

MEAL PARTICIPATION PREDICTION WITH BAYESIAN HIERARCHICAL MODELS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
MASTER OF SCIENCE
IN
INDUSTRIAL ENGINEERING

By
Aleyna Kof
December 2021

Meal Participation Prediction with Bayesian Hierarchical Models

By Aleyna Kof

December 2021

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.



Savaş Dayanık(Advisor)

Özlem Çavuş İyigün

Mustafa Gökçe Baydoğan

Approved for the Graduate School of Engineering and Science:

Ezhan Karaşan
Director of the Graduate School

ABSTRACT

MEAL PARTICIPATION PREDICTION WITH BAYESIAN HIERARCHICAL MODELS

Aleyna Kof

M.S. in Industrial Engineering

Advisor: Savaş Dayanık

December 2021

Forecasting sales in the catering industry helps authorities to organize daily transactions efficiently to prevent both waste and business loss. In this study, we focused on predicting meal sales in Bilintur Catering Centre with the dataset which is collected through five academic years. To forecast the meal sales, we constructed two Bayesian hierarchical models. The first model does not differentiate effects of predictors in different academic years, while the second does. We derived the full conditional distributions and employed Gibbs sampling in an extensive MCMC study. We tested two models along with a benchmark multiple regression model on the held-out academic year. We concluded that multiple regression and first model provide more accurate results.

Keywords: Bayesian Hierarchical Models, Gibbs Sampling, Markov Chain Monte Carlo, forecasting, catering.

ÖZET

TÜRKÇE BAŞLIK

Aleyna Kof
Endüstri Mühendisliği, Yüksek Lisans
Tez Danışmanı: Savaş Dayanık
Aralık 2021

Yemek endüstrisinde satışları tahmin etmek, hem israfı hem de iş kaybını önleyerek yetkililerin günlük transferleri verimli bir şekilde organize etmesine yardımcı olur. Bu çalışmada, beş akademik yıl boyunca Bilintur Tabldot Yemek Merkezi (BCC-Bilintur Catering Centre)’nden elde edilen veri seti ile yemek satışlarını tahmin etmeye odaklanılmıştır. Yemek satışlarını tahmin etmek için iki tane bayesci hiyerarşik model oluşturulmuştur. İlk model, farklı akademik yıllarda tahminleyicilerin etkilerini farklılaştırmazken, ikincisi bunu yapar. Kapsamlı bir MCMC çalışmasında tam koşullu dağılımları elde edilmiştir ve Gibbs örnekleme kullanılmıştır. Test seti olarak ayrılan akademik yılda iki modeli bir karşılaştırmalı çoklu regresyon modeliyle birlikte test edilmiştir. Sonuç olarak bazı tahminleyicilerin etkilerinin zamanla değişmediği, Bayesci hiyerarşik modellerinin kullanımındansa çoklu regresyon modeli ya da ilk modelin daha doğru ve istikrarlı tahminler verdiği çıkarılmıştır.

Anahtar sözcükler: Hiyerarşik Bayesci modeller, Gibbs örneklendirmesi, Markov Zinciri Monte Carlo, önkestirim, yemek hizmetleri.

Acknowledgement

First and foremost, I would like to express my deepest gratitude for my advisor Prof. Savaş Dayanık. His mentorship, patience and expertise guided me throughout my studies. This work would not be possible without his guidance.

I would like to thank Asst. Prof. Özlem Çavuş İyigün and Asst. Prof. Mustafa Gökçe Baydoğan for accepting to be in my thesis committee.

I would like to express my sincere gratitude for my family. Their endless love and support have given me the strength to become the person I am today.

I also would like to express my appreciation for the supportive and nurturing environment that Bilkent University provided throughout my academic journey. I would like to thank all my professors, colleagues and friends for creating a community that I am so proud to be a part of.

Contents

1	Introduction	1
2	Dataset	4
2.1	Exploring Dataset	9
2.1.1	Examining Date Effect on Sales	9
2.1.2	Examining Sales with Menu Ingredients	11
2.1.3	Examining Sales with Events	13
2.2	Dataset Pre-processing	13
3	Literature Review	16
4	Linear Regression Model	24
4.1	Variable Selection	24
4.2	Diagnostics	34
4.3	Prediction	52

5	Single Layer Hierarchical Model	54
5.1	Convergence Analysis	56
5.2	Diagnostics	59
6	Two Layers Bayesian Regression Model	62
6.1	Convergence Analysis	65
6.2	Posterior Distributions	69
7	Forecasting	70
8	Conclusion	79
A	Derivations of Full Conditional Distributions	80
A.1	Variances and Covariance Matrices for Single Layer Model	80
A.2	Covariates' Effects for Single Layer Model	81
A.3	Mean Covariates' Effects for Single Layer Model	83

List of Figures

2.1	Daily number of fix lunch servings sold in different academic years	9
2.2	Number of fix lunch sold through semesters	10
2.3	Number of fix lunch being sold for summer semesters	11
2.4	Menu ingredients compared with the days of a Week	12
2.5	Sales changes regarding to registration events	13
2.6	Fix lunch density	14
2.7	Fix lunch density for weekdays and weekends	15
4.1	Diagnostic plots for the model with formula	26
4.2	Diagnostics after log-transformation is applied	27
4.3	Partial residual plots for quantitative variables with current linear model	29
4.4	Partial residual plots for quantitative variables after log transformation with current linear model	30

4.5	Partial residual plots after PrevFixLunch ² is included to current linear regression model	33
4.6	t-statistics. The white thick vertical lines mark -2 and +2 values. The bars show the t-statistics for only those predictors whose t-statistics are greater than two in absolute value. The second plot is with sorted values	34
4.7	Diagnostic plots for linear regression model	36
4.8	Diagnostic plots show influence measures calculated after each observation (x-axis) is removed	37
4.9	Studentized residual distributions conditional on semester. Dates for extreme points are expressed as yy.m.d here and later.	38
4.10	Studentized residual distributions conditional on Ramadan or holiday	39
4.11	Studentized residual distributions conditional on week day	40
4.12	Studentized residual distributions conditional on week	41
4.13	Studentized residual distributions conditional on week part	42
4.14	Studentized residual distributions conditional on semester part	43
4.15	Studentized residual distributions conditional on semester week	44
4.16	Studentized residual distributions conditional on temperature	45
4.17	Studentized residuals vs fix lunch servings on previous two days	46
4.18	Studentized residuals vs temperature and fitted values	47
4.19	Q-Q plot with studentized residuals	49

4.20	Influential points with dates. 7 cases (<i>12.12.31</i> , <i>13.8.7</i> , <i>15.3.18</i> , <i>15.7.16</i> , <i>15.9.23</i> , <i>16.5.19</i> and <i>16.7.4</i>) have leverage equal to one. Plot has to be created without them because studentized residual is undefined when leverage equals one.	50
4.21	DFBETAS for different effects	51
4.22	Log and normal scale estimated vs observed values	53
5.1	DAG for hierarchical linear model	55
5.2	Trace plots for coefficients of year	57
5.3	Trace plot for variance of coefficients	58
5.4	R-hat statistics for variance and all covariates with different interval lengths	59
5.5	Observed vs predicted on train data	60
5.6	t-statistics for single layer model. The white thick vertical lines mark -2 and +2 values. The bars show the t-statistics for only those predictors whose t-statistics are greater than two in absolute value.	61
6.1	DAG for multilevel hierarchical bayesian regression model	63
6.2	Trace plots of <code>FixLunchhas.kof1</code> (meatball is not offered), <code>DAYSTOSTART1</code> (1 day is left to beginning of the academic year) and variance of coefficients	66
6.3	t-statistics for two layers model for first chain and first academic year. The white thick vertical lines mark -2 and +2 values. The bars show the t-statistics for only those predictors whose t-statistics are greater than two in absolute value.	67

6.4	Gelman-Rubin statistic for different intervals for two layer bayesian regression	68
6.5	DAYSTOSTART5 density for through years under single layer model and two layers model respectively	69
7.1	Fix lunch sales vs predicted sales for single layer model on test data	71
7.2	Variance values through Academic Years	72
7.3	Sampling covariates of with last year variance	73
7.4	Sampling covariates of with mean of last two years variance	74

List of Tables

2.1	List of ingredients	5
2.2	Description of covariates	6
2.3	Lagging variables	9
3.1	U-statistic	18
3.2	$\hat{R}$ -statistic	22
4.1	AIC and BIC values for different models	31
4.2	T test results for quadratic effect	48
7.1	Measures of accuracy comparison for different prediction techniques	75
7.2	The comparisons of linear regression, single layer and two layer models updated every four week block of 2016-2017 and tested on the next four week block	76

Chapter 1

Introduction

This thesis introduces a methodology to forecast the meal participation of the university cafeteria. Meals are distributed three times which are breakfast, lunch and dinner in a day at cafeteria. Lunch times are the most crowded times of the cafeteria compared to others. Therefore, this study aims to develop a forecasting method for the meal participation during lunch times.

Forecasting accurately are highly beneficial in every industry because of that it contributes to production and scheduling decisions especially with resource allocation and cost control. Because of short shelf life of foods, precise forecasts are significant in the food related industries so if overproduced, it is highly to result in economic loss. However under forecasting is risky as well as over forecasting. With this approach, forecasting for transactions of Bilintur Catering Centre (BCC) is studied. BCC is an establishment of food service which operates in Bilkent University dining halls. These dining halls are followed with their respecting located campus: Marmara Restaurant located in Main Campus, Sultan Restaurant located in East Campus and lastly Table'd Hote Cafeteria located in Middle Campus in Bilkent University.

Major concern of BCC for forecasting is predicting of meal participation due

to inconstant demands, student population and regulations. BCC currently forecast by using last year's sales as a basis. Apart from this, weather conditions, personal experiences of the officials, parking occupancy are considered during forecasting processes. Considering the increasing student population in campus and insufficient capacity of current dining halls, Marmara restaurant opened in a new building in March 2014. After the opening of new dining hall, BCC's requirement for systematic approach in forecasting sales comes into focus.

Weak forecasting results lead to either over or under production of respecting food. This situation results in poor satisfaction of customers, resources waste and capital loss. If over forecasting occurs, staff for serving more than needed is assigned which is an issue of human capital loss. For the aspect of food, the resources which can be served in other times are wasted. In the case of under forecasting, the problem is mainly the shortage of meals to be served. If shortage of meals occurs, the authorities are likely to serve other solutions since regulatory actions are to be faced in the scope of unsatisfied demand amount. Other solutions to replace of unmet demand includes changing the menu items which can be with frozen of other available items and this results in immediate integration of stock out elements. It also leads to detracting the credibility of the establishment because of changing the menu in the absence of announcement. Therefore, constructing an accurate forecasting approach is an important element of the establishment since it has a critical impact on both resource allocation and customer satisfaction.

To come up with a forecasting model with small error rates, different approaches were used. Firstly, the number of fixed lunch servings is determined as response variable and the parameters which are probably related with the response variable are studied. The relation between the parameters and response variable is checked by analysis of variance and as a result variable selection is applied. Linear regression is applied between response and selected parameters. Later, output analysis is performed and error rates were observed. Secondly, to capture the heterogeneity in the dataset, bayesian hierarchical model is applied to dataset. The hierarchical model is aimed to allow effects of the same variables to change in different academic years. After developing this model, Gibbs

sampling is used to obtain the posterior distribution of effects in the hierarchical model. The remaining part of this study is organized as follows. In chapter 2, we explore data set using graphical analysis. In last section of this chapter, we present pre-processing steps. In chapter 3, we reviewed the literature and explain related papers such as forecasting in catering industry and hierarchical bayesian model studies. In chapter 4, hierarchical bayesian models and the full conditionals distributions are explained. In chapter 5, posterior distributions via Gibbs sampling and convergence analysis are examined. In chapter 6, forecasting studies are performed on test data and results are discussed. Lastly, we concluded by discussing the results and the models' performances.

