 0.586 | 201814 | 199652 | 204049 | 201935 |
| units_total_HB1 | 0.23* | 0.26* | 0.369 | 0.363 | 95490 | 91675 | 97725 | 93958 |

* The *p*-value of the *F* test was significant at $p < 2.2e-16$

# Model Comparison

| | |
|---|---|
| **GLM 1** | $\log(Sales\ Volume) = \beta_0 + \sum_i \beta_{1i}(day\ of\ week)_i + \sum_j \beta_{2j}(holiday)_j + \sum_k \beta_{3k}(weather)_k + \sum_l \beta_{4l}(national\ promotion)_l + \sum_m \beta_{5m}(tactic\ promotion)_m + \sum_n \beta_{6n} log\left(\frac{promo\_price_n}{regular\_price_n}\right)$ |
| **GLM 2** | $\log(Sales\ Volume) = \beta_0 + \sum_i \beta_{1i}(day\ of\ week)_i + \sum_j \beta_{2j}(holiday)_j + \sum_k \beta_{3k}(weather)_k + \sum_l \beta_{4l}(national\ promotion)_l + \sum_m \beta_{5m}(tactic\ promotion)_m + \sum_n \beta_{6n} logDiscount_n$ |
| **GLM 3** | $\log(Sales\ Volume) = \beta_0 + \sum_i \beta_{1i}(day\ of\ week)_i + \sum_j \beta_{2j}(holiday)_j + \sum_k \beta_{3k}(weather)_k + \sum_l \beta_{4l}(national\ promotion)_l + \sum_m \beta_{5m}(tactic\ promotion)_m + \sum_n \beta_{6n} logDiscount_n + \sum_o \beta_{7o} logRegular\_price_o$ |
| | |

# Inclusion of log(regular price) significantly improve the prediction of the models

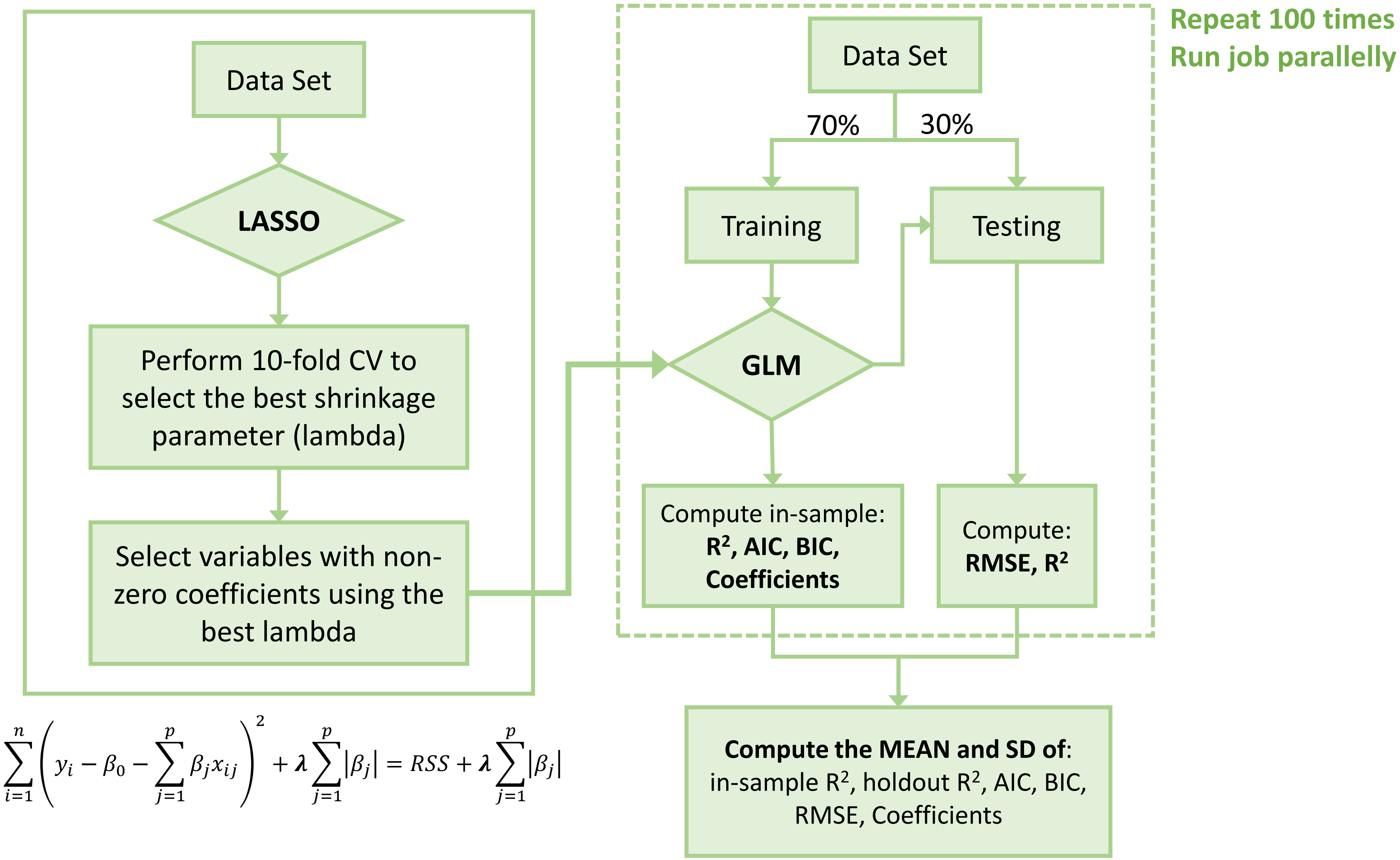
| | in-sample $R^2$ | | | RMSE | | | AIC | | | BIC | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Product Name | GLM 1 | GLM 2 | GLM 3 | GLM 1 | GLM 2 | GLM 3 | GLM 1 | GLM 2 | GLM 3 | GLM 1 | GLM 2 | GLM 3 |
| units_total_BF2 | 0.15* | 0.21* | 0.31* | 0.344 | 0.331 | 0.309 | 79207 | 70570 | 55351 | 81442 | 72853 | 57790 |
| units_total_BF3 | 0.26* | 0.30* | 0.45* | 0.374 | 0.362 | 0.321 | 98135 | 91236 | 63361 | 100370 | 93520 | 65791 |
| units_total_BF4 | 0.33* | 0.33* | 0.42* | 0.380 | 0.379 | 0.353 | 101856 | 101713 | 85200 | 104091 | 103997 | 87553 |
| units_total_BF5 | 0.23* | 0.26* | 0.32* | 0.396 | 0.388 | 0.371 | 111303 | 106557 | 96218 | 113539 | 108841 | 98609 |
| units_total_BK1 | 0.26* | 0.30* | 0.41* | 0.507 | 0.494 | 0.454 | 167376 | 161430 | 142119 | 169611 | 163713 | 144539 |
| units_total_CB4 | 0.24* | 0.29* | 0.37* | 0.368 | 0.355 | 0.333 | 94672 | 86648 | 72191 | 96907 | 88931 | 74702 |
| units_total_CB5 | 0.18* | 0.19* | 0.23* | 0.528 | 0.525 | 0.512 | 176464 | 175374 | 169466 | 178699 | 177656 | 171835 |
| units_total_CK1 | 0.25* | 0.31* | 0.39* | 0.412 | 0.394 | 0.371 | 120361 | 109915 | 96207 | 122596 | 112198 | 98714 |
| units_total_CK2 | 0.40* | 0.43* | 0.59* | 0.721 | 0.702 | 0.600 | 247038 | 240884 | 205221 | 249273 | 243167 | 207620 |
| units_total_CK3 | 0.24* | 0.26* | 0.32* | 0.393 | 0.388 | 0.372 | 109589 | 106520 | 97062 | 111824 | 108803 | 99617 |
| units_total_DS1 | 0.34* | 0.38* | 0.41* | 0.609 | 0.592 | 0.575 | 208952 | 202511 | 195859 | 211187 | 204794 | 198251 |
| units_total_DS2 | 0.19* | 0.25* | 0.35* | 0.431 | 0.415 | 0.385 | 130443 | 122057 | 105269 | 132678 | 124340 | 107681 |
| units_total_DS3 | 0.21* | 0.23* | 0.28* | 0.591 | 0.586 | 0.565 | 201814 | 199652 | 191889 | 204049 | 201935 | 194325 |
| units_total_HB1 | 0.23* | 0.26* | 0.37* | 0.369 | 0.363 | 0.335 | 95490 | 91675 | 73947 | 97725 | 93958 | 76462 |

* The *p*-value of the *F* test was significant at $p < 2.2e-16$

# Model Comparison

| | |
|---|---|
| **GLM 1** | $\log(Sales\ Volume) = \beta_0 + \sum_i \beta_{1i}(day\ of\ week)_i + \sum_j \beta_{2j}(holiday)_j + \sum_k \beta_{3k}(weather)_k + \sum_l \beta_{4l}(national\ promotion)_l + \sum_m \beta_{5m}(tactic\ promotion)_m + \sum_n \beta_{6n} log\left(\frac{promo\_price_n}{regular\_price_n}\right)$ |
| **GLM 2** | $\log(Sales\ Volume) = \beta_0 + \sum_i \beta_{1i}(day\ of\ week)_i + \sum_j \beta_{2j}(holiday)_j + \sum_k \beta_{3k}(weather)_k + \sum_l \beta_{4l}(national\ promotion)_l + \sum_m \beta_{5m}(tactic\ promotion)_m + \sum_n \beta_{6n} logDiscount_n$ |
| **GLM 3** | $\log(Sales\ Volume) = \beta_0 + \sum_i \beta_{1i}(day\ of\ week)_i + \sum_j \beta_{2j}(holiday)_j + \sum_k \beta_{3k}(weather)_k + \sum_l \beta_{4l}(national\ promotion)_l + \sum_m \beta_{5m}(tactic\ promotion)_m + \sum_n \beta_{6n} logDiscount_n + \sum_o \beta_{7o} logRegular\_price_o$ |
| **GLM 4** | $\log(Sales\ Volume) = \beta_0 + \sum_i \beta_{1i}(day\ of\ week)_i + \underbrace{\boxed{\sum_j \beta_{2j}(holiday)_j + \sum_k \beta_{3k}(weather)_k + \sum_l \beta_{4l}(national\ promotion)_l + \sum_m \beta_{5m}(tactic\ promotion)_m}}_{\textbf{LASSO}} + \sum_n \beta_{6n} logDiscount_n + \sum_o \beta_{7o} logRegular\_price_o$ |