Chapter 2

Dataset

The data being used in this study were collected by Bilintur Catering Centre (BCC). BCC is an institutional and industrial catering service provider and held accountable for the catering services at university campus. BCC provides catering services at three different locations in campus which are Marmara Restaurant in main campus, Sultan Restaurant in east campus, Table'd Hote Restaurant in middle campus. BCC keeps track of each meal purchase instance by recording the students' ID information, the place and time of purchase in a remote database. Therefore, the data enclose the total number of meal sold per meal at different locations where BCC is on service. The data set contains the number of meal purchased from these restaurants during five academic years and arranged as tabular data. Since this study aims to predict total number of sold meal servings, the data are summarized as the daily count of meal purchases. Moreover, menu information are collected from BCC meal plans and cafeteria regulations including sufficient calorie intake.

Other than information provided by BCC, we benefit from academic calendars available online. The information in the academic calendar are embedded to previously mentioned data set to gain better insight about students' eating preferences under distinctive events. After combining these datasets, information are extracted including registration events, menu ingredients, exams, holidays

etc. As a result, we formed data set with 102 columns and 1803 rows. Each row represents each day during five academic years while each column encloses the information about the events in these days. The covariate variables can mainly be described in three categories. The covariates in first category describe the menu ingredients of that day and are summarized as in Table 2.1

Table 2.1: List of ingredients

Covariate	Description	Examples	Levels
has fish	Does the menu have fish in it?	mezgit, alabalık, uskumru	Logical: TRUE or FALSE
has wmeat	Does the menu have white meat in it?	piliç, tavuk, hindi	Logical: TRUE or FALSE
has rmeat	Does the menu have red meat in it?	piliç, tavuk, hindi	Logical: TRUE or FALSE
has veg	Does the menu have only vegetables in it?	köfte, kebab, güveç, dana, et güveç, kavurma, et sote, elbasan tava, ankara tava, macar gulas	Logical: TRUE or FALSE

(continued on the next page)

(continued from the previous page)

Covariate	Description	Examples	Levels
has pulse	Does the menu have pulse?	mercimek (soup is excluded), no-hut, fasulye	Logical: TRUE or FALSE
has dolma	Does the menu have dolma?	kabak/biber dolma, etli biber dolma	Logical: TRUE or FALSE
has kof	Does the menu have kofte?	madalyon köfte, sumaklı köfte, belen köfte	Logical: TRUE or FALSE

The second category covariates describe the events of that day ranging from weather to registration events in Table 2.2.

Table 2.2: Description of covariates

Covariate	Description and Levels
Date	(The date of the transaction is made) Date object
Acadyr	(Academic year) Factor with 5 levels
Sem	(Semester) Factor with 6 levels FD:fall semester, FFB:fall final and break, SD:spring semester, SFB:spring final and break, SUD:summer semester, SUFB:summer final and break
Part	(The part of any semester with equal splits) Factor with 4 levels: 1: First part, 2: Second part, 3: Third part, 4: Fourth part splitting the semester into equal parts

(continued on the next page)

(continued from the previous page)

Covariate	Description and Levels
Sempart	(The part of the current semester) Factor with 15 levels FB01: First part of fall break, FD01: First part of fall semester, FD02: Second part of fall semester, FD03: Third part of fall semester, FD04: Fourth part of fall semester, FF01: First part of fall final exam period, SB01: First part of spring break, SD01: First part of spring semester, SD02: Second part of spring semester, SD03: Third part of spring semester, SF01: First part of spring final exam period, SUB01: First part of fall break, SUD01: First part of summer semester, SUD02: Second part of summer semester, SUF01: First part of summer final exam period
Semweek	(The week of the current semester) Factor with 14 levels: FD01, ..., FD14 and 2,3 or 5 weeks for semester breaks depending on the semester
Week	(The week number based on academic calendar) Factor with 14 levels: 1: First week, 2: Second week,... 14: Fourteenth week
Rmdn	(Whether it is in Ramadan month) Factor with 2 levels: NR: Not Ramadan, R: Ramadan
Wday	(Day of Week) Factor with 7 levels: 1: Monday, 2: Tuesday, 3: Wednesday, 4: Thursday, 5: Friday, 6: Saturday, 7: Sunday
Holiday	(Holiday different from weekends) Factor with 5 levels H: Holiday, H05: Half-day holiday, HEPnU: Holiday for preparatory school and non-holiday for departments, HUnEP: Holiday for departments and non-holiday for preparatory school, NH: Not Holiday
Temp	(Temperature for the day in C) continuous

(continued on the next page)

(continued from the previous page)

Covariate	Description and Levels
Precip	(Amount of rainfall in liters) continuous
Events	(Weather events) Factor with 3 levels: none: No weather event, rain: rainy, snow: snowy, both: rainy and snowy
COPEcope1	(Whether the first stage of English proficiency exam is held on that day or not) Logical
COPEcope2	(Whether the second stage of English proficiency exam is held on that day or not) Logical
Daystostart	(The number of days being left to the beginning of a new semester) Factor with 7 levels: 1,...5,6 or more
GE100	(GE100 Introduction to Academic Life program is held on that day) Factor with 2 levels: NO and YES
Hldytogo	(The number of days being left until the end of the holiday) Ordered factor with levels $1 < 2 < \dots < 9$
Pdaystostart	(The number of days being left to begin a new semester in preparatory school) Factor with 5 levels: 1,...,3,4 or more
REG	(Registration events) Factor with 5 levels: enroll: Registration day for new students, exc: Registration day for exchange students, int: Registration day for full-time international students, reg: Registration day for other students
LastDayCope1	(Whether it is the last day of first stage of English proficiency exam) Logical
LastDayCope2	(Whether it is the last day of second stage of English proficiency exam) Logical

Lastly, there are lagging variables such as previous week's/day's fix menu sales to explain time series effect as shown in Table 2.3.

Table 2.3: Lagging variables

Covariate	Description and Levels
Prev	(The number of meals purchased in previous day) Continuous
Prev2	(The number of meals purchased in two days before) Continuous
Weekago	(The number of meals purchased a week ago) Continuous

2.1 Exploring Dataset

Following data collection and transformation, we used graphical analysis to investigate the potential effects of variables on sales, with the findings helping us to create a sensible hierarchical model having multiple predictors.

2.1.1 Examining Date Effect on Sales

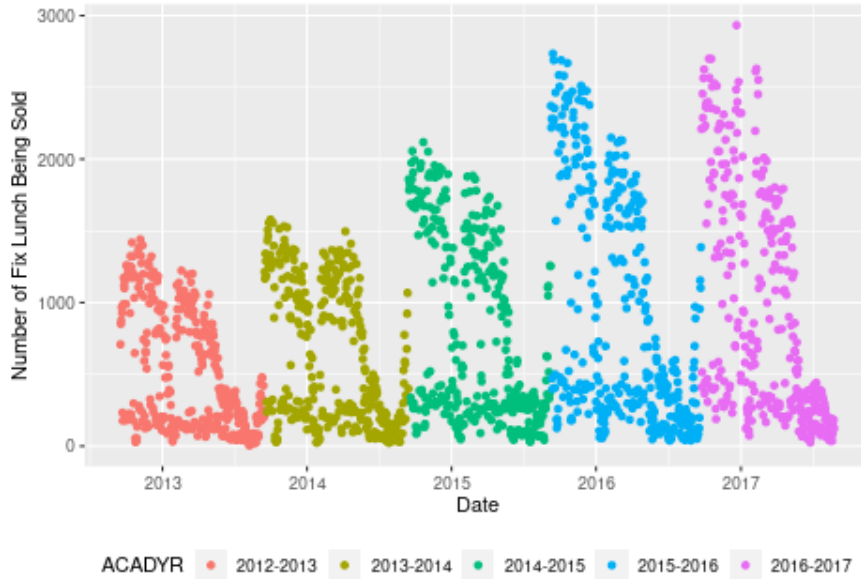


Figure 2.1: Daily number of fix lunch servings sold in different academic years

The number of fix lunch being sold has increased during academic years. One can deduce from Figure 2.1 that the sales have a tendency to increase at the beginning of each year and decrease towards the middle of that year and increase again at the very end of that year. To identify this pattern, we first start investigating the semester effect.

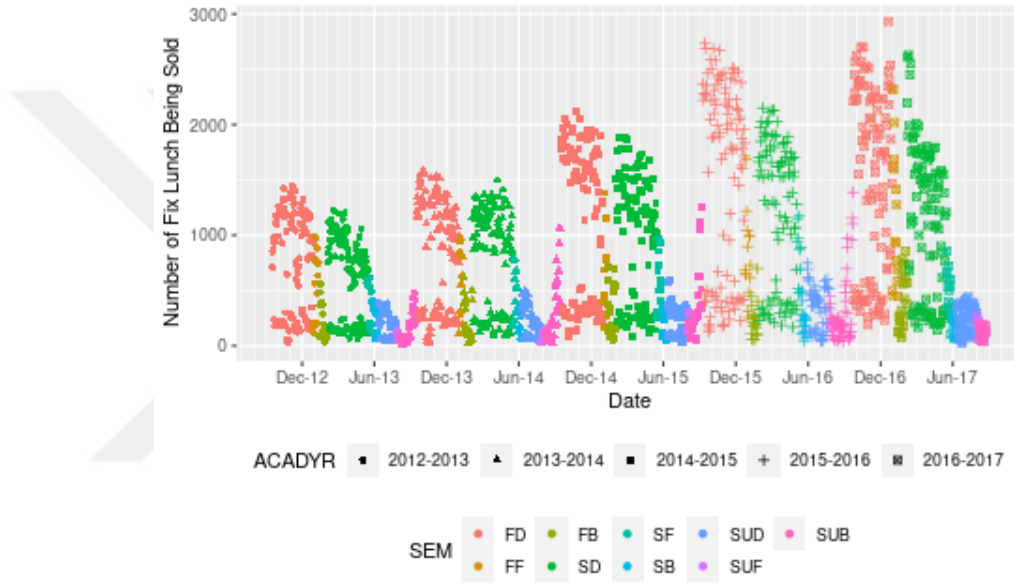


Figure 2.2: Number of fix lunch sold through semesters

One can deduce from Figure 2.2 that the number of meal sold is likely to increase at the beginning of fall and spring semesters (FD and SD respectively). Moreover, it has a tendency to decrease towards the end of these semesters and is affected by a sharp decrease during summer semester (SUD) since the number of students being enrolled for summer school is low compared to fall and spring semesters. Additionally, Figure 2.3 shows during summer breaks (SUB), number of sales raises compared to the summer semester and probable reasons for the increments might be course registration, early start of preparatory school, orientation events and PAE exam. Within each semester, meal purchases are high during the week and low during weekends. The purchases tend to decrease as the semester is coming to its end.