# Working Flow of Lasso Regression




$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j|$$

# Lasso regression did not improve the predicting performance

| | in-sample $R^2$ | | | | RMSE | | | | AIC | | | | BIC | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Product Name | GLM 1 | GLM 2 | GLM 3 | GLM 4 | GLM 1 | GLM 2 | GLM 3 | GLM 4 | GLM 1 | GLM 2 | GLM 3 | GLM 4 | GLM 1 | GLM 2 | GLM 3 | GLM 4 |
| units_total_BF2 | 0.15* | 0.21* | 0.31* | 0.31* | 0.344 | 0.331 | 0.309 | 0.309 | 79207 | 70570 | 55351 | 55362 | 81442 | 72853 | 57790 | 57751 |
| units_total_BF3 | 0.26* | 0.30* | 0.45* | 0.45* | 0.374 | 0.362 | 0.321 | 0.321 | 98135 | 91236 | 63361 | 63363 | 100370 | 93520 | 65791 | 65731 |
| units_total_BF4 | 0.33* | 0.33* | 0.42* | 0.42* | 0.380 | 0.379 | 0.353 | 0.353 | 101856 | 101713 | 85200 | 85125 | 104091 | 103997 | 87553 | 87599 |
| units_total_BF5 | 0.23* | 0.26* | 0.32* | 0.32* | 0.396 | 0.388 | 0.371 | 0.370 | 111303 | 106557 | 96218 | 96181 | 113539 | 108841 | 98609 | 98607 |
| units_total_BK1 | 0.26* | 0.30* | 0.41* | 0.41* | 0.507 | 0.494 | 0.454 | 0.454 | 167376 | 161430 | 142119 | 142111 | 169611 | 163713 | 144539 | 144518 |
| units_total_CB4 | 0.24* | 0.29* | 0.37* | 0.37* | 0.368 | 0.355 | 0.333 | 0.333 | 94672 | 86648 | 72191 | 72274 | 96907 | 88931 | 74702 | 74599 |
| units_total_CB5 | 0.18* | 0.19* | 0.23* | 0.23* | 0.528 | 0.525 | 0.512 | 0.512 | 176464 | 175374 | 169466 | 169407 | 178699 | 177656 | 171835 | 171894 |
| units_total_CK1 | 0.25* | 0.31* | 0.39* | 0.39* | 0.412 | 0.394 | 0.371 | 0.371 | 120361 | 109915 | 96207 | 96286 | 122596 | 112198 | 98714 | 98596 |
| units_total_CK2 | 0.40* | 0.43* | 0.59* | 0.59* | 0.721 | 0.702 | 0.600 | 0.600 | 247038 | 240884 | 205221 | 205192 | 249273 | 243167 | 207620 | 207600 |
| units_total_CK3 | 0.24* | 0.26* | 0.32* | 0.32* | 0.393 | 0.388 | 0.372 | 0.372 | 109589 | 106520 | 97062 | 97189 | 111824 | 108803 | 99617 | 99480 |
| units_total_DS1 | 0.34* | 0.38* | 0.41* | 0.41* | 0.609 | 0.592 | 0.575 | 0.575 | 208952 | 202511 | 195859 | 195823 | 211187 | 204794 | 198251 | 198267 |
| units_total_DS2 | 0.19* | 0.25* | 0.35* | 0.35* | 0.431 | 0.415 | 0.385 | 0.385 | 130443 | 122057 | 105269 | 105253 | 132678 | 124340 | 107681 | 107667 |
| units_total_DS3 | 0.21* | 0.23* | 0.28* | 0.28* | 0.591 | 0.586 | 0.565 | 0.565 | 201814 | 199652 | 191889 | 191898 | 204049 | 201935 | 194325 | 194278 |
| units_total_HB1 | 0.23* | 0.26* | 0.37* | 0.37* | 0.369 | 0.363 | 0.335 | 0.336 | 95490 | 91675 | 73947 | 74035 | 97725 | 93958 | 76462 | 76318 |

* The *p*-value of the *F* test was significant at $p < 2.2\text{e-}16$

# It is because the dimension of variables was not greatly reduced using Lasso

**Number of Variables Selected Using Lasso:**

| Product name | Number of Variables **(original=260)** |
|---|---|
| units_total_BF2 | 254 |
| units_total_BF3 | 254 |
| units_total_BF4 | 252 |
| units_total_BF5 | 255 |
| units_total_BK1 | 253 |
| units_total_CB4 | 254 |
| units_total_CB5 | 257 |
| units_total_CK1 | 255 |
| units_total_CK2 | 249 |
| units_total_CK3 | 255 |
| units_total_DS1 | 255 |
| units_total_DS2 | 256 |
| units_total_DS3 | 252 |
| units_total_HB1 | 251 |

# Model Comparison

| | |
|---|---|
| **GLM 1** | $\log(Sales\ Volume) = \beta_0 + \sum_i \beta_{1i}(day\ of\ week)_i + \sum_j \beta_{2j}(holiday)_j + \sum_k \beta_{3k}(weather)_k + \sum_l \beta_{4l}(national\ promotion)_l + \sum_m \beta_{5m}(tactic\ promotion)_m + \sum_n \beta_{6n} log\left(\frac{promo\_price_n}{regular\_price_n}\right)$ |
| **GLM 2** | $\log(Sales\ Volume) = \beta_0 + \sum_i \beta_{1i}(day\ of\ week)_i + \sum_j \beta_{2j}(holiday)_j + \sum_k \beta_{3k}(weather)_k + \sum_l \beta_{4l}(national\ promotion)_l + \sum_m \beta_{5m}(tactic\ promotion)_m + \sum_n \beta_{6n} logDiscount_n$ |
| **GLM 3** | $\log(Sales\ Volume) = \beta_0 + \sum_i \beta_{1i}(day\ of\ week)_i + \sum_j \beta_{2j}(holiday)_j + \sum_k \beta_{3k}(weather)_k + \sum_l \beta_{4l}(national\ promotion)_l + \sum_m \beta_{5m}(tactic\ promotion)_m + \sum_n \beta_{6n} logDiscount_n + \sum_o \beta_{7o} logRegular\_price_o$ |
| **GLM 5** | $\log(Sales\ Volume) = \beta_0 + \sum_i \beta_{1i}(day\ of\ week)_i + \sum_j \beta_{2j}(holiday)_j + \sum_k \beta_{3k}(weather)_k + \sum_l \beta_{4l}(national\ promotion)_l + \sum_m \beta_{5m}(tactic\ promotion)_m + \sum_n \beta_{6n} logDiscount_n + \sum_o \beta_{7o} logRegular\_price_o + \sum_p \beta_{8p}(week\ of\ year)_p$ |

# Inclusion of week_of_year slightly improves the models

| | in-sample $R^2$ | | | | RMSE | | | | AIC | | | | BIC | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Product Name | GLM1 | GLM 2 | GLM 3 | GLM 5 | GLM1 | GLM 2 | GLM 3 | GLM 5 | GLM1 | GLM 2 | GLM 3 | GLM 5 | GLM1 | GLM 2 | GLM 3 | GLM 5 |
| units_total_BF2 | 0.15* | 0.21* | 0.31* | 0.32* | 0.344 | 0.331 | 0.309 | 0.307 | 79207 | 70570 | 55351 | 53361 | 81442 | 72853 | 57790 | 56288 |
| units_total_BF3 | 0.26* | 0.30* | 0.45* | 0.46* | 0.374 | 0.362 | 0.321 | 0.319 | 98135 | 91236 | 63361 | 61471 | 100370 | 93520 | 65791 | 64398 |
| units_total_BF4 | 0.33* | 0.33* | 0.42* | 0.45* | 0.380 | 0.379 | 0.353 | 0.346 | 101856 | 101713 | 85200 | 80088 | 104091 | 103997 | 87553 | 83016 |
| units_total_BF5 | 0.23* | 0.26* | 0.32* | 0.34* | 0.396 | 0.388 | 0.371 | 0.365 | 111303 | 106557 | 96218 | 92537 | 113539 | 108841 | 98609 | 95464 |
| units_total_BK1 | 0.26* | 0.30* | 0.41* | 0.42* | 0.507 | 0.494 | 0.454 | 0.449 | 167376 | 161430 | 142119 | 138928 | 169611 | 163713 | 144539 | 141855 |
| units_total_CB4 | 0.24* | 0.29* | 0.37* | 0.40* | 0.368 | 0.355 | 0.333 | 0.327 | 94672 | 86648 | 72191 | 67440 | 96907 | 88931 | 74702 | 70367 |
| units_total_CB5 | 0.18* | 0.19* | 0.23* | 0.27* | 0.528 | 0.525 | 0.512 | 0.498 | 176464 | 175374 | 169466 | 162003 | 178699 | 177656 | 171835 | 164930 |
| units_total_CK1 | 0.25* | 0.31* | 0.39* | 0.41* | 0.412 | 0.394 | 0.371 | 0.366 | 120361 | 109915 | 96207 | 93048 | 122596 | 112198 | 98714 | 95975 |
| units_total_CK2 | 0.40* | 0.43* | 0.59* | 0.60* | 0.721 | 0.702 | 0.600 | 0.590 | 247038 | 240884 | 205221 | 200440 | 249273 | 243167 | 207620 | 203367 |
| units_total_CK3 | 0.24* | 0.26* | 0.32* | 0.34* | 0.393 | 0.388 | 0.372 | 0.366 | 109589 | 106520 | 97062 | 92946 | 111824 | 108803 | 99617 | 95873 |
| units_total_DS1 | 0.34* | 0.38* | 0.41* | 0.55* | 0.609 | 0.592 | 0.575 | 0.506 | 208952 | 202511 | 195859 | 165942 | 211187 | 204794 | 198251 | 168869 |
| *units_total_DS2 | 0.19* | 0.25* | 0.35* | 0.38* | 0.431 | 0.415 | 0.385 | 0.378 | 130443 | 122057 | 105269 | 100003 | 132678 | 124340 | 107681 | 102930 |
| units_total_DS3 | 0.21* | 0.23* | 0.28* | 0.32* | 0.591 | 0.586 | 0.565 | 0.549 | 201814 | 199652 | 191889 | 184077 | 204049 | 201935 | 194325 | 187004 |
| units_total_HB1 | 0.23* | 0.26* | 0.37* | 0.38* | 0.369 | 0.363 | 0.335 | 0.330 | 95490 | 91675 | 73947 | 69322 | 97725 | 93958 | 76462 | 72250 |

* The *p*-value of the *F* test was significant at $p < 2.2e-16$

# Model Comparison

| | |
|---|---|
| **GLM 5** | $\log(Sales\ Volume)$ $= \beta_0 + \sum_i \beta_{1i}(day\ of\ week)_i + \sum_j \beta_{2j}(holiday)_j + \sum_k \beta_{3k}(weather)_k + \sum_l \beta_{4l}(national\ promotion)_l + \sum_m \beta_{5m}(tactic\ promotion)_m + \sum_n \beta_{6n} logDiscount_n + \sum_o \beta_{7o} logRegular\_price_o + \sum_p \beta_{8p}(week\ of\ year)_p$ |
| **GLM 6** | $log(Sales\ Volume)_{storeID}$ $= \beta_{0,storeID} + \sum_i \beta_{1i,storeID}(day\ of\ week)_{i,storeID} + \sum_j \beta_{2j,storeID}(holiday)_{j,storeID} + \sum_k \beta_{3k,storeID}(weather)_{k,storeID} + \sum_l \beta_{4l,storeID}(national\ promotion)_{l,storeID} + \sum_m \beta_{5m,storeID}(tactic\ promotion)_{m,storeID} + \sum_n \beta_{6n,storeID} logDiscount_{n,storeID} + \sum_o \beta_{7o,storeID} logRegular\_price_{o,storeID} + \sum_p \beta_{8p,storeID}(week\ of\ year)_{p,storeID}$ |
| | |