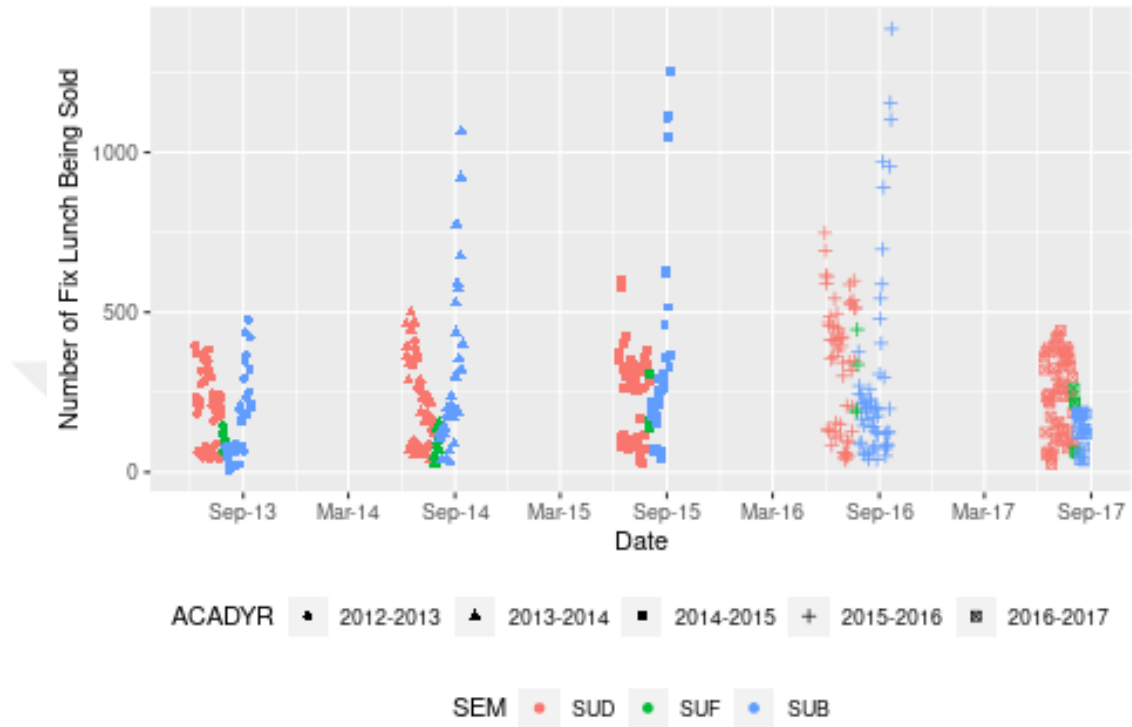


Figure 2.3: Number of fix lunch being sold for summer semesters

2.1.2 Examining Sales with Menu Ingredients

After examining the date-related features, we considered that the students' eating preferences might be affected by the menu on offer.

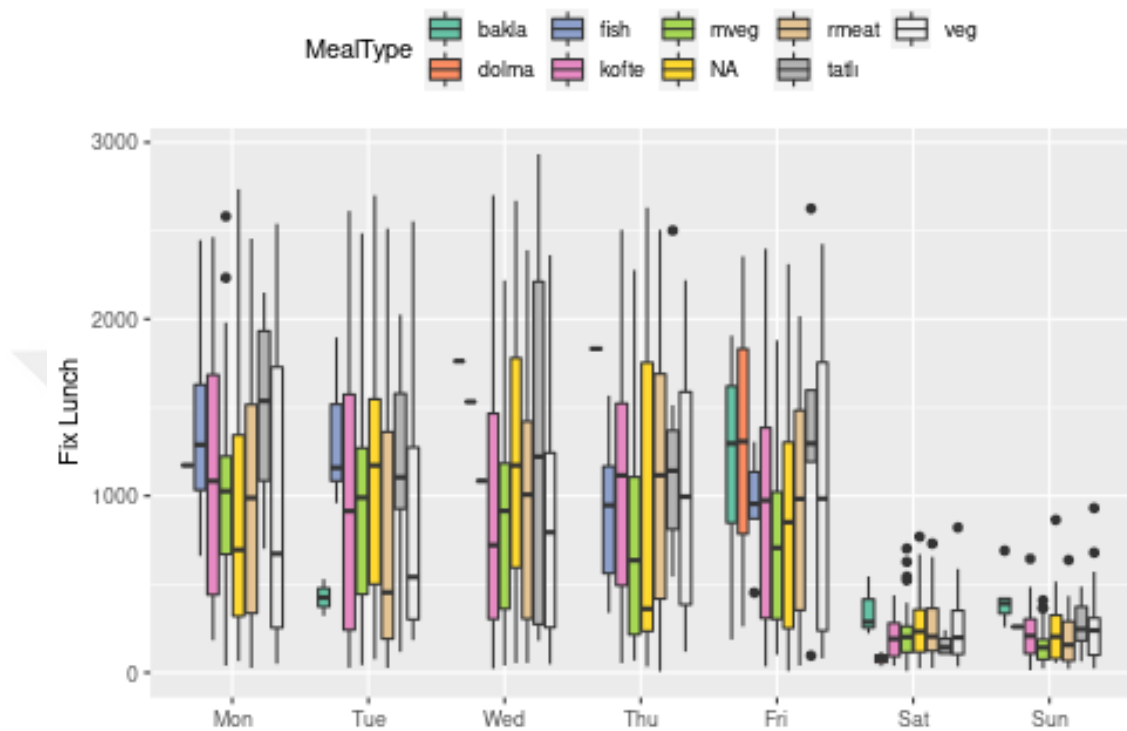


Figure 2.4: Menu ingredients compared with the days of a Week

Figure 2.4 shows certain ingredients are relatively preferred by the students more than others. To exemplify, fish are favored more than pulse. However, this depends on the weekday and semester as well.

2.1.3 Examining Sales with Events

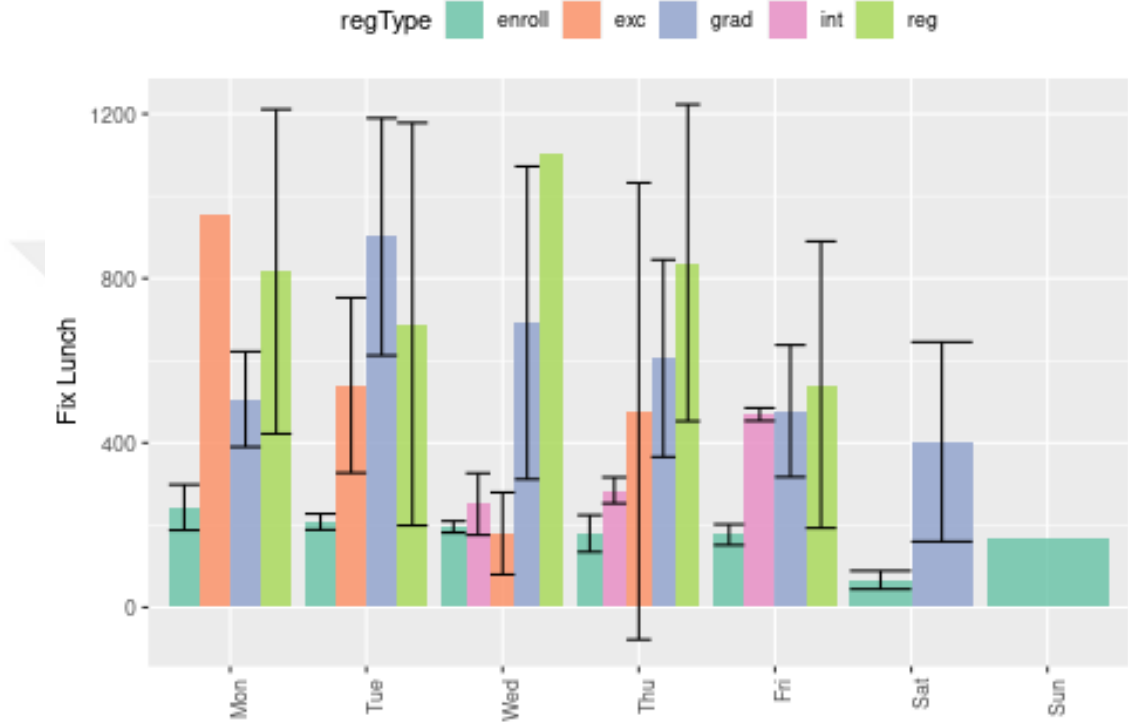


Figure 2.5: Sales changes regarding to registration events

In Figure 2.5, the changes regarding the registration events are indicated. The sales address similar trends as considering the different registration events. Some registration events only have instance in data, therefore, it is not possible to calculate lower and upper bounds for these events. To exemplify, registration of new students occurs on Sunday only once, therefore, standard deviation for this event is zero.

2.2 Dataset Pre-processing

Before explaining our model, it is important to mention pre-processing steps. As it was stated before, BCC has been collected data for different purposes. The raw

data contain each student's ID number, time and place of the sales and menu list is available as a separate source. Since the purpose of this study is to forecast the total number of fix lunch sales, we used grouped data as a starting point instead of using individual level data [gulin]. After obtaining data, feature engineering steps are applied to extract relevant features from data such as menu ingredients, registration or orientation events etc.

We represented factors with dummy variables using sum contrast. All variables are centered at their median and scaled with their median absolute deviation (MAD) which is a robust measure of variability. Figure 2.6 shows the density plot for fix lunch servings.

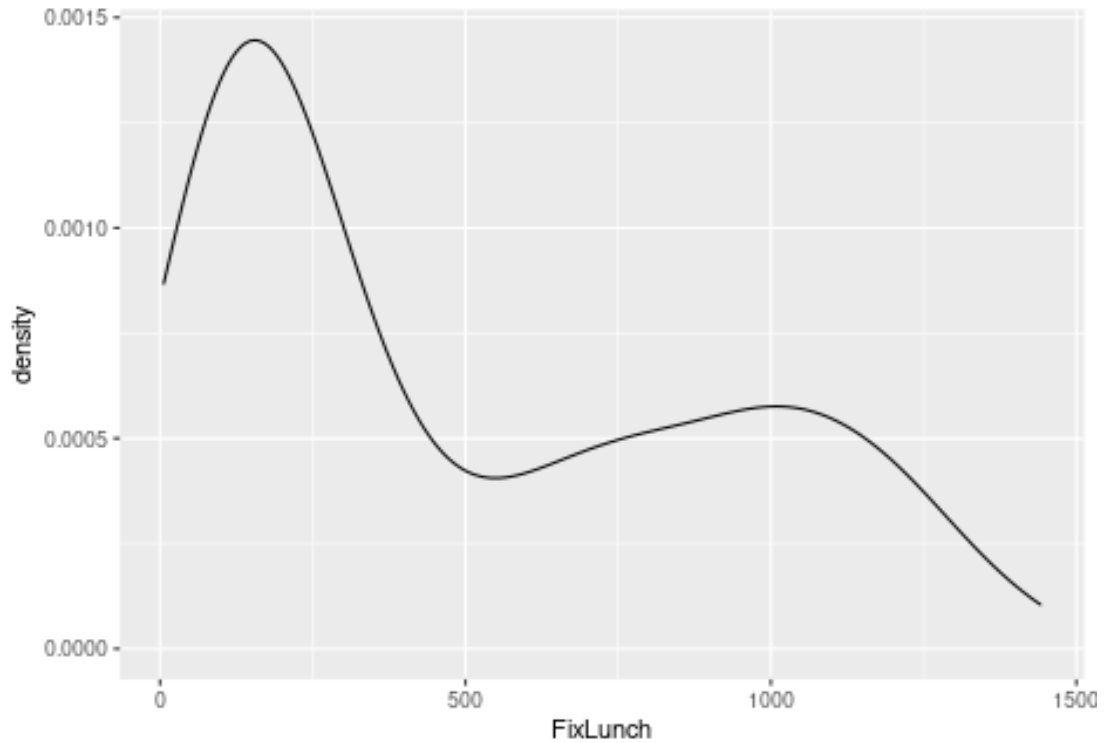


Figure 2.6: Fix lunch density

One can observe that there are two lumps. First lump has smaller mean and variance compared to second lump. Considering first lump's density, it is probable to take negative values, however, the sales cannot be lower than zero. Therefore, we used log-transformation to remove the non-negativity constraint

and help mitigate any remaining right skewness. Moreover, these lumps are likely to represent the sales differentiation during weekdays and weekends.

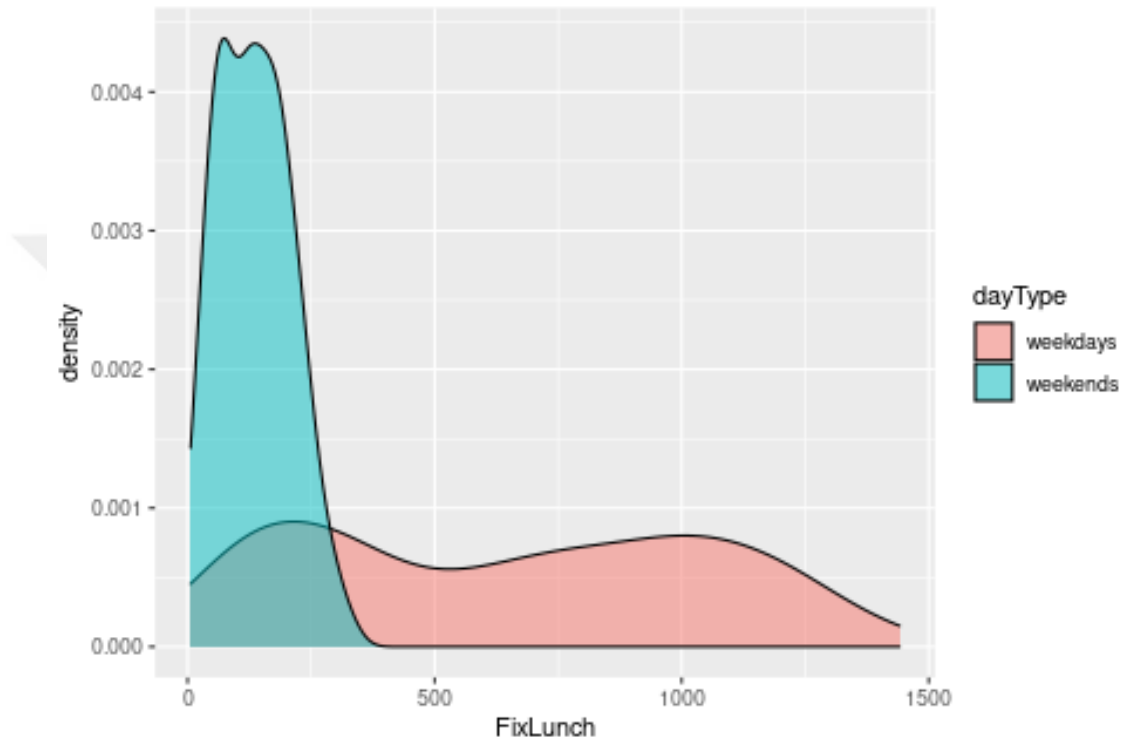


Figure 2.7: Fix lunch density for weekdays and weekends

Chapter 3

Literature Review

Catering industry has been researched in the literature for the topic of forecasting in many projects. The authors studies meal plan forecasting using time series analysis, causal analysis and complex analysis. Time series methods used in the catering industry include moving average, simple exponential smoothing, naïve, Holt's and Winter's methods.

Moving average models are summarized as calculating the average of the observations to use as the predictor for the next period which means that the analysis is greatly dependent on the number of terms chosen to construct the average [1]. Simple exponential smoothing is a technique that employs a weighted moving average of previous data as a forecasting basis. The algorithm considers heaviest weight to the recent observations and assigns smaller weight to the more distant past observations [2]. Naïve methods are based on the most recent information available and consider the dependent variable by using data from the previous week to forecast the current week [3]. Winters' method and prophet model provide a useful way to explain the seasonality when time-series data have a seasonal pattern [2] while Holt's Method [4] incorporates the trend in data in addition to smoothing the historical pattern.

Causal methods utilized for the catering industry for forecasting purposes

mainly use regression models. Linear regression aims to estimate the causality between a dependent variable and independent variables with linear predictors [1]. Multiple linear regression differs from simple linear regression with using several independent variables to predict a dependent variable. Complex methods use artificial intelligence methods such as neural networks and random forest. Random forest is a tree-based method involving stratification or segmentation of the space of inputs into a number of simple regions. With the aim of making a prediction for a given observation, the mean of the training observations in the region to which it belongs is used.

The results of examinations are compared generally with the metrics of mean absolute deviations (MAD), mean squared errors (MSE), mean absolute percentage errors (MAPE) and cumulative mean absolute percentage errors (cMAPE).

In the study conducted by Blecher and Yeh [5], the purpose is to determine which forecasting model predicts meal participation at university residential dining facilities most precisely. The authors used data collected from a large urban university to test naive, three different moving average, and simple exponential smoothing forecasting methods. The moving average models differ according to the participants considered for the analysis. Moving average method, A and B, respectively, use the average of the actual number of participants from the previous three and five weeks for the same day of the next week forecasting, while method C uses the average of the observed number of participants for a particular day from the previous five weeks. The data used for this study consist of the actual number of students who attended the service of food at two residential dining establishments for fifteen weeks. At these dining establishments, breakfast, lunch, and dinner meals are offered on weekdays while brunch and dinner are served on weekends. Meal participation at each dining hall ranged between approximately 300 and 700 students changing with the day of the week while utilizing a four-week select cycle menu. Also on the closed days, the data used for forecasting purposes is the previous same day of the week. The results imply that the moving average methods do better than exponential smoothing in the vast majority of the cases and provide 6.4% or less MAPE values. Method C surpassed other algorithms for 75% of the cases for forecasting predictions in

terms of MAPE. Although the difference between using methods A and C is fairly small for some cases, the data support the moving average algorithm C to be the preferred method for forecasting participation.

In another study to analyze the algorithm of forecasting meal participation which is conducted by Kisang and Sanchez [4], the forecasting methods of naive model, moving average, double moving average, simple exponential smoothing, double exponential smoothing, Holt's, and Winter's methods and linear and multiple regressions analysis are considered. The data are collected daily at a dining establishment of a university during 2000 fall semester and 2001 spring semester for 13 weeks. Even if the data contain the counts of every meal, the research examined solely counts for dinner meal participation for all the days except Sundays. The result of this study indicates that multiple regression is the most accurate forecasting method and the methods with regards to the effects of seasonality comparatively perform better. The authors use a performance metric, namely U-statistic which acknowledges relative comparison of the forecasting approaches with naive methods [6]. U-statistic can be defined as the RMSE from forecasting method to the RMSE of the naive method [7][8]. Then, U-statistics can be categorized into three groups as in Table 3.1

Table 3.1: U-statistic

$U < 1$	the forecasting technique being used is better than the naive approach
$U = 1$	the naive approach is as good as the forecasting technique being evaluated
$U > 1$	there is no reason to use a formal forecasting method, since using a naive method will generate better results

In this study, multiple regression with U-statistic of 0.78 score outperformed the rest of the methods since this method appropriately fitted the time series data which consists of the influence of seasonality. Secondly, Winter's exponential smoothing method provides U-statistic of 0.83 value with the optimal alpha,

beta, and gamma values. Thirdly, naïve model with a lag of one week regards seasonality since it utilizes the past week of data for the next week forecasting. This method achieves the value of U-statistic of 1.

A study is carried out for Swedish catering system by Malefors, Strid, Hansson and Eriksson [9], the authors examined conventional forecasting methods such as moving-average models and complex models such as prophet model, neural network model. The study also considered the associated margins for all models. Data used in the forecasting models are the number of portions registered at 21 schools in the period 2010–2019 provided by the municipalities regarding that a school year consists of at least 178 days. The students are offered school meals without cost and with two menu options. In the analysis, the past year data points are employed to test the models with current observations. Conventional forecasting methods are simple to implement as well as providing the lowest MAPE results of 2–3% in seven cases, however the neural network models are applied with the lowest MAPE value of 2–3% for eleven kitchens. With a two-day window analysis, the moving-average model and the prophet model are the performing better models for seven kitchens and for three kitchens respectively. However, there exists small differences between the models in three of the cases. To exemplify, the moving average model with a five-day window with MAPE value of %1.3 performs similar to its moving average equivalent with a two-day window with MAPE value of %1.4 and the neural network model with MAPE value of %1.5.

A similar study [10] which is carried out for the Swedish public catering system is also undertaken during the Covid-19 pandemic with the purpose of forecasting the guest participation throughout the pandemic by collecting data from school and preschool catering establishments before and during the pandemic, and employing machine learning methods based on historical data. The dataset for evaluating covered 18 primary school kitchens and 16 preschool kitchens starting on the beginning of the pandemic of March 18, 2020 and ending on June 9, 2020 which is regarded as the end of the spring semester. The complete data set includes the number of guests participation from 2010 to the beginning of June 2020 and is performed by counting plates after every lunch since plate data are

an estimation of the number of pupils benefited from meal serving. The results of various learning methods comparison imply that random forest algorithm with cMAPE is less than 0.15 for the trained dataset for one of the kitchens issued more accurate predictions than a neural network. The authors analyzed the dataset with both pooled and kitchen level. The pooled train data set has sales with daily mean 248 for one year with 125 and 119 missing days for first and second kitchen respectively while the test data covers eight days period with no missing days with daily mean 215. The random forest approach is employed and MSE is equal to 59322 and 3036 for training and test data, respectively. cMAPE score for the pooled dataset for all kitchens studied is equal to 0.39 for the training data and 1.28 for the test data. At kitchen level, MSE for first kitchen is around 175 and 609 for the training data and test data; and cMAPE is around 0.448 and 0.137 for the training and test data, respectively. Comparatively with a neural network with one hidden layer, the MSE of the test data is around 11600 and cMAPE is around 4.104. For the second kitchen, the MSE is around 19 and 107 for the training data and test data, while cMAPE is 0.95 and 0.487 for the training data and test data respectively.

Other than machine learning methods, Markov Chain Monte Carlo (MCMC) methods are widely used for prediction. The idea behind MCMC methods is to formulate is a Markov chain for its parameters and to use it to build simulation-based estimates in case it has long-run distribution [11]. The algorithm starts from a random point in parameter space and proceeds forward through the simulated chain. Since we use random variables, this proceeding step requires the definition of conditional distribution, $\theta_{r+1} \mid \theta_r$. After repetition of the conditional distribution forward, a large sample from joint distribution of all variables is built. One can notice that the conditional distribution depends only on the previous step which is defined as Markov property[11]. In a study which is carried out by Radchenko, MCMC based Bayesian inference is used to forecast the prices in energy industry. The author uses Gibbs sampling algorithm to predict models with changing trends. This study concludes that Bayesian framework combined with Gibbs sampling algorithm results in decrease in mean forecast squared error at 36% for 10-period-ahead forecasts [12].

In another study [13], the authors develop a model to forecast used product returns. They use MCMC-based Bayesian inference approach for parameter estimation. Firstly, they produce initial predictions for the parameters of the model using simulation techniques. Secondly, using Markov chain approach, they generate new samples of parameters. Thirdly, posterior probabilities are determined for generated values via Metropolis-Hasting algorithm which is one of MCMC methods and lastly, parameter estimation from the posterior distribution is performed. They conclude that compared to maximum likelihood estimation (MLE) approach, MCMC-based Bayesian inference approach offers fast method to draw samples from posterior distribution in high-dimension.

Apart from the applications of MCMC methods in literature, it is important to use the statistics to check convergence. Convergence of the Gibbs sampler is ensured under fairly mild status. Tierney's Theorem 1 guarantees convergence of the n -step-ahead distribution to the invariant distribution for practically all starting locations if the Gibbs sampler is irreducible. The majority of reducible Gibbs samplers have some type of state space constraint. A Gibbs sampler whose conditional densities are strictly positive everywhere and whose marginal densities exist will be irreducible if the state space is the Cartesian product of intervals on each coordinate axis. Furthermore, this sampler will converge to the stationary distribution from the beginning points and will be ergodic, according to Tierney's Theorem 2, Corollary 1 [11].

By definition, a Markov chain generates samples from the target distribution only after it has converged to equilibrium. In theory, convergence is only guaranteed asymptotically as the number of draws grows without bound. In practice, experimenter should use diagnostics to check convergence for the finite number of draws available in real life problems[14].

One method to monitor convergence is to compare the chains to the other chains with random initial points. This forms the motivation for Gelman-Rubin statistic, $\hat{R}$. The statistic measures the ratio of the average variance of samples within each chain to the variance of the pooled samples across chains; if all chains are at equilibrium, these will be the same and $\hat{R}$ will be one. If the chains have

not converged to a common distribution, the $\hat{R}$ statistic will be greater than one [15].

1. For each chain, first discard D values as "burn-in" and keep the remaining L values, $x_D^{(j)}, x_{D+1}^{(j)}, \dots, x_{D+L-1}^{(j)}$. For example, you might set D = L.