# Optimal approach: fit models for each product in each store

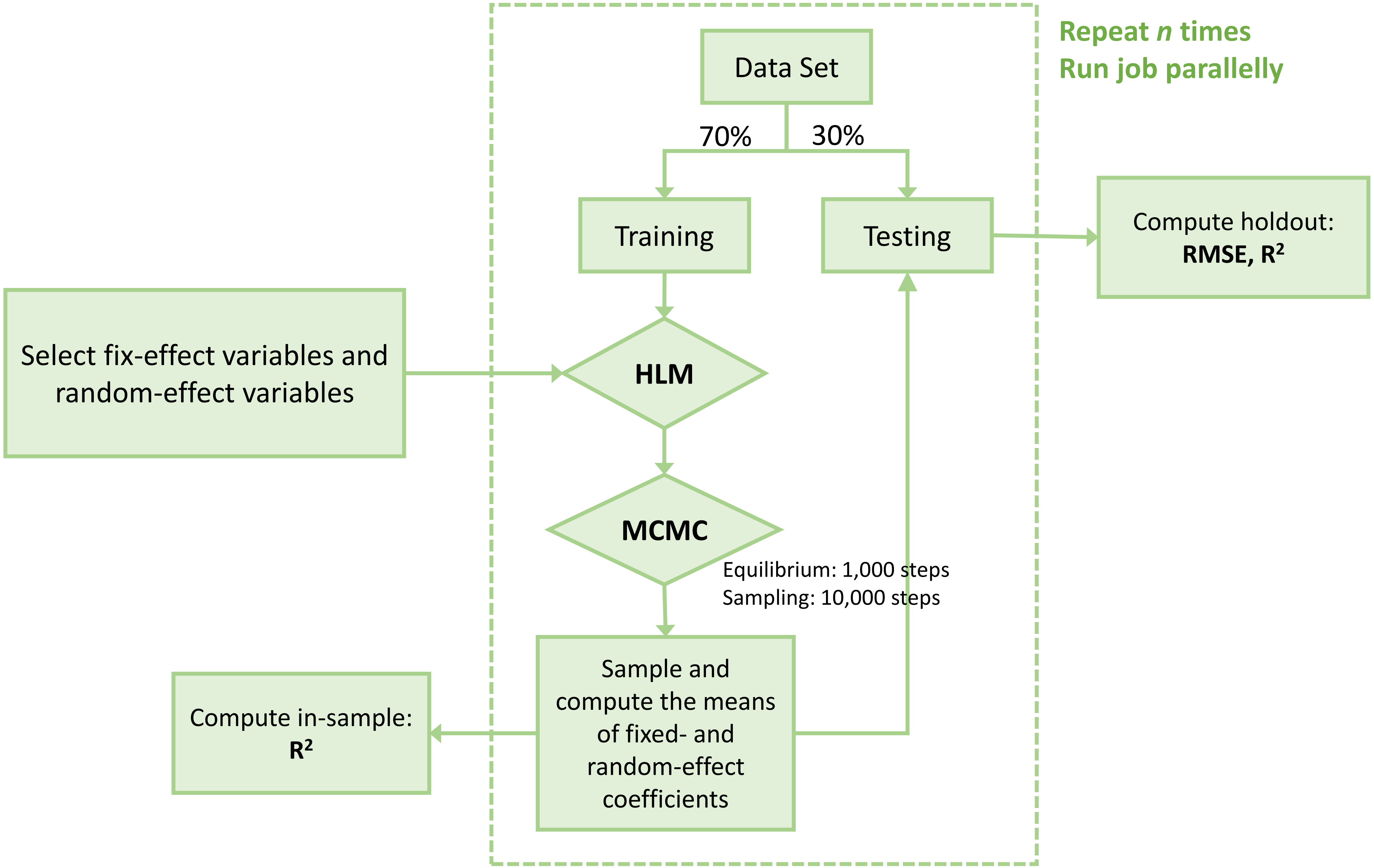
| | in-sample $R^2$ | | | | | holdout $R^2$ | | | | | RMSE | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Product Name | GLM 1 | GLM 2 | GLM 3 | GLM 5 | GLM 6 | GLM 1 | GLM 2 | GLM 3 | GLM 5 | GLM 6 | GLM 1 | GLM 2 | GLM 3 | GLM 5 | GLM 6 |
| units_total_BF2 | 0.15* | 0.21* | 0.31* | 0.32* | **0.86** | 0.14 | 0.21 | 0.31 | 0.32 | **0.77** | 0.344 | 0.331 | 0.309 | 0.307 | **0.177** |
| units_total_BF3 | 0.26* | 0.30* | 0.45* | 0.46* | **0.90** | 0.25 | 0.30 | 0.45 | 0.46 | **0.83** | 0.374 | 0.362 | 0.321 | 0.319 | **0.176** |
| units_total_BF4 | 0.33* | 0.33* | 0.42* | 0.45* | **0.86** | 0.33 | 0.33 | 0.42 | 0.44 | **0.77** | 0.380 | 0.379 | 0.353 | 0.346 | **0.223** |
| units_total_BF5 | 0.23* | 0.26* | 0.32* | 0.34* | **0.82** | 0.22 | 0.26 | 0.32 | 0.34 | **0.71** | 0.396 | 0.388 | 0.371 | 0.365 | **0.241** |
| units_total_BK1 | 0.26* | 0.30* | 0.41* | 0.42* | **0.88** | 0.26 | 0.29 | 0.40 | 0.42 | **0.81** | 0.507 | 0.494 | 0.454 | 0.449 | **0.259** |
| units_total_CB4 | 0.24* | 0.29* | 0.37* | 0.40* | **0.94** | 0.23 | 0.29 | 0.37 | 0.40 | **0.90** | 0.368 | 0.355 | 0.333 | 0.327 | **0.136** |
| units_total_CB5 | 0.18* | 0.19* | 0.23* | 0.27* | **0.66** | 0.18 | 0.19 | 0.23 | 0.27 | **0.44** | 0.528 | 0.525 | 0.512 | 0.498 | **0.435** |
| units_total_CK1 | 0.25* | 0.31* | 0.39* | 0.41* | **0.85** | 0.24 | 0.31 | 0.39 | 0.40 | **0.74** | 0.412 | 0.394 | 0.371 | 0.366 | **0.243** |
| units_total_CK2 | 0.40* | 0.43* | 0.59* | 0.60* | **0.86** | 0.40 | 0.43 | 0.59 | 0.60 | **0.77** | 0.721 | 0.702 | 0.600 | 0.590 | **0.448** |
| units_total_CK3 | 0.24* | 0.26* | 0.32* | 0.34* | **0.82** | 0.24 | 0.26 | 0.32 | 0.34 | **0.69** | 0.393 | 0.388 | 0.372 | 0.366 | **0.249** |
| units_total_DS1 | 0.34* | 0.38* | 0.41* | 0.55* | **0.80** | 0.34 | 0.38 | 0.41 | 0.54 | **0.66** | 0.609 | 0.592 | 0.575 | 0.506 | **0.439** |
| *units_total_DS2 | 0.19* | 0.25* | 0.35* | 0.38* | **0.84** | 0.19 | 0.25 | 0.35 | 0.37 | **0.73** | 0.431 | 0.415 | 0.385 | 0.378 | **0.249** |
| units_total_DS3 | 0.21* | 0.23* | 0.28* | 0.32* | **0.70** | 0.21 | 0.23 | 0.28 | 0.32 | **0.52** | 0.591 | 0.586 | 0.565 | 0.549 | **0.462** |
| units_total_HB1 | 0.23* | 0.26* | 0.37* | 0.38* | **0.88** | 0.23 | 0.26 | 0.36 | 0.38 | **0.81** | 0.369 | 0.363 | 0.335 | 0.330 | **0.184** |

* The *p*-value of the *F* test was significant at $p < 2.2e\text{-}16$

# Model Comparison

| | |
|---|---|
| **GLM 5** | $\log(Sales\ Volume) = \beta_0 + \sum_i \beta_{1i}(day\ of\ week)_i + \sum_j \beta_{2j}(holiday)_j + \sum_k \beta_{3k}(weather)_k + \sum_l \beta_{4l}(national\ promotion)_l + \sum_m \beta_{5m}(tactic\ promotion)_m + \sum_n \beta_{6n} logDiscount_n + \sum_o \beta_{7o} logRegular\_price_o + \sum_p \beta_{8p}(week\ of\ year)_p$ |
| **GLM 6** | $log(Sales\ Volume)_{storeID} = \beta_{0,storeID} + \sum_i \beta_{1i,storeID}(day\ of\ week)_{i,storeID} + \sum_j \beta_{2j,storeID}(holiday)_{j,storeID} + \sum_k \beta_{3k,storeID}(weather)_{k,storeID} + \sum_l \beta_{4l,storeID}(national\ promotion)_{l,storeID} + \sum_m \beta_{5m,storeID}(tactic\ promotion)_{m,storeID} + \sum_n \beta_{6n,storeID} logDiscount_{n,storeID} + \sum_o \beta_{7o,storeID} logRegular\_price_{o,storeID} + \sum_p \beta_{8p,storeID}(week\ of\ year)_{p,storeID}$ |
| **MCMC 1** **(HLM/MCMC)** | $log(Sales\ Volume) = \beta_0 + \sum_i \beta_{1i}(day\ of\ week)_i + \sum_j \beta_{2j}(holiday)_j + \sum_k \beta_{3k}(weather)_k + \sum_l \beta_{4l}(national\ promotion)_l + \sum_m \beta_{5m}(tactic\ promotion)_m + \sum_n \beta_{6n} logDiscount_n + \sum_o \beta_{7o} logRegular\_price_o + \sum_{p,storeID} \alpha_{0p,storeID} + \sum_{q,storeID} \alpha_{1q} logDiscount_{q,storeID}$ |