2. Calculate

$$\begin{aligned}
 \bar{x}_j &= \frac{1}{L} \sum_{t=1}^L x_t^{(j)} && (\text{chain mean}) \\
 \bar{x} &= \frac{1}{J} \sum_{j=1}^J \bar{x}_j && (\text{grand mean}) \\
 B &= \frac{L}{J-1} \sum_{j=1}^J (\bar{x}_j - \bar{x})^2 && (\text{between chain variance}) \\
 s_j^2 &= \frac{1}{L-1} \sum_{t=1}^L (x_t^{(j)} - \bar{x}_j)^2 && (\text{within chain variance}) \\
 W &= \frac{1}{J} \sum_{j=1}^J s_j^2
 \end{aligned} \tag{3.1}$$

3. The Gelman-Rubin statistic is then

$$R = \frac{\frac{L-1}{L}W + \frac{1}{L}B}{W} \tag{3.2}$$

Gelman and Rubin's recommendation is that the independent Markov chains should be initialized with diffuse starting values for the parameters and sampled until all values for $\hat{R}$ are below 1.1. $\hat{R}$ statistic can be summarized as in Table 3.2

Table 3.2: $\hat{R}$ -statistic

$\hat{R} = 1$	all chains are at equilibrium
$\hat{R} > 1$	the chains have not converged to a common distribution

We have built linear regression model and two hierarchical models to solve our

problem. These models are hierarchical Bayesian linear regression and hierarchical Bayesian with multilevel linear regression model.



Chapter 4

Linear Regression Model

In this chapter, we will build a multiple regression model as benchmark. At first we start with a model of FixLunch regressed on main effects. Diagnostic plot suggested that log transformation of FixLunch may be regressed on the predictors and their interactions. To avoid overfitting, we select predictors with backward elimination. These steps are described in Section 4.1 while for the final model the regression assumptions are checked in Section 4.2. Variable selection and diagnostics for linear regression model will be examined.

4.1 Variable Selection

Since the data set is formed by menu ingredients and academic calendar events, we build the model using these variables and their interaction terms. Then, we use `step` function [16]. This function is used to select variables in the final model with backward elimination. This elimination method starts with the full model and iteratively disregards the variables which provides least contribution to the model by checking Akaike Information Criteria (AIC). AIC is a mathematical method to decide the best fitted model for the data. One can use the formula to calculate AIC

$$AIC = 2\mathbb{K} - 2\log(L)$$

where $\mathbb{K}$ and L represent the number of variables in the model and the maximum likelihood estimate of the model. We selected the model with the following formula to put into **step** function:

$$\begin{aligned} \text{FixLunch} = & \alpha + \text{SEM}_i + \text{ACADYR}_i + \\ & \text{RMDN}_i + \text{WDAY}_i + \text{HOLIDAY}_i + \\ & \text{Temp} + \text{FL_bakla}_i + \text{FL_dolma}_i + \\ & \text{FixLunchhas_fish}_i + \text{FL_mveg}_i + \text{FL_veg}_i + \\ & \text{FL_rmeat}_i + \text{FL_wmeat}_i + (\text{SEM}_i \times \text{Temp}) + \\ & \epsilon \end{aligned} \tag{4.1}$$

After comparing the AIC values, the model formula for the selected model becomes:

$$\begin{aligned} \text{FixLunch} = & \alpha + \text{SEM}_i + \text{ACADYR}_i + \\ & \text{RMDN}_i + \text{WDAY}_i + \text{HOLIDAY}_i + \\ & \text{Temp} + \text{FixLunchhas_fish}_i + \text{FL_mveg}_i + \\ & \text{FL_veg}_i + \text{FL_rmeat}_i + \text{FL_wmeat}_i + \\ & (\text{SEM}_i \times \text{Temp}) + \epsilon \end{aligned} \tag{4.2}$$

Figure 4.1 shows the diagnostic plots for this regression model. One can see that there is nonlinearity problem in residuals vs fitted model. To overcome this problem, we consider nonlinear transformation for the response variable. One can deduce that the residuals are fairly normal based on Q-Q plot. Scale-Location plot shows that the residuals have non-constant variance. Transformation may help eliminating this problem as well. Residuals vs. leverage plot shows that there are cases with high residual and low leverage. High residual is an implication of outliers in the data. Therefore, log-transformation is applied to the response variable and the formula for the new regression model becomes:

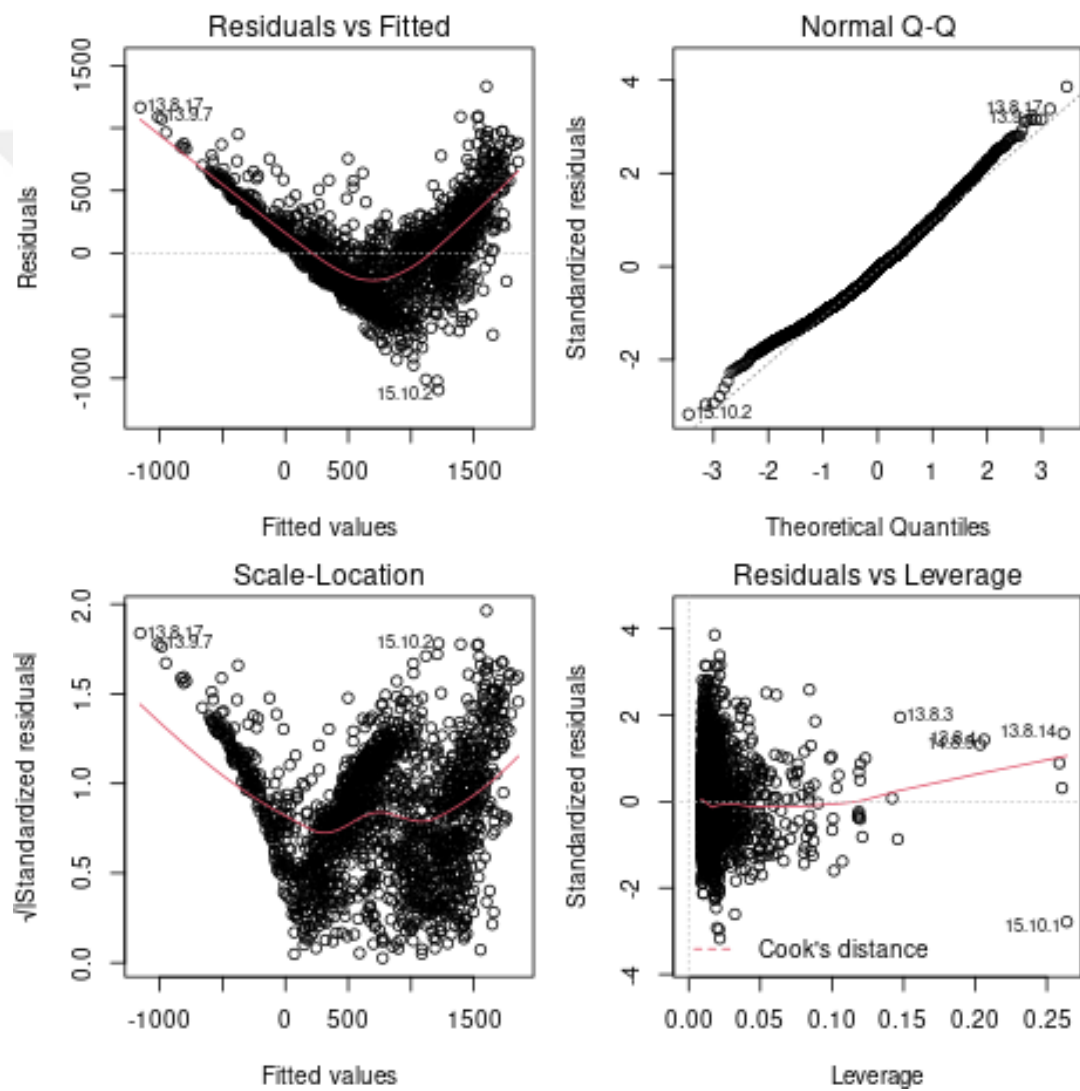


Figure 4.1: Diagnostic plots for the model with formula

$$\begin{aligned}
\log(\text{FixLunch}) = & \alpha + \text{SEM}_i + \text{ACADYR}_i + \\
& \text{RMDN}_i + \text{WDAY}_i + \text{HOLIDAY}_i + \\
& \text{Temp} + \text{FixLunchhas_fish}_i + \text{FL_mveg}_i + \\
& \text{FL_veg}_i + \text{FL_rmeat}_i + \text{FL_wmeat}_i + \\
& (\text{SEM}_i \times \text{Temp}) + \epsilon
\end{aligned} \tag{4.3}$$

Figure 4.2 shows the diagnostics plots for the regression model which log-transformation is applied to response variable.

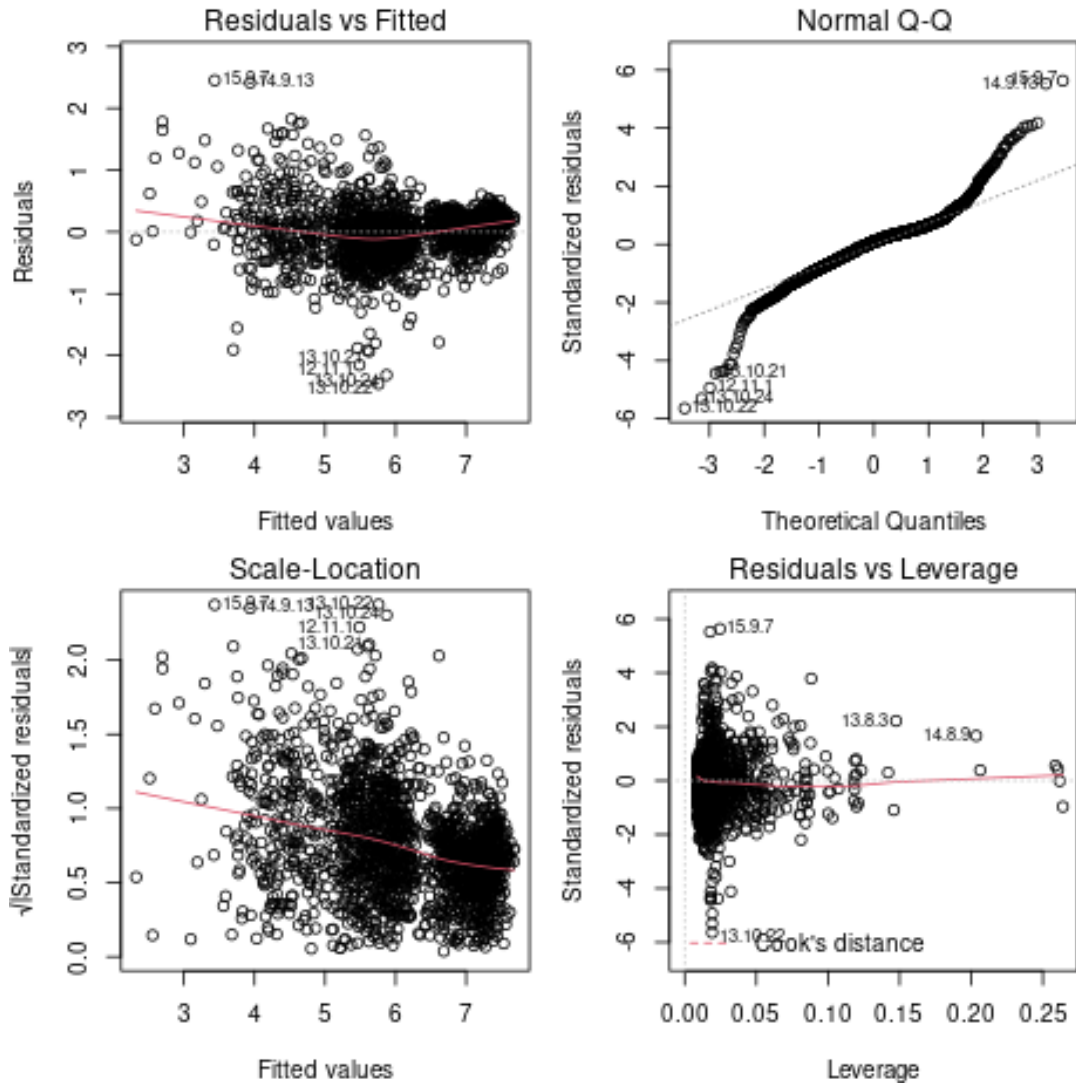


Figure 4.2: Diagnostics after log-transformation is applied

Log-transformation alleviates non-linearity problem in residuals vs fitted plot. However, there is still a slight pattern in the plot. Normal Q-Q plot has fat tails and residuals appear highly non-normal.