# Working Flow of Combined Hierarchical Linear Modeling/MCMC Simulation



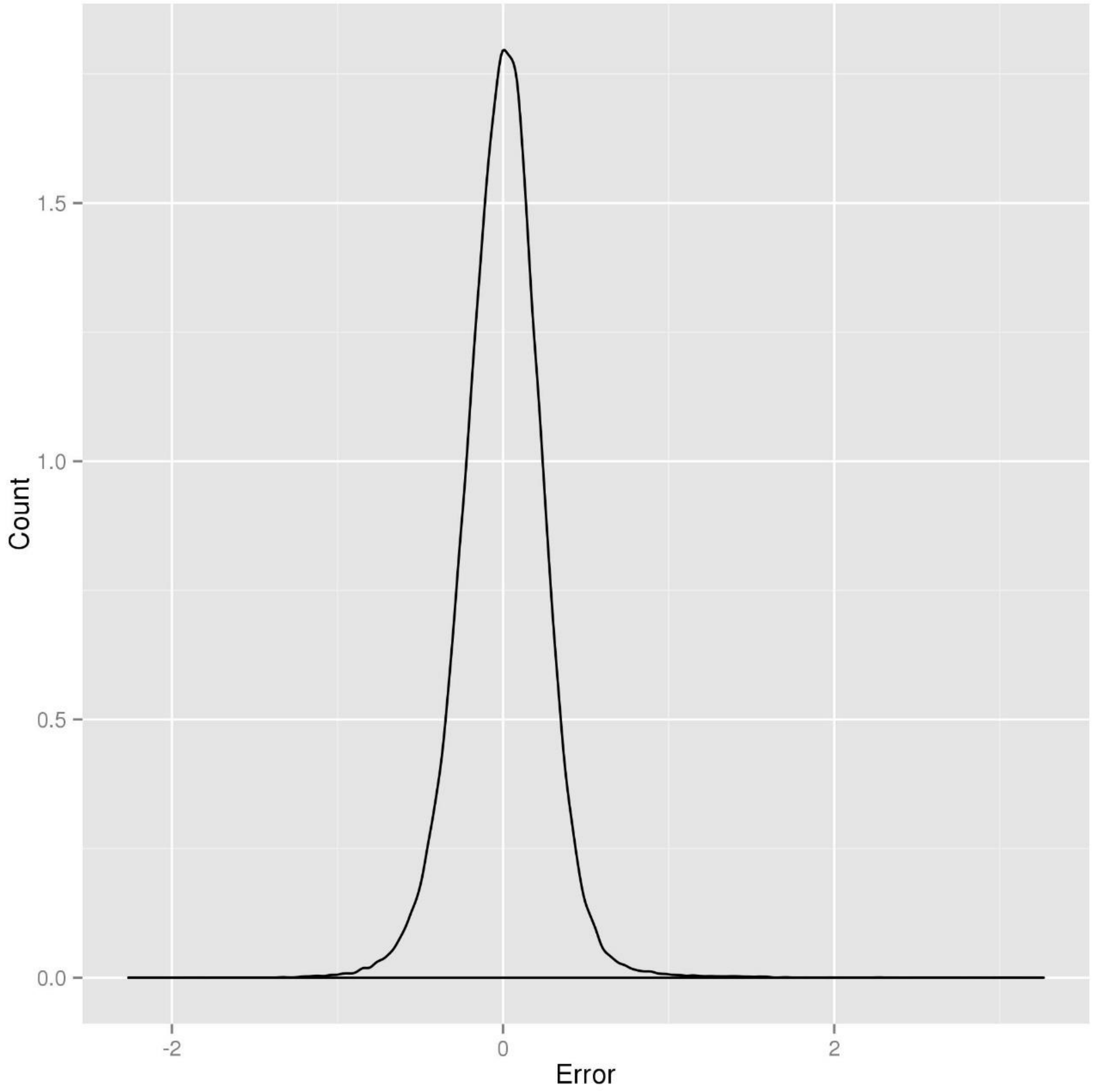
# Model Comparison

| | in-sample $R^2$ | | | | | | holdout $R^2$ | | | | | | RMSE | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Product Name | GLM 1 | GLM 2 | GLM 3 | GLM 5 | GLM 6 | MCMC 1 | GLM 1 | GLM 2 | GLM 3 | GLM 5 | GLM 6 | MCMC 1 | GLM 1 | GLM 2 | GLM 3 | GLM 5 | GLM 6 | MCMC 1 |
| units_total_BF2 | 0.15* | 0.21* | 0.31* | 0.32* | **0.86** | 0.76 | 0.14 | 0.21 | 0.31 | 0.32 | **0.77** | 0.58 | 0.344 | 0.331 | 0.309 | 0.307 | **0.177** | 0.241 |
| units_total_BF3 | 0.26* | 0.30* | 0.45* | 0.46* | **0.90** | 0.81 | 0.25 | 0.30 | 0.45 | 0.46 | **0.83** | 0.67 | 0.374 | 0.362 | 0.321 | 0.319 | **0.176** | 0.246 |
| units_total_BF4 | 0.33* | 0.33* | 0.42* | 0.45* | **0.86** | 0.75 | 0.33 | 0.33 | 0.42 | 0.44 | **0.77** | 0.63 | 0.380 | 0.379 | 0.353 | 0.346 | **0.223** | 0.280 |
| units_total_BF5 | 0.23* | 0.26* | 0.32* | 0.34* | **0.82** | 0.70 | 0.22 | 0.26 | 0.32 | 0.34 | **0.71** | 0.57 | 0.396 | 0.388 | 0.371 | 0.365 | **0.241** | 0.292 |
| units_total_BK1 | 0.26* | 0.30* | 0.41* | 0.42* | **0.88** | 0.80 | 0.26 | 0.29 | 0.40 | 0.42 | **0.81** | 0.72 | 0.507 | 0.494 | 0.454 | 0.449 | **0.259** | 0.308 |
| units_total_CB4 | 0.24* | 0.29* | 0.37* | 0.40* | **0.94** | 0.85 | 0.23 | 0.29 | 0.37 | 0.40 | **0.90** | 0.71 | 0.368 | 0.355 | 0.333 | 0.327 | **0.136** | 0.226 |
| units_total_CB5 | 0.18* | 0.19* | 0.23* | 0.27* | **0.66** | 0.51 | 0.18 | 0.19 | 0.23 | 0.27 | **0.44** | 0.42 | 0.528 | 0.525 | 0.512 | 0.498 | **0.435** | 0.443 |
| units_total_CK1 | 0.25* | 0.31* | 0.39* | 0.41* | **0.85** | 0.75 | 0.24 | 0.31 | 0.39 | 0.40 | **0.74** | 0.63 | 0.412 | 0.394 | 0.371 | 0.366 | **0.243** | 0.286 |
| units_total_CK2 | 0.40* | 0.43* | 0.59* | 0.60* | **0.86** | 0.76 | 0.40 | 0.43 | 0.59 | 0.60 | **0.77** | 0.72 | 0.721 | 0.702 | 0.600 | 0.590 | **0.448** | 0.491 |
| units_total_CK3 | 0.24* | 0.26* | 0.32* | 0.34* | **0.82** | 0.69 | 0.24 | 0.26 | 0.32 | 0.34 | **0.69** | 0.56 | 0.393 | 0.388 | 0.372 | 0.366 | **0.249** | 0.298 |
| units_total_DS1 | 0.34* | 0.38* | 0.41* | 0.55* | **0.80** | 0.71 | 0.34 | 0.38 | 0.41 | 0.54 | **0.66** | 0.65 | 0.609 | 0.592 | 0.575 | 0.506 | **0.439** | 0.445 |
| units_total_DS2 | 0.19* | 0.25* | 0.35* | 0.38* | **0.84** | 0.71 | 0.19 | 0.25 | 0.35 | 0.37 | **0.73** | 0.59 | 0.431 | 0.415 | 0.385 | 0.378 | **0.249** | 0.304 |
| units_total_DS3 | 0.21* | 0.23* | 0.28* | 0.32* | **0.70** | 0.56 | 0.21 | 0.23 | 0.28 | 0.32 | **0.52** | 0.49 | 0.591 | 0.586 | 0.565 | 0.549 | **0.462** | 0.473 |
| units_total_HB1 | 0.23* | 0.26* | 0.37* | 0.38* | **0.88** | 0.78 | 0.23 | 0.26 | 0.36 | 0.38 | **0.81** | 0.64 | 0.369 | 0.363 | 0.335 | 0.330 | **0.184** | 0.253 |

* The *p*-value of the *F* test was significant at $p < 2.2e-16$

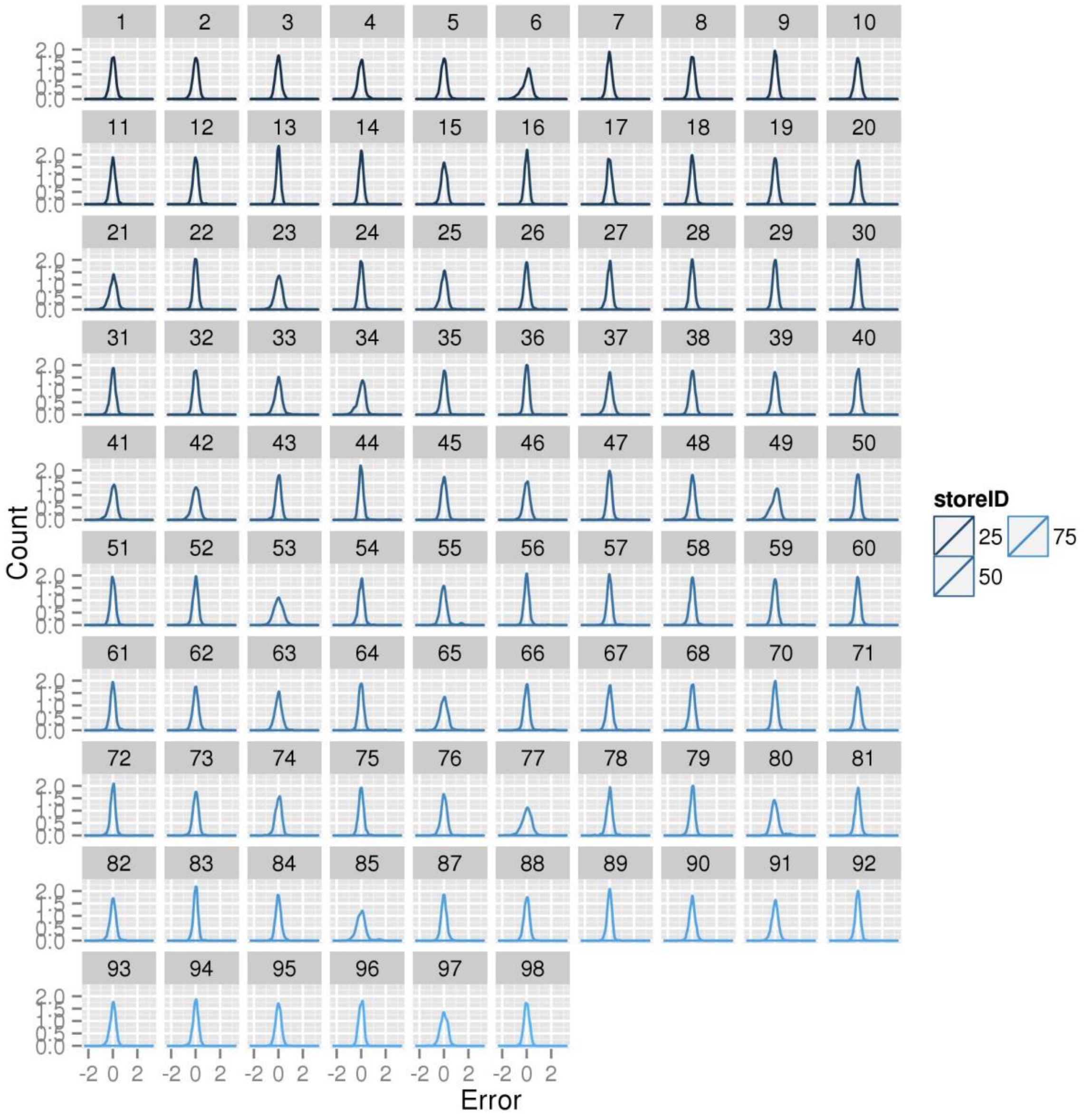
# Density Distribution of Residuals for the Prediction of units_total_BF2

Example: units_total_BF2



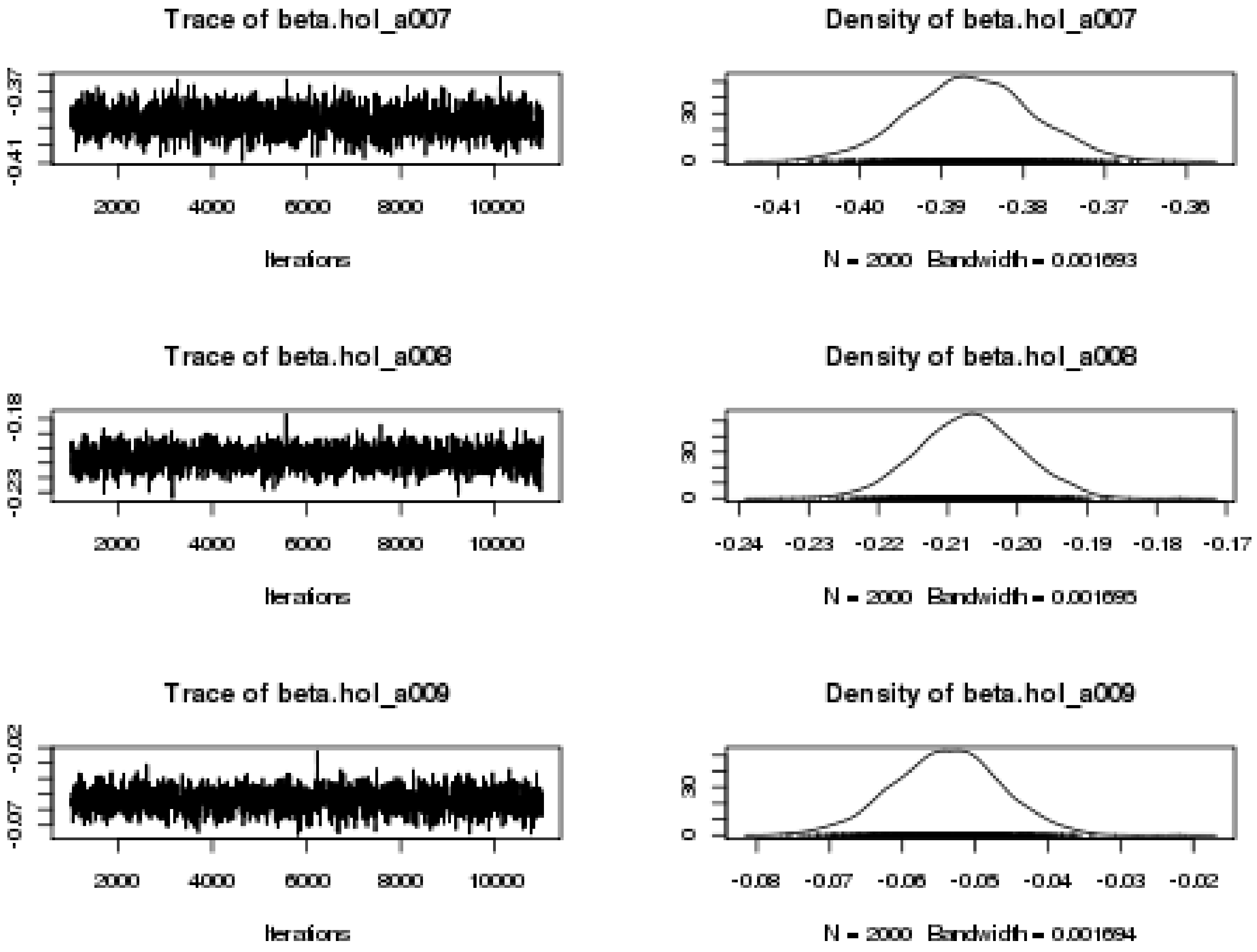
# Density Distribution of Residuals for the Prediction of units_total_BF2 for Each Store