In the final model, the response variable is log-sales and the formula for the model as follows:

$$\begin{aligned}
 \log\text{FixLunch} = & \alpha + \text{SEM}_i + \\
 & \text{ACADYR}_i + \text{RMDN}_i + \\
 & \text{WDAY}_i + \text{HOLIDAY}_i + \\
 & \text{FL_rmeat}_i + \text{FL_wmeat}_i + \\
 & \text{PrevFixLunch} + \text{WeekagoFixLunch} + \\
 & \text{Prev2FixLunch} + \text{FL_veg2}_i + \\
 & \text{FL_kof}_i + \text{SEMPART}_i + \\
 & \text{SEMWEEK}_i + \text{COPEcope1}_i + \\
 & \text{GE100}_i + \text{REGgrad}_i + \\
 & \text{REGint}_i + \text{DAYSTOSTART}_i + \\
 & \text{SB_tatl}_i + \text{FL_bakla}_i + \\
 & \text{HLDYTOGO}_j + (\text{WDAY}_i \times \text{PrevFixLunch}) + \\
 & (\text{SEM}_i \times \text{WeekagoFixLunch}) + (\text{WDAY}_i \times \text{Prev2FixLunch}) + \\
 & (\text{WDAY}_i \times \text{HOLIDAY}_i) + (\text{SEM}_i \times \text{Prev2FixLunch}) + \\
 & (\text{SEM}_i \times \text{HOLIDAY}_i) + (\text{ACADYR}_i \times \text{HOLIDAY}_i) + \\
 & (\text{SB_tatl}_i \times \text{FL_bakla}_i) + (\text{SEM}_i \times \text{WDAY}_i \times \text{Prev2FixLunch}) + \\
 & \epsilon
 \end{aligned}
 \tag{4.4}$$

WEEK and PART were aliased by the SEMWEEK and SEMPART.

When performing a linear regression with a single independent variable, a scatter plot of the response variable versus the independent variable provides a good indication of the nature of the relationship. However, when the model contains more than one independent variables, a scatter plot is not adequate to present the relationship. Therefore, partial residual plots are useful. Partial

residual plots attempt to show the relationship between a given independent variable and the response variable given that other independent variables are also in the model.

Partial residual plots are meaningful with quantitative variables. Therefore, we use plots for *PrevFixLunch*, *Prev2FixLunch*, *WeekagoFixLunch*, *Temp*, *logPrevFixLunch*, *logPrev2FixLunch*, *logWeekagoFixLunch*. Figure 4.3 and Figure 4.4 show partial residual plots for these variables respectively. First plot shows that smoother does not follow fitted line, which is a probable indication of the current model needs higher order of *PrevFixLunch*. Similarly, third plot has same problem indicating the requirement for higher order of *WeekagoFixLunch*. Therefore, we decided to update the model by adding $PrevFixLunch^2$ and $WeekagoFixLunch^2$.

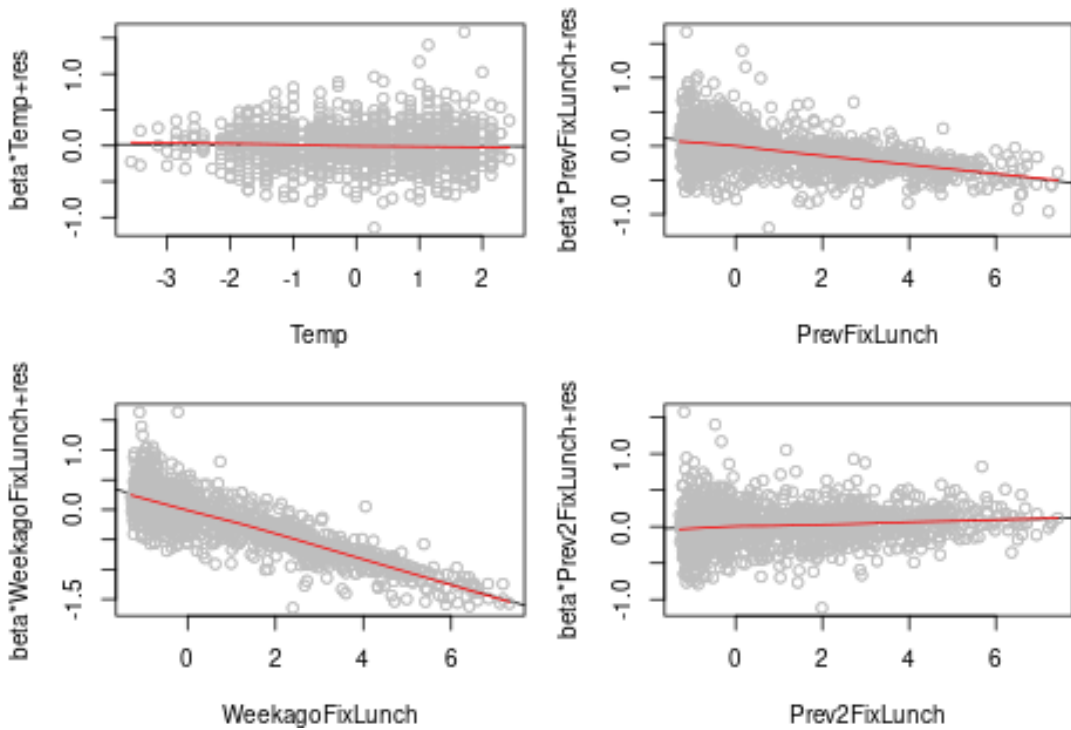


Figure 4.3: Partial residual plots for quantitative variables with current linear model

We used AIC and Bayesian Information Criteria (BIC) to compare the models and decide whether the new variables are required in the model or not. BIC is a

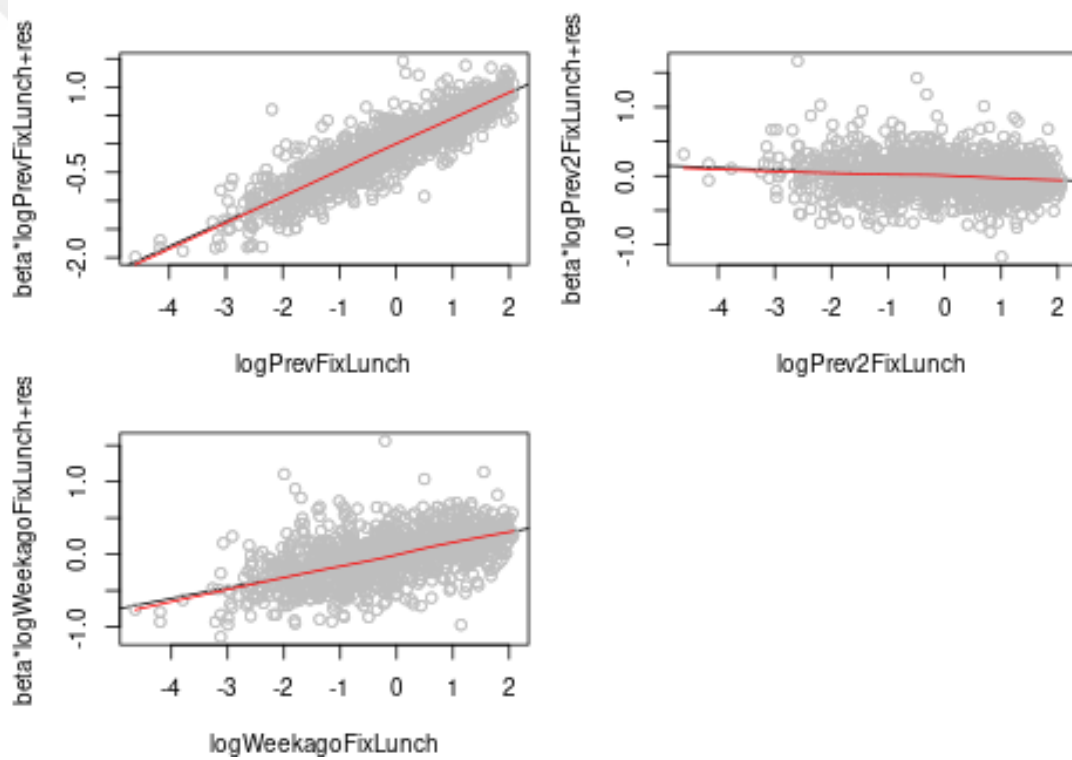


Figure 4.4: Partial residual plots for quantitative variables after log transformation with current linear model

similar method to AIC, with the formula

$$BIC = \log(n)\mathbb{K} - 2\log(L)$$

where n is the number of data points.

Table 4.1: AIC and BIC values for different models

Model Description	AIC	BIC
Current model	875.4180	1784.289
year*opened+ $I(year^2)*opened$ +	456.8214	1370.977
logPrevFixLunch*(WDAY+SEM)+		
logPrev2FixLunch*(WDAY+SEM)		
+logWeek-		
agoFixLunch*WDAY+logWeekagoFixLunch*SEM		
included		
$I(logPrevFixLunch^2)$ is added	456.9612	1376.401
$I(logPrevFixLunch^2)*WDAY$ is added	450.8662	1402.010
$I(logPrevFixLunch^2)$ is	456.9612	1376.401
included and logPre-		
vFixLunch,logPrevFixLunch:WDAY		
and logPrevFixLunch:SEM are disre-		
garded		
logPrevFixLunch:WDAY, logPre-	766.3567	1601.250
vFixLunch and logPrevFixLunch:SEM		
are disregarded		
logPrevFixLunch:(WDAY+SEM)	544.214	1458.369
and logPrevFixLunch disregarded		
and $I(logPrevFixLunch^2)$ and		
$I(logPrevFixLunch^2):(WDAY+SEM)$		
included		
$I(WeekagoFixLunch^2)$ is added	457.8152	1382.539

$I(\logPrevFixLunch^2),$	450.8662	1402.01
$I(\logPrevFixLunch^2):WDAY$	and	
$\logPrevFixLunch:WDAY$ included		

Based on AIC and BIC values, one can conclude current model and updated model with $PrevFixLunch^2$ has distinctly better score in terms of AIC. The probable reason the gap between scores of AIC and BIC is that BIC punishes more severely than AIC when the number of variables increase in the model. Therefore, we decided to continue with updated model. Figure 4.5 shows the partial residual plots for updated model, and one can deduce that smoother and fitted line are aligned for the first plot.

Figure 4.6 is the bar chart of t-statistics of significant effects whose t-statistic is greater than two in absolute value. The t-statistic is the effect divided by its standard error. The standard error is an estimate of the standard deviation of the effect. It can be thought as a measure of the precision with which the regression coefficient is measured. If a coefficient is large compared to its standard error, then t-statistic is probably different from 0. The greater the magnitude of t is, the stronger the evidence against the null hypothesis. This means there is greater evidence for a significant difference. Then the variables such as *year*, *WDAY6*, *SEMWEEK10*, *SEM2:logPrevFixLunch*, *logPrevFixLunch*, *DAYSTOSTART2*, *FixLunchhas_kof1*, *COPEcope11* have higher absolute value of t-statistic. The effect of *FixLunchhas_wmeat1* meaning that lunch menu does not offer white meat such as chicken appears to have highest t-statistics. Since t-statistics is negative, not having white meat in menu has a negative impact on sales. Similarly, *FixLunchhas_kof1* refers to the lunches with no meatballs offered in the menu and it has negative t-statistics meaning lunches with no meatballs has a negative impact on sales. Moreover, *year*, *DAYSTOSTART2* have positive t-statistics meaning that they are related with the increase in sales. As the year passes, the new cafeteria opened in the campus and the number of students raises which result in increase in sales, therefore, *year* has positive t-statistics. Similarly, *DAYSTOSTART2* represents how many days are left to the start of the academic year and considering that the students start to arrive to the campus, the sales

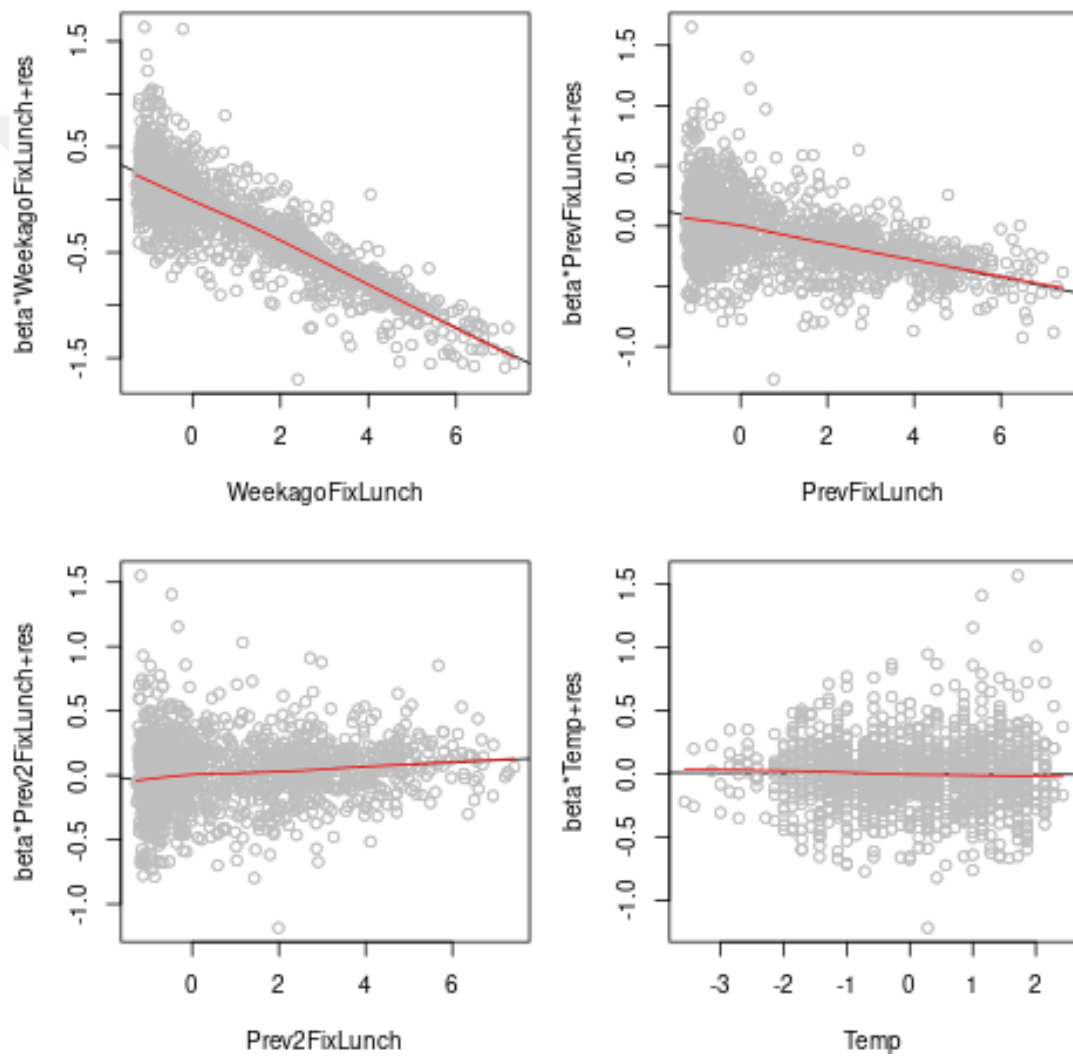


Figure 4.5: Partial residual plots after PrevFixLunch^2 is included to current linear regression model

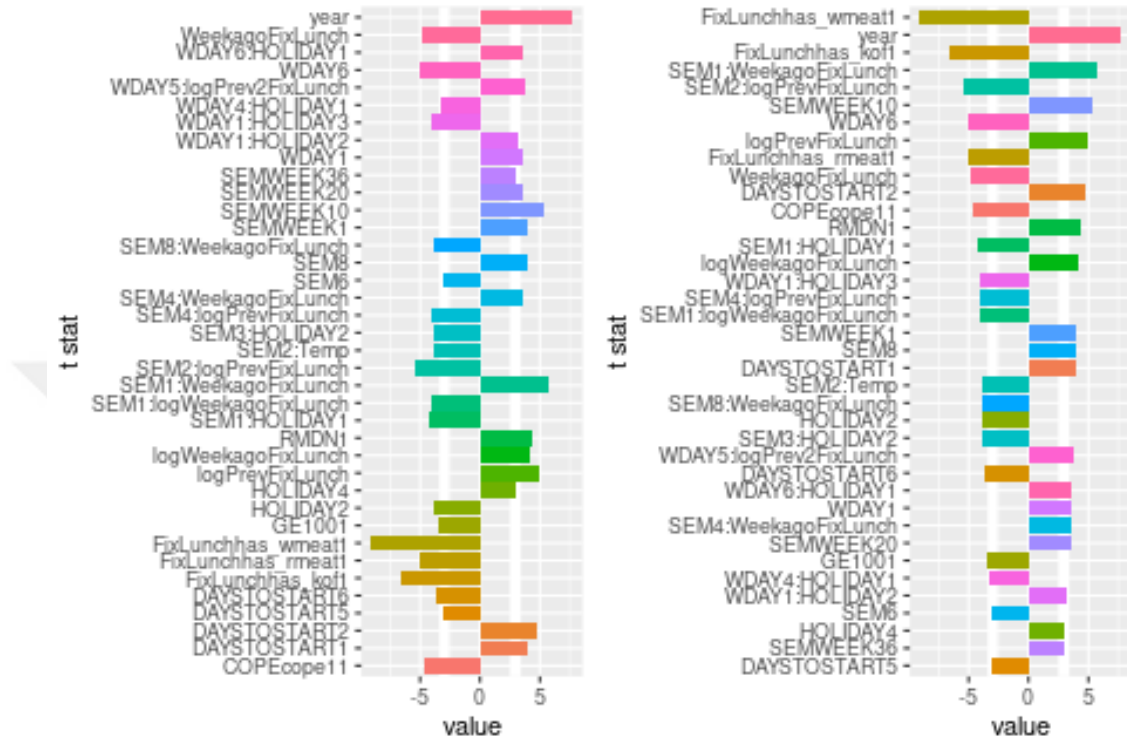


Figure 4.6: t-statistics. The white thick vertical lines mark -2 and +2 values. The bars show the t-statistics for only those predictors whose t-statistics are greater than two in absolute value. The second plot is with sorted values

increase. Therefore, it is coherent to have positive t-statistics. As interpreting the effect signs, consult Table 2.1 and 2.2 for factor levels. The effect of the last level is always minus the sum of the effects of remaining levels. Therefore, we can conclude these predictors have greater evidence to be significantly different than zero. Therefore, we can conclude that they are likely to have impact on the response variable.

4.2 Diagnostics

For the statistical tests and confidence intervals to remain valid, we must confirm that the final model satisfies the regression model assumptions. Here, we check if the model suffers any shortcomings and we find remedies if it does. Figure

4.7 shows the important diagnostic plots in assessing whether the linear model assumptions are satisfied. Residuals vs. fitted values plot shows the residuals vs fitted values on log-scale with scaled train data set. In linear model regression model, the error terms $(\varepsilon_1 \dots \varepsilon_n)$ are assumed independent and identically distributed with constant variance. Therefore, we expect the error term (ε_i) will have constant variance when plotted against fitted values and the errors and fitted values will be uncorrelated.

Lower left plot in Figure 4.7 the error variance does not remain constant. Normal Q-Q plot shows that the residuals appear highly non-normal. Both the lower and upper tail are heavier than we would expect under normality. This may be due to the non-constant variance issue we observed in the plot of residuals vs. fitted values. In scale-location plot there should be no discernible trends. Residuals vs leverage plot is useful to locate influential samples. Leverage is a measure of dissimilarity of predictor values of a sample from the remainder. It's a one-number summary of how different the model fit would be if the given observation was excluded, compared to the model fit where the observation is included. Cases with high standardized residual (poorly described by the model) and high leverage (high potential for influence on model fit) are influential. They are skewing the model fit away from the rest of the data and don't really seem to fit with the rest of the data. Cook's distances higher than 1 indicate influential cases.

Since the residuals' variance is not constant, residuals vs. fitted plot can be deceptive. Therefore studentized vs. fitted plot will be more informative. We use `residualPlots()` from `car` package in [17]. This function provides plots against all first-order terms, both regressors and factors and skips higher order terms. Figures 4.16, 4.17, 4.18 show that quantitative variables such as *Temp*, *PrevFixLunch*, *WeekagoFixLunch* have non-linear effects.

Table 4.2 shows the results for quadratic effect. For fitted values in linear models, the test is Tukey's one-degree-of-freedom test for nonadditivity [17] and it tests whether adding quadratic term to the model formula contributes to the model or not. Table 4.2 shows that we fail to reject the hypothesis which is