Example: units_total_BF2



# Convergence: trace of the movement of the means over each iteration of the chain

Example: units_total_BF2



# Hierarchical linear modeling was not converged due to the curse of variable dimension

| | |
|---|---|
| **GLM 6** | $log(Sales\ Volume)_{storeID} = \beta_{0,storeID} + \sum_i \beta_{1i,storeID}(day\ of\ week)_{i,storeID} + \sum_j \beta_{2j,storeID}(holiday)_{j,storeID} + \sum_k \beta_{3k,storeID}(weather)_{k,storeID} + \sum_l \beta_{4l,storeID}(national\ promotion)_{l,storeID} + \sum_m \beta_{5m,storeID}(tactic\ promotion)_{m,storeID} + \sum_n \beta_{6n,storeID} logDiscount_{n,storeID} + + \sum_o \beta_{7o,storeID} logRegular\_price_{o,storeID} + \sum_p \beta_{8p,storeID}(week\ of\ year)_{p,storeID}$ |
| **MCMC 1** **(HLM/MCMC)** | $log(Sales\ Volume) = \beta_0 + \sum_i \beta_{1i}(day\ of\ week)_i + \sum_j \beta_{2j}(holiday)_j + \sum_k \beta_{3k}(weather)_k + \sum_l \beta_{4l}(national\ promotion)_l + \sum_m \beta_{5m}(tactic\ promotion)_m + \sum_n \beta_{6n} logDiscount_n + \sum_o \beta_{7o} logRegular\_price_o + \sum_{p,storeID} \alpha_{0p,storeID} + \sum_{q,storeID} \alpha_{1q} logDiscount_{q,storeID}$ |
| **HLM 1** | $log(Sales\ Volume) = \beta_0 + \sum_i \beta_{1i}(day\ of\ week)_i + \sum_j \beta_{2j}(holiday)_j + \sum_k \beta_{3k}(weather)_k + \sum_l \beta_{4l}(national\ promotion)_l + \sum_m \beta_{5m}(tactic\ promotion)_m + \sum_n \beta_{6n} logDiscount_n + \sum_o \beta_{7o} logRegular\_price_o + \sum_{p,storeID} \alpha_{0p,storeID} + \sum_{q,storeID} \alpha_{1q} logDiscount_{q,storeID}$ <br> *Over time limitation (36 hrs) on RCC!* |

# Model Comparison

| | |
|---|---|
| **GLM 6** | $\log(\text{Sales Volume})_{storeID} = \beta_{0,storeID} + \sum_i \beta_{1i,storeID}(\text{day of week})_{i,storeID} + \sum_j \beta_{2j,storeID}(\text{holiday})_{j,storeID} + \sum_k \beta_{3k,storeID}(\text{weather})_{k,storeID} + \sum_l \beta_{4l,storeID}(\text{national promotion})_{l,storeID} + \sum_m \beta_{5m,storeID}(\text{tactic promotion})_{m,storeID} + \sum_n \beta_{6n,storeID} logDiscount_{n,storeID} + + \sum_o \beta_{7o,storeID} logRegular\_price_{o,storeID} + \sum_p \beta_{8p,storeID}(\text{week of year})_{p,storeID}$ |
| **MCMC 1** (HLM/MCMC) | $\log(\text{Sales Volume}) = \beta_0 + \sum_i \beta_{1i}(\text{day of week})_i + \sum_j \beta_{2j}(\text{holiday})_j + \sum_k \beta_{3k}(\text{weather})_k + \sum_l \beta_{4l}(\text{national promotion})_l + \sum_m \beta_{5m}(\text{tactic promotion})_m + \sum_n \beta_{6n} logDiscount_n + \sum_o \beta_{7o} logRegular\_price_o + \sum_{p,storeID} \alpha_{0p,storeID} + \sum_{q,storeID} \alpha_{1q} logDiscount_{q,storeID}$ |
| **MCMC 2** (HLM/MCMC) | $\log(\text{Sales Volume}) = \beta_0 + \sum_i \beta_{1i}(\text{day of week})_i + \sum_j \beta_{2j}(\text{holiday})_j + \sum_k \beta_{3k}(\text{weather})_k + \sum_l \beta_{4l}(\text{national promotion})_l + \sum_m \beta_{5m}(\text{tactic promotion})_m + \sum_n \beta_{6n} logDiscount_n + \sum_o \beta_{7o} logRegular\_price_o + \sum_p \beta_{8p}(\text{week of year})_p + \sum_{p,storeID} \alpha_{0p,storeID} + \sum_{q,storeID} \alpha_{1q} logDiscount_{q,storeID} + \sum_{r,storeID} \alpha_{2r} logRegular\_price_{r,storeID}$ |

# MCMC2 has comparative holdout $R^2$ and in-sample RMSE relative to GLM6

| | in-sample $R^2$ | | | | | | | holdout $R^2$ | | | | | | | RMSE | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Product Name | GLM 1 | GLM 2 | GLM 3 | GLM 5 | **GLM 6** | MCMC 1 | MCMC 2 | GLM 1 | GLM 2 | GLM 3 | GLM 5 | **GLM 6** | MCMC 1 | MCMC 2 | GLM 1 | GLM 2 | GLM 3 | GLM 5 | **GLM 6** | MCMC 1 | MCMC 2 |
| units_total_BF2 | 0.15* | 0.21* | 0.31* | 0.32* | **0.86** | 0.76 | 0.78 | 0.14 | 0.21 | 0.31 | 0.32 | **0.77** | 0.58 | 0.76 | 0.344 | 0.331 | 0.309 | 0.307 | **0.177** | 0.241 | 0.181 |
| units_total_BF3 | 0.26* | 0.30* | 0.45* | 0.46* | **0.90** | 0.81 | 0.84 | 0.25 | 0.30 | 0.45 | 0.46 | **0.83** | 0.67 | 0.83 | 0.374 | 0.362 | 0.321 | 0.319 | **0.176** | 0.246 | 0.179 |
| units_total_BF4 | 0.33* | 0.33* | 0.42* | 0.45* | **0.86** | 0.75 | 0.77 | 0.33 | 0.33 | 0.42 | 0.44 | **0.77** | 0.63 | 0.77 | 0.380 | 0.379 | 0.353 | 0.346 | **0.223** | 0.280 | 0.222 |
| units_total_BF5 | 0.23* | 0.26* | 0.32* | 0.34* | **0.82** | 0.70 | 0.72 | 0.22 | 0.26 | 0.32 | 0.34 | **0.71** | 0.57 | 0.70 | 0.396 | 0.388 | 0.371 | 0.365 | **0.241** | 0.292 | 0.244 |
| units_total_BK1 | 0.26* | 0.30* | 0.41* | 0.42* | **0.88** | 0.80 | 0.82 | 0.26 | 0.29 | 0.40 | 0.42 | **0.81** | 0.72 | 0.81 | 0.507 | 0.494 | 0.454 | 0.449 | **0.259** | 0.308 | 0.259 |
| units_total_CB4 | 0.24* | 0.29* | 0.37* | 0.40* | **0.94** | 0.85 | 0.88 | 0.23 | 0.29 | 0.37 | 0.40 | **0.90** | 0.71 | 0.87 | 0.368 | 0.355 | 0.333 | 0.327 | **0.136** | 0.226 | 0.148 |
| units_total_CB5 | 0.18* | 0.19* | 0.23* | 0.27* | **0.66** | 0.51 | 0.54 | 0.18 | 0.19 | 0.23 | 0.27 | **0.44** | 0.42 | 0.51 | 0.528 | 0.525 | 0.512 | 0.498 | **0.435** | 0.443 | 0.407 |
| units_total_CK1 | 0.25* | 0.31* | 0.39* | 0.41* | **0.85** | 0.75 | 0.77 | 0.24 | 0.31 | 0.39 | 0.40 | **0.74** | 0.63 | 0.75 | 0.412 | 0.394 | 0.371 | 0.366 | **0.243** | 0.286 | 0.237 |
| units_total_CK2 | 0.40* | 0.43* | 0.59* | 0.60* | **0.86** | 0.76 | 0.79 | 0.40 | 0.43 | 0.59 | 0.60 | **0.77** | 0.72 | 0.78 | 0.721 | 0.702 | 0.600 | 0.590 | **0.448** | 0.491 | 0.433 |
| units_total_CK3 | 0.24* | 0.26* | 0.32* | 0.34* | **0.82** | 0.69 | 0.72 | 0.24 | 0.26 | 0.32 | 0.34 | **0.69** | 0.56 | 0.70 | 0.393 | 0.388 | 0.372 | 0.366 | **0.249** | 0.298 | 0.245 |
| units_total_DS1 | 0.34* | 0.38* | 0.41* | 0.55* | **0.80** | 0.71 | 0.72 | 0.34 | 0.38 | 0.41 | 0.54 | **0.66** | 0.65 | 0.70 | 0.609 | 0.592 | 0.575 | 0.506 | **0.439** | 0.445 | 0.408 |
| units_total_DS2 | 0.19* | 0.25* | 0.35* | 0.38* | **0.84** | 0.71 | 0.75 | 0.19 | 0.25 | 0.35 | 0.37 | **0.73** | 0.59 | 0.74 | 0.431 | 0.415 | 0.385 | 0.378 | **0.249** | 0.304 | 0.245 |
| units_total_DS3 | 0.21* | 0.23* | 0.28* | 0.32* | **0.70** | 0.56 | 0.59 | 0.21 | 0.23 | 0.28 | 0.32 | **0.52** | 0.49 | 0.56 | 0.591 | 0.586 | 0.565 | 0.549 | **0.462** | 0.473 | 0.438 |
| units_total_HB1 | 0.23* | 0.26* | 0.37* | 0.38* | **0.88** | 0.78 | 0.81 | 0.23 | 0.26 | 0.36 | 0.38 | **0.81** | 0.64 | 0.80 | 0.369 | 0.363 | 0.335 | 0.330 | **0.184** | 0.253 | 0.188 |

* The *p*-value of the *F* test was significant at $p < 2.2e-16$.

# Model Comparison

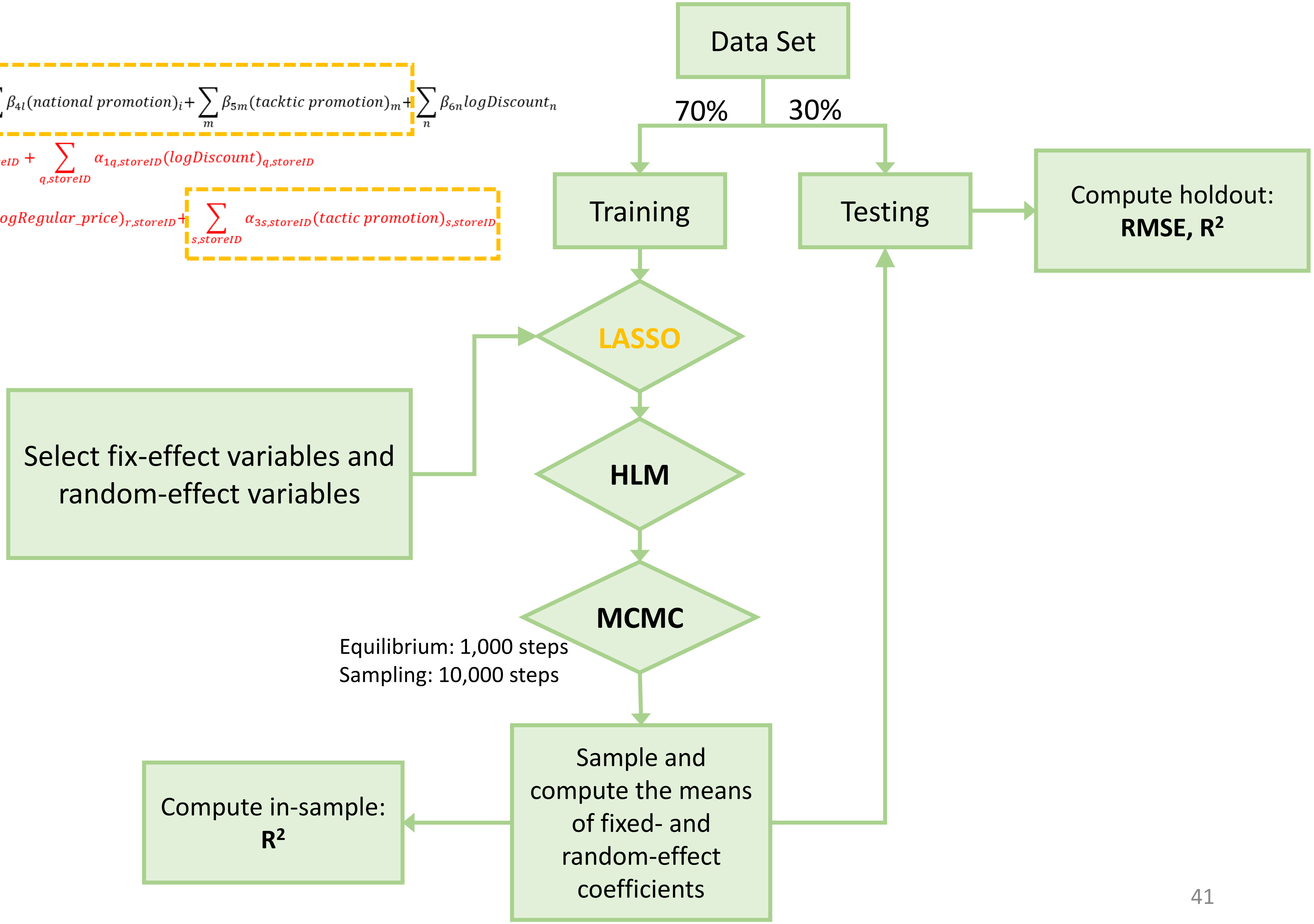
| | |
|---|---|
| **GLM 6** | $\log(Sales\ Volume)_{storeID} = \beta_{0,storeID} + \sum_i \beta_{1i,storeID}(day\ of\ week)_{i,storeID} + \sum_j \beta_{2j,storeID}(holiday)_{j,storeID} + \sum_k \beta_{3k,storeID}(weather)_{k,storeID} + \sum_l \beta_{4l,storeID}(national\ promotion)_{l,storeID} + \sum_m \beta_{5m,storeID}(tactic\ promotion)_{m,storeID} + \sum_n \beta_{6n,storeID} logDiscount_{n,storeID} + + \sum_o \beta_{7o,storeID} logRegular\_price_{o,storeID} + \sum_p \beta_{8p,storeID}(week\ of\ year)_{p,storeID}$ |
| **MCMC 6** (HLM/MMC) | $\log(Sales\ Volume) = \beta_0 + \sum_i \beta_{1i}(day\ of\ week)_i + \sum_j \beta_{2j}(holiday)_j + \sum_k \beta_{3k}(weather)_k + \sum_l \beta_{4l}(national\ promotion)_l + \sum_m \beta_{5m}(tacktic\ promotion)_m + \sum_n \beta_{6n} logDiscount_n + \sum_o \beta_{7o}(logRegular\_price)_o + \sum_p \beta_{8p}(week\ of\ year)_p + \sum_{p,storeID} \alpha_{0p,storeID} + \sum_{q,storeID} \alpha_{1q,storeID}(day\ of\ week)_{q,storeID} + \sum_{r,storeID} \alpha_{2r,storeID}(holiday)_{r,storeID} + \sum_{s,storeID} \alpha_{3s,storeID}(weather)_{s,storeID} + \sum_t \alpha_{4t,storeID}(national\ promotion)_{t,storeID} + \sum_{u,storeID} \alpha_{5u,storeID}(tactic\ promotion)_{u,storeID} + \sum_{v,storeID} \alpha_{6v,storeID}(logDiscount)_{v,storeID} + \sum_w \alpha_{7w,storeID}(logRegular\_price)_{w,storeID} + \sum_{x,storeID} \alpha_{8x,storeID}(week\ of\ year)_{x,storeID}$ <br> *3.5hrs/200 MCMC steps => 8 days/11000 MCMC steps* |

# NEXT: Model Comparison

| | |
|---|---|
| **MCMC 1** (HLM/MCMC) | $log(Sales\ Volume)$ $= \beta_0 + \sum_i \beta_{1i}(day\ of\ week)_i + \sum_j \beta_{2j}(holiday)_j + \sum_k \beta_{3k}(weather)_k + \sum_l \beta_{4l}(national\ promotion)_l + \sum_m \beta_{5m}(tactic\ promotion)_m + \sum_n \beta_{6n} logDiscount_n + \sum_o \beta_{7o} logRegular\_price_o + \sum_{p,storeID} \alpha_{0p,storeID} + \sum_{q,storeID} \alpha_{1q} logDiscount_{q,storeID}$ |
| **MCMC 2** (HLM/MCMC) | $log(Sales\ Volume)$ $= \beta_0 + \sum_i \beta_{1i}(day\ of\ week)_i + \sum_j \beta_{2j}(holiday)_j + \sum_k \beta_{3k}(weather)_k + \sum_l \beta_{4l}(national\ promotion)_l + \sum_m \beta_{5m}(tactic\ promotion)_m + \sum_n \beta_{6n} logDiscount_n + \sum_o \beta_{7o} logRegular\_price_o + \sum_p \beta_{8p}(week\ of\ year)_p + \sum_{p,storeID} \alpha_{0p,storeID} + \sum_{q,storeID} \alpha_{1q} logDiscount_{q,storeID} + \sum_{r,storeID} \alpha_{2r} logRegular\_price_{r,storeID}$ |
| **MCMC 3** (HLM/MCMC) | $log(Sales\ Volume)$ $= \beta_0 + \sum_i \beta_{1i}(day\ of\ week)_i + \sum_j \beta_{2j}(holiday)_j + \sum_k \beta_{3k}(weather)_k + \sum_l \beta_{4l}(national\ promotion)_i + \sum_m \beta_{5m}(tacktic\ promotion)_m + \sum_n \beta_{6n} logDiscount_n + \sum_o \beta_{7o}(logRegular\_price)_o + \sum_p \beta_{8p}(week\ of\ year)_p + \sum_{p,storeID} \alpha_{0p,storeID} + \sum_{q,storeID} \alpha_{1q,storeID}(logDiscount)_{q,storeID} + \sum_r \alpha_{2r,storeID}(logRegular\_price)_{r,storeID} + \sum_{s,storeID} \alpha_{3s,storeID}(tactic\ promotion)_{s,storeID}$ |

# Working Flow of Combined Hierarchical Linear Modeling/MCMC Simulation

$$\log(Sales\ Volume) = \beta_0 + \sum_i \beta_{1i}(day\ of\ week)_i + \sum_j \beta_{2j}(holiday)_j + \sum_k \beta_{3k}(weather)_k + \sum_l \beta_{4l}(national\ promotion)_i + \sum_m \beta_{5m}(tacktic\ promotion)_m + \sum_n \beta_{6n} logDiscount_n$$
$$+ \sum_o \beta_{7o}(logRegular\_price)_o + \sum_p \beta_{8p}(week\ of\ year)_p + \sum_{p,storeID} \alpha_{0p,storeID} + \sum_{q,storeID} \alpha_{1q,storeID}(logDiscount)_{q,storeID}$$
$$+ \sum_r \alpha_{2r,storeID}(logRegular\_price)_{r,storeID} + \sum_{s,storeID} \alpha_{3s,storeID}(tactic\ promotion)_{s,storeID}$$



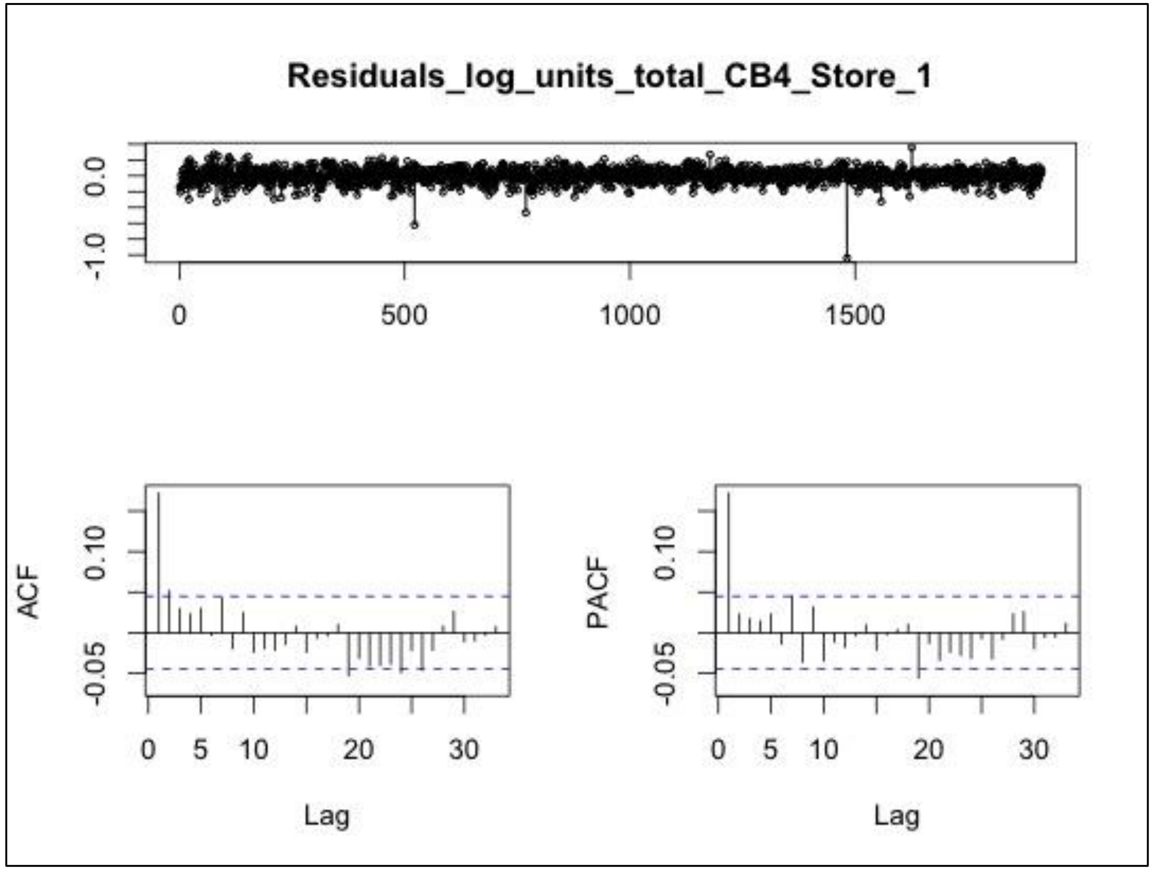
# MCMC2 has comparative performance relative to MCMC3 and is less time-consuming

| | in-sample $R^2$ | | | | | holdout $R^2$ | | | | | RMSE | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Product Name | GLM 5 | GLM 6 | MCMC 1 | MCMC 2 | MCMC 3 | GLM 5 | GLM 6 | MCMC 1 | MCMC 2 | MCMC 3 | GLM 5 | GLM 6 | MCMC 1 | MCMC 2 | MCMC 3 |
| units_total_BF2 | 0.32* | **0.86** | 0.76 | 0.78 | 0.79 | 0.32 | **0.77** | 0.58 | 0.76 | 0.76 | 0.307 | **0.177** | 0.241 | 0.181 | 0.180 |
| units_total_BF3 | 0.46* | **0.90** | 0.81 | 0.84 | 0.85 | 0.46 | **0.83** | 0.67 | 0.83 | 0.83 | 0.319 | **0.176** | 0.246 | 0.179 | 0.179 |
| units_total_BF4 | 0.45* | **0.86** | 0.75 | 0.77 | 0.79 | 0.44 | **0.77** | 0.63 | 0.77 | 0.77 | 0.346 | **0.223** | 0.280 | 0.222 | 0.222 |
| units_total_BF5 | 0.34* | **0.82** | 0.70 | 0.72 | 0.74 | 0.34 | **0.71** | 0.57 | 0.70 | 0.71 | 0.365 | **0.241** | 0.292 | 0.244 | 0.242 |
| units_total_BK1 | 0.42* | **0.88** | 0.80 | 0.82 | 0.83 | 0.42 | **0.81** | 0.72 | 0.81 | 0.81 | 0.449 | **0.259** | 0.308 | 0.259 | 0.258 |
| units_total_CB4 | 0.40* | **0.94** | 0.85 | 0.88 | 0.89 | 0.40 | **0.90** | 0.71 | 0.87 | 0.87 | 0.327 | **0.136** | 0.226 | 0.148 | 0.149 |
| units_total_CB5 | 0.27* | **0.66** | 0.51 | 0.54 | 0.56 | 0.27 | **0.44** | 0.42 | 0.51 | 0.51 | 0.498 | **0.435** | 0.443 | 0.407 | 0.406 |
| units_total_CK1 | 0.41* | **0.85** | 0.75 | 0.77 | 0.78 | 0.40 | **0.74** | 0.63 | 0.75 | 0.76 | 0.366 | **0.243** | 0.286 | 0.237 | 0.234 |
| units_total_CK2 | 0.60* | **0.86** | 0.76 | 0.79 | 0.81 | 0.60 | **0.77** | 0.72 | 0.78 | 0.79 | 0.590 | **0.448** | 0.491 | 0.433 | 0.430 |
| units_total_CK3 | 0.34* | **0.82** | 0.69 | 0.72 | 0.73 | 0.34 | **0.69** | 0.56 | 0.70 | 0.71 | 0.366 | **0.249** | 0.298 | 0.245 | 0.243 |
| units_total_DS1 | 0.55* | **0.80** | 0.71 | 0.72 | 0.73 | 0.54 | **0.66** | 0.65 | 0.70 | 0.70 | 0.506 | **0.439** | 0.445 | 0.408 | 0.409 |
| units_total_DS2 | 0.38* | **0.84** | 0.71 | 0.75 | 0.77 | 0.37 | **0.73** | 0.59 | 0.74 | 0.75 | 0.378 | **0.249** | 0.304 | 0.245 | 0.237 |
| units_total_DS3 | 0.32* | **0.70** | 0.56 | 0.59 | 0.61 | 0.32 | **0.52** | 0.49 | 0.56 | 0.56 | 0.549 | **0.462** | 0.473 | 0.438 | 0.441 |
| units_total_HB1 | 0.38* | **0.88** | 0.78 | 0.81 | 0.82 | 0.38 | **0.81** | 0.64 | 0.80 | 0.80 | 0.330 | **0.184** | 0.253 | 0.188 | 0.190 |

* The *p*-value of the *F* test was significant at $p < 2.2e-16$.

# Remaining Issues I: Residuals of both models (GLM6 and MCM2) show seasonal patterns

Example: Residuals, ACF, and PACF of the linear model for the prediction of product CB4 at store 1 (GLM6)




The residuals of linear models GLM6 show strong autocorrelation. So linear assumption of independence of residuals is not valid in these models, which may lead to the wrong coefficients.

# **Remaining Issues II:** Both models have a large percentage of coefficients with incorrect signs. MCMC2, however, has zero instances of N/A's

## Price Elasticity: Incorrect vs. Correct Sign

### GLM 6

| | BF2 | BF3 | BF4 | BF5 | BK1 | CB4 | CB5 | CK1 | CK2 | CK3 | DS1 | DS2 | DS3 | HB1 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| >0 (Incorrect) | 24 | 21 | 63 | 37 | 40 | 12 | 54 | 54 | 8 | 44 | 14 | 48 | 34 | 1 | **454** |
| <0 (Correct) | 43 | 74 | 32 | 47 | 53 | 27 | 39 | 34 | 86 | 47 | 77 | 45 | 43 | 91 | **738** |
| N/A | 29 | 1 | 1 | 12 | 3 | 57 | 3 | 8 | 2 | 5 | 5 | 3 | 19 | 4 | **152** |
| % With Incorrect Sign | 35.8% | 22.1% | 66.3% | 44.0% | 43.0% | 30.8% | 58.1% | 61.4% | 8.5% | 48.4% | 15.4% | 51.6% | 44.2% | 1.1% | **29.1%** |

### HLM MCMC2

| | BF2 | BF3 | BF4 | BF5 | BK1 | CB4 | CB5 | CK1 | CK2 | CK3 | DS1 | DS2 | DS3 | HB1 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| >0 (Incorrect) | 64 | 20 | 55 | 47 | 45 | 28 | 62 | 46 | 28 | 43 | 7 | 29 | 25 | 0 | **499** |
| <0 (Correct) | 32 | 76 | 41 | 49 | 51 | 68 | 34 | 50 | 68 | 53 | 89 | 67 | 71 | 96 | **845** |
| N/A | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **0** |
| % With Incorrect Sign | 66.7% | 20.8% | 57.3% | 49.0% | 46.9% | 29.2% | 64.6% | 47.9% | 29.2% | 44.8% | 7.3% | 30.2% | 26.0% | 0.0% | **31.4%** |

# Next

- **For Issue I:** Diagnosing the seasonality pattern in the dependent variables
  Deseasonalization by the means of seasonal adjustment

- **For Issue II:** Perform constrained optimization using non negative least squares (different from Lasso)
  Perform best subset regression

# Appendix

# Working Flow of Validation Process

All variables

Generate **GLM** model with **ALL** the independent variables

Detect important **interactions** using **Regression Trees**

Criteria for Selection of Interactions:

- Prune the tree using the CP value with the minimum error.
- Frequency of child-parent pair ≥ 10
- 10-fold CV was applied to select #interactions (having the lowest RMSE) put in the model

Generate model formula which contains all the independent variables and the selected interaction terms

Repeat 100 times

Data Set

70% 30%

Training

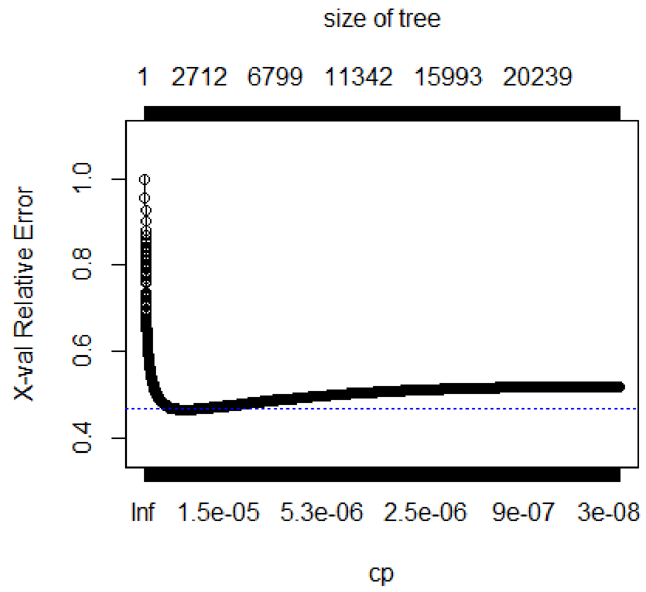
Testing

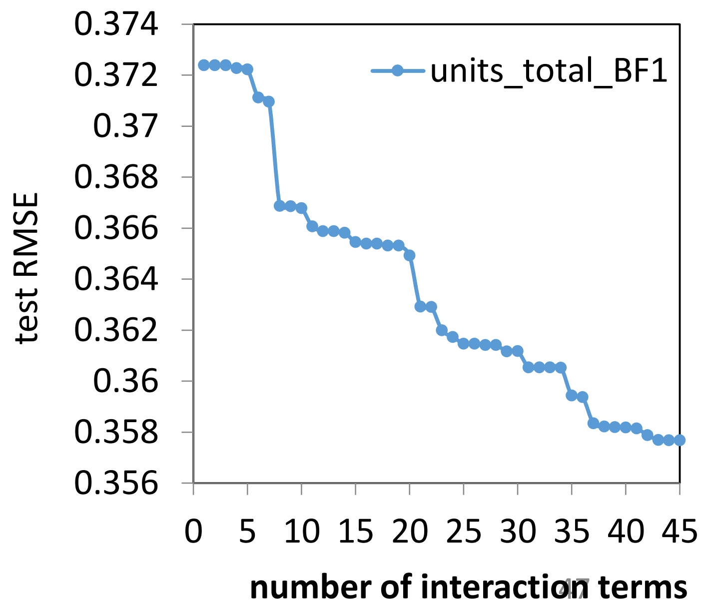
GLM

Compute: **R2, AIC, BIC, Coefficients**

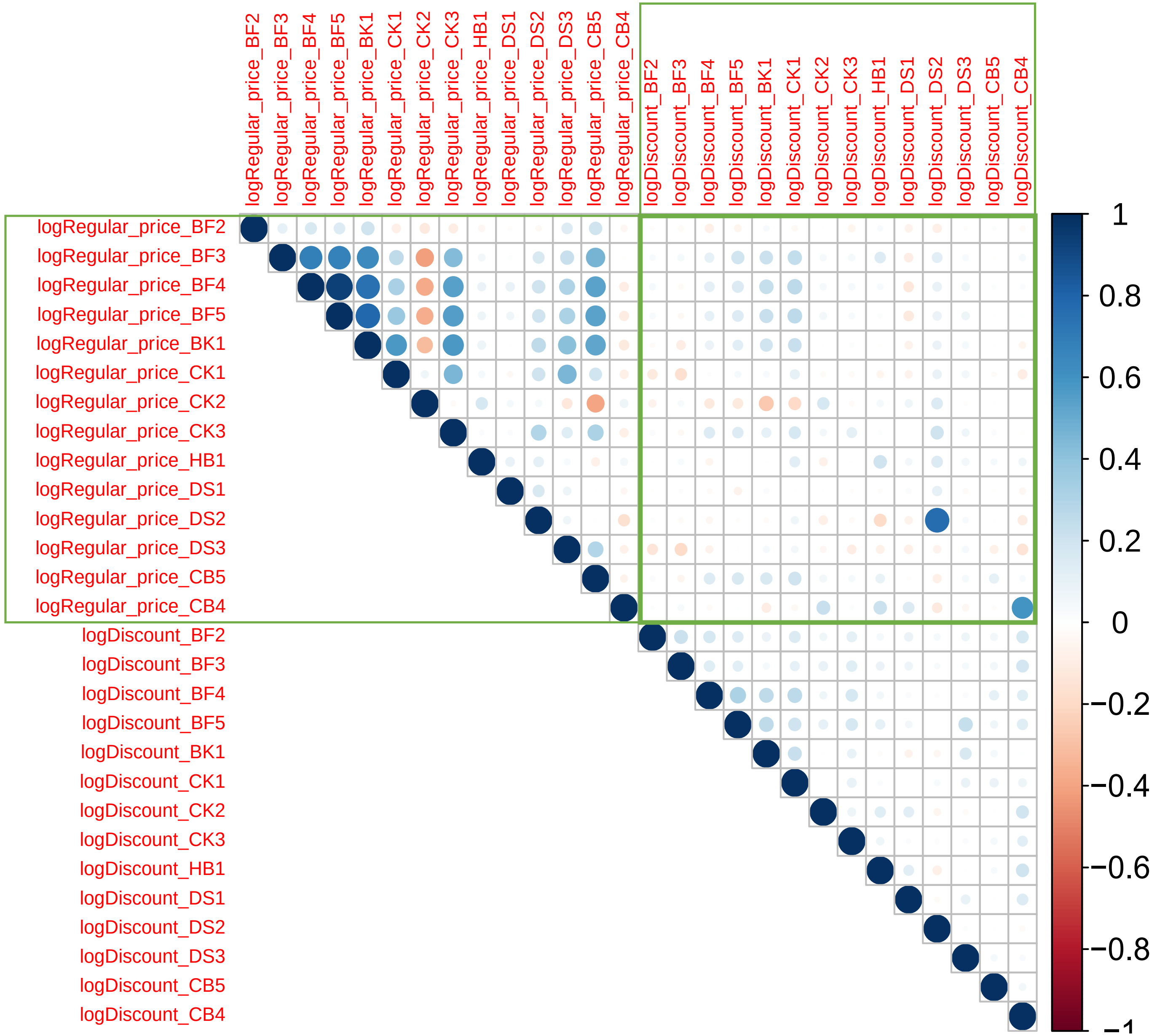
Compute: **RMSE**

**Compute the MEAN and SD of**: R2, AIC, BIC, RMSE, Coefficients

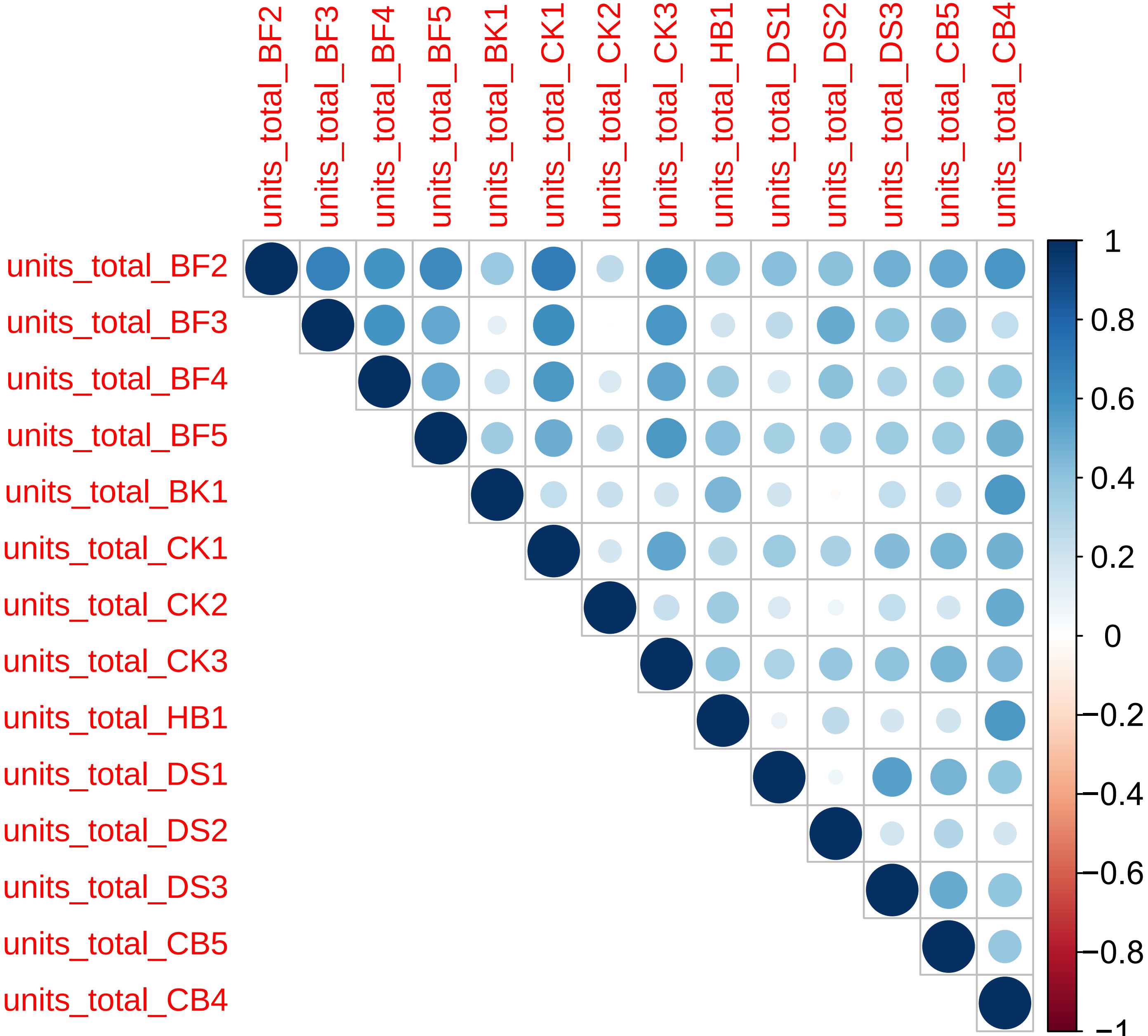


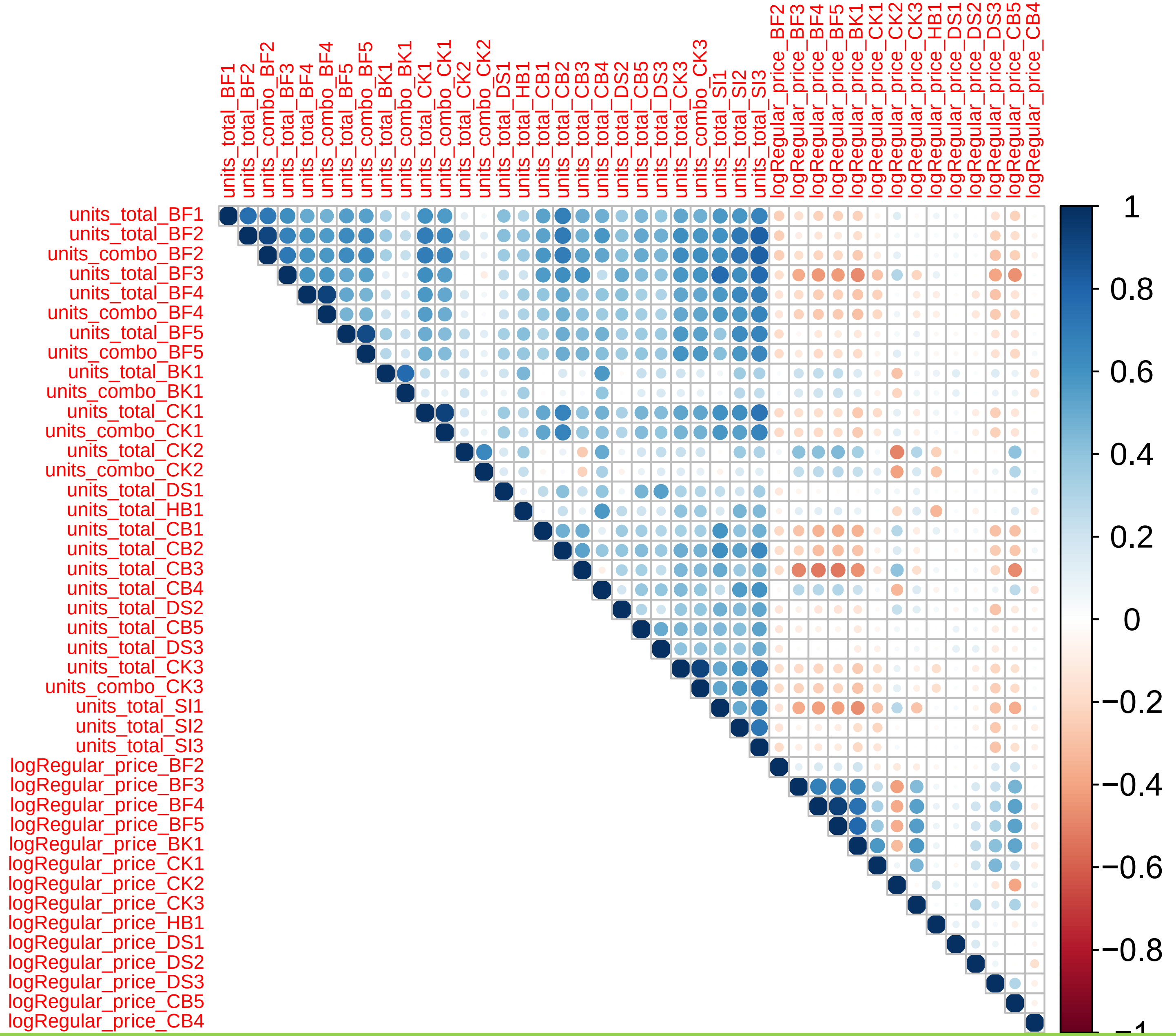
# Correlation matrix of log regular prices and log Discount prices



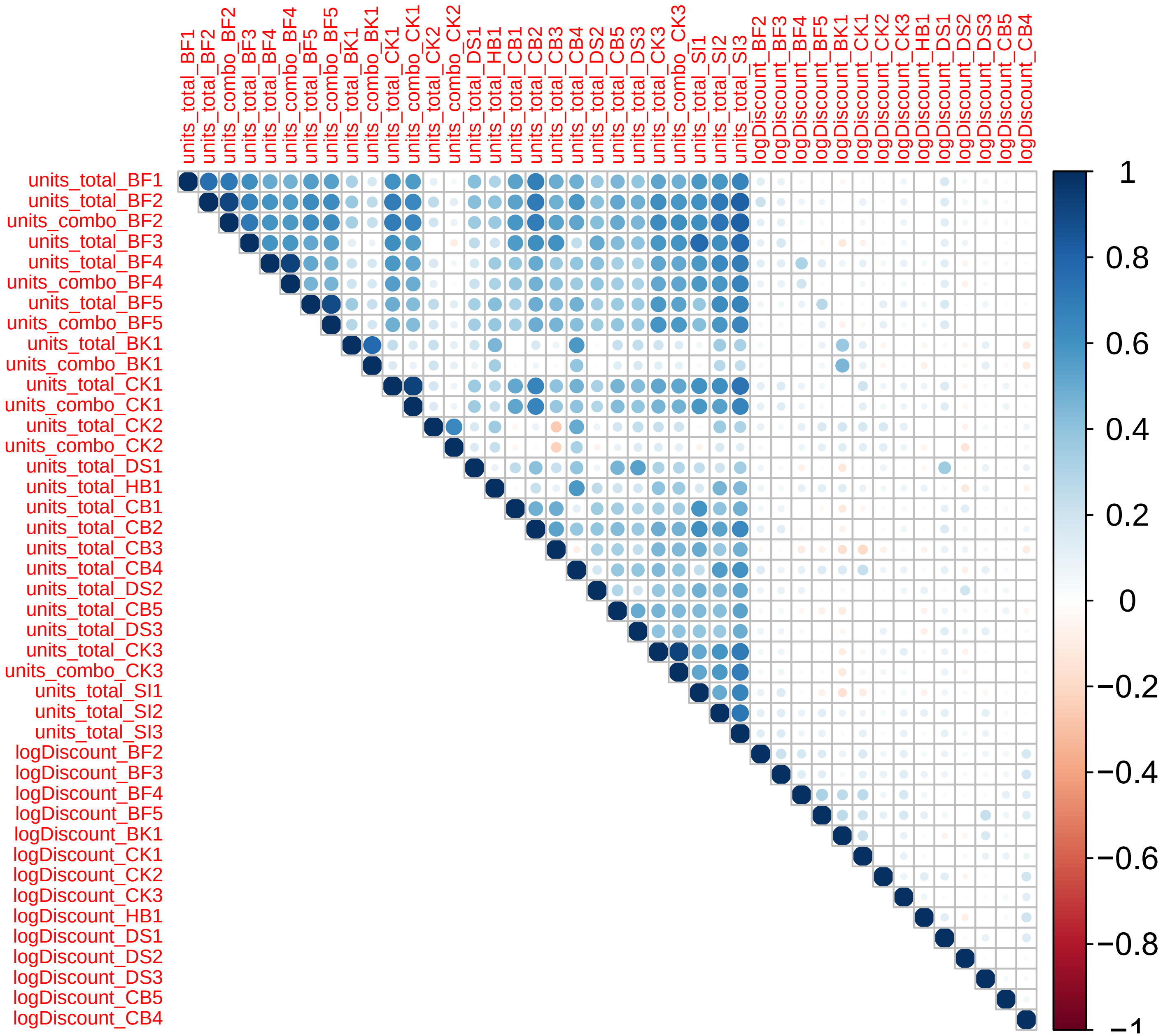
# Correlation matrix of 14 Dependent Variables for modeling



# Correlation matrix of log units and log regular prices



# Correlation matrix of log units and log Discount prices



# Correlation matrix of log Units, log regular prices and log Discount prices