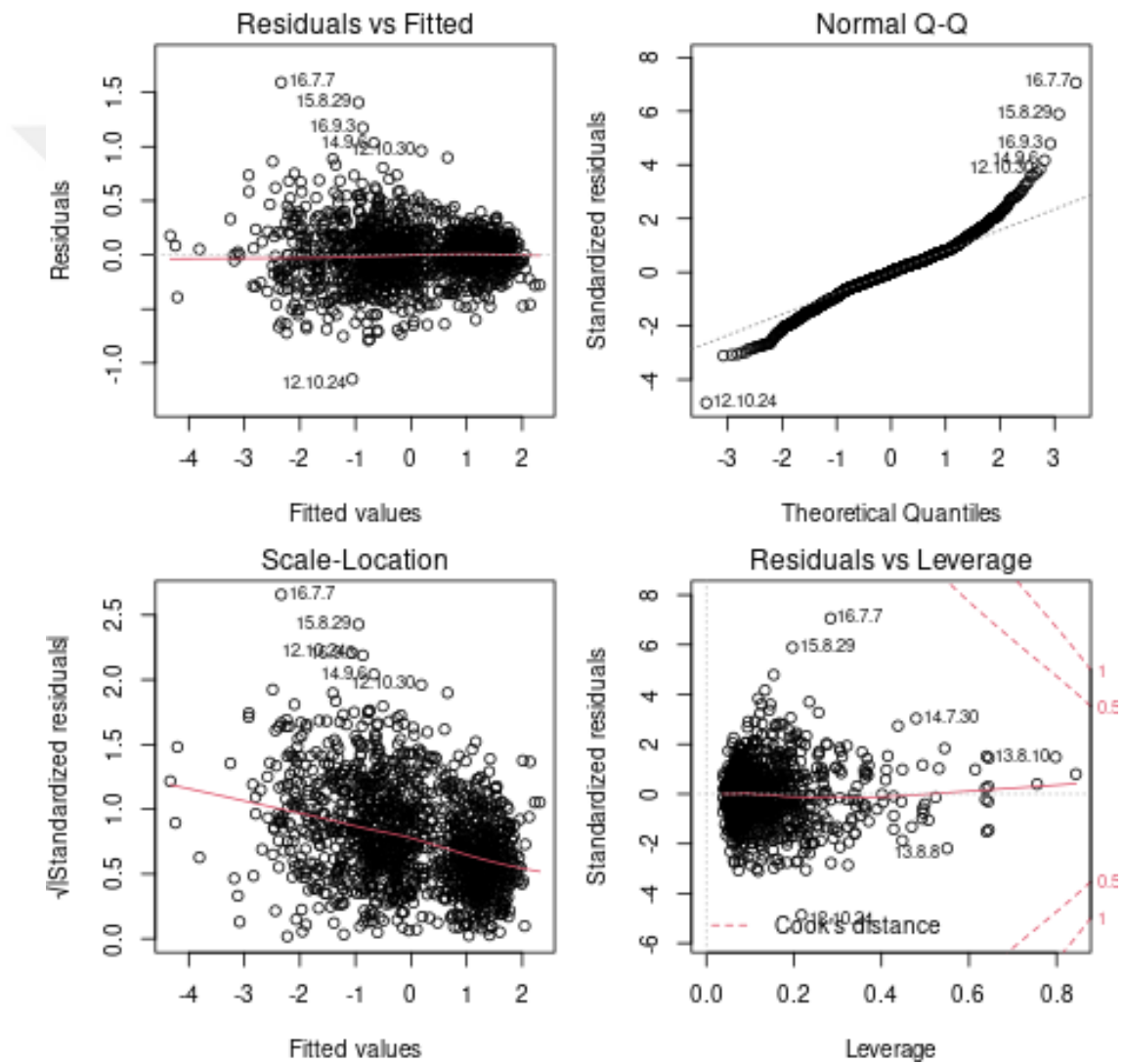


Figure 4.7: Diagnostic plots for linear regression model

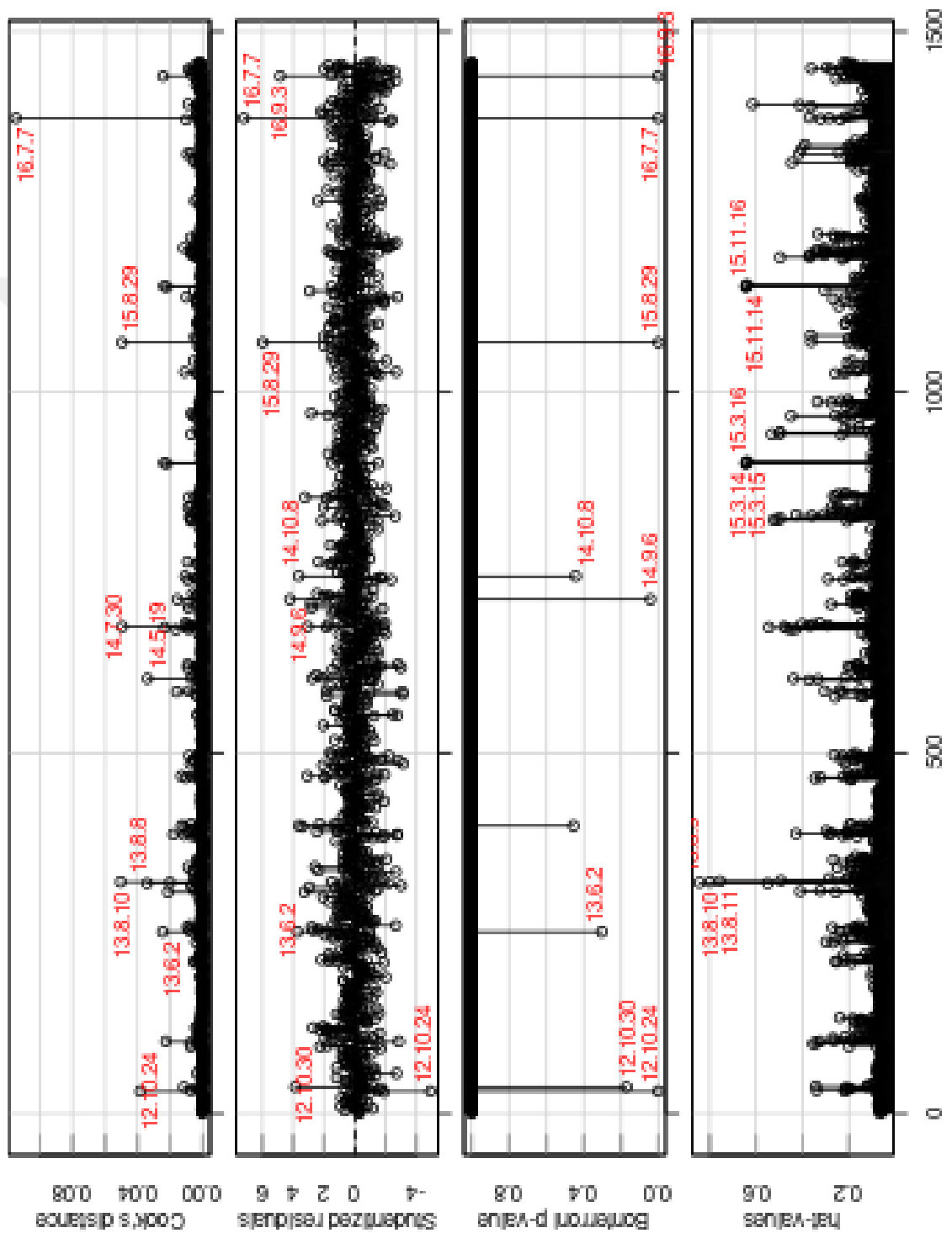


Figure 4.8: Diagnostic plots show influence measures calculated after each observation (x-axis) is removed

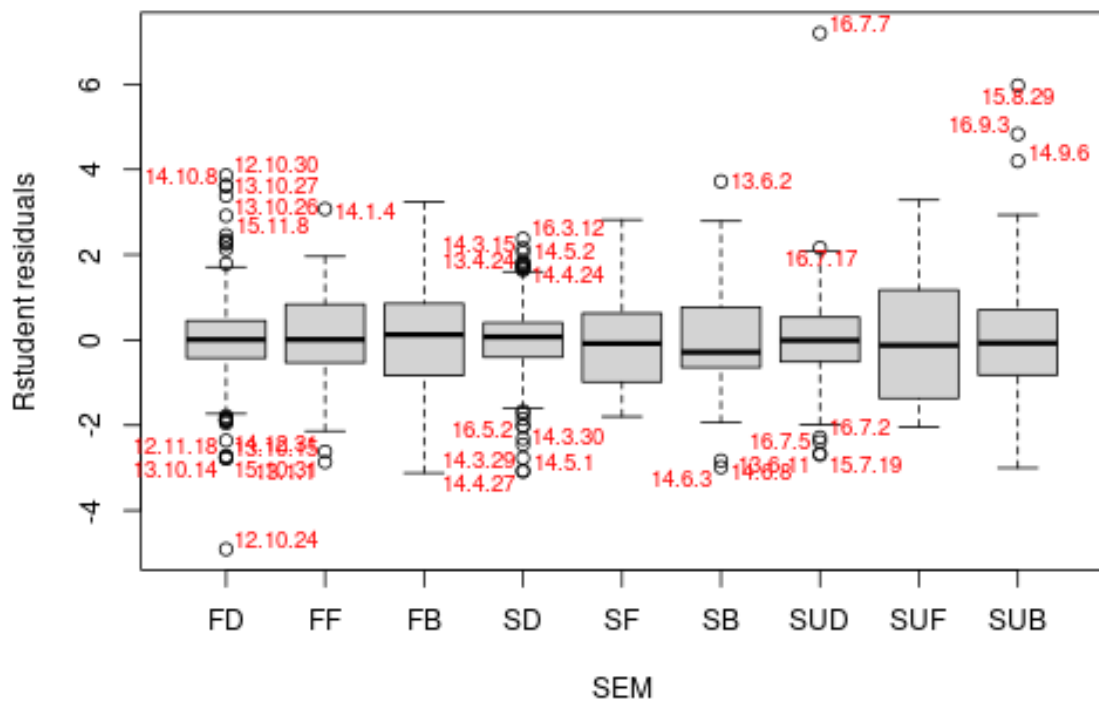


Figure 4.9: Studentized residual distributions conditional on semester. Dates for extreme points are expressed as yy.m.d here and later.

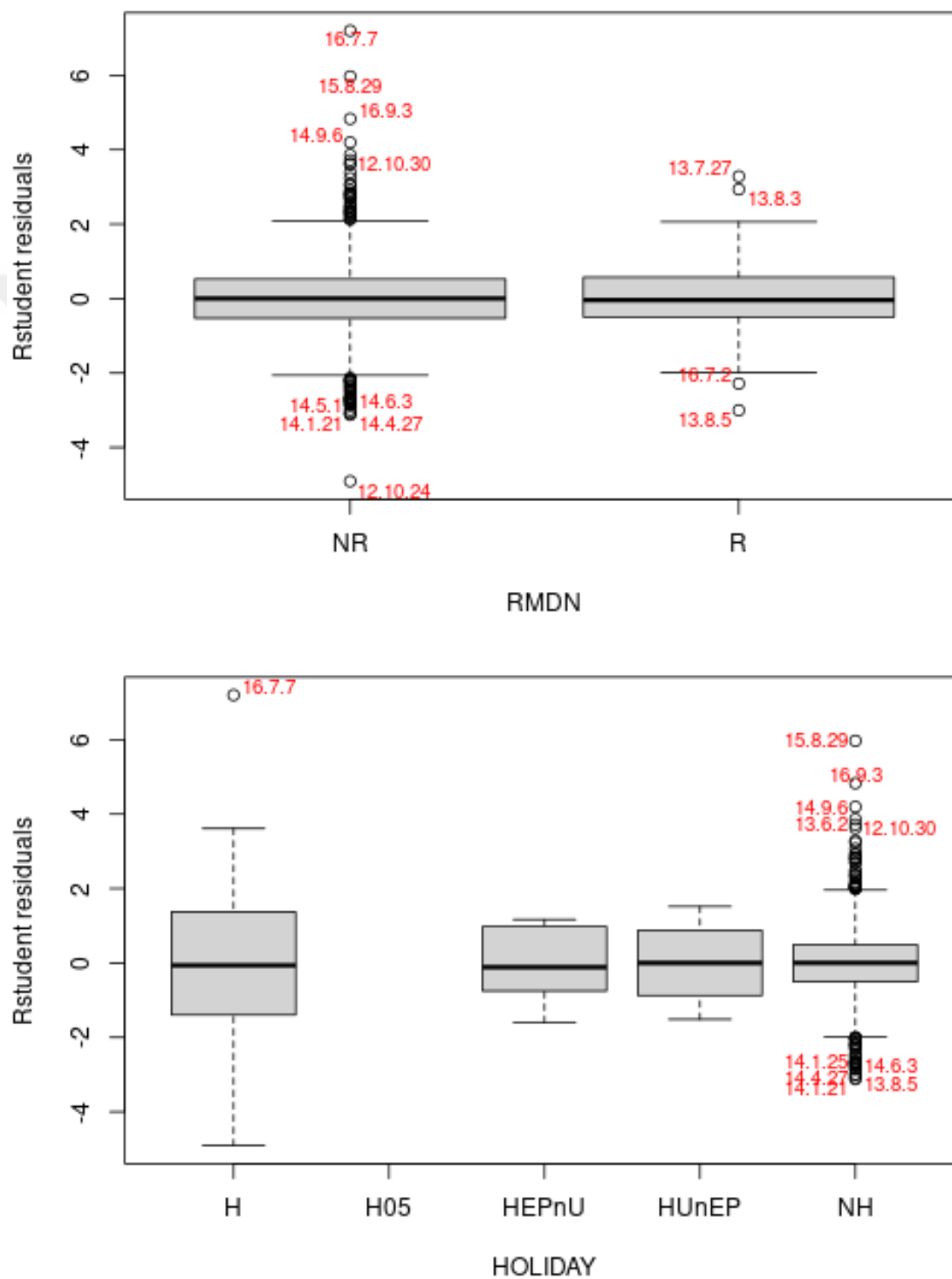


Figure 4.10: Studentized residual distributions conditional on Ramadan or holiday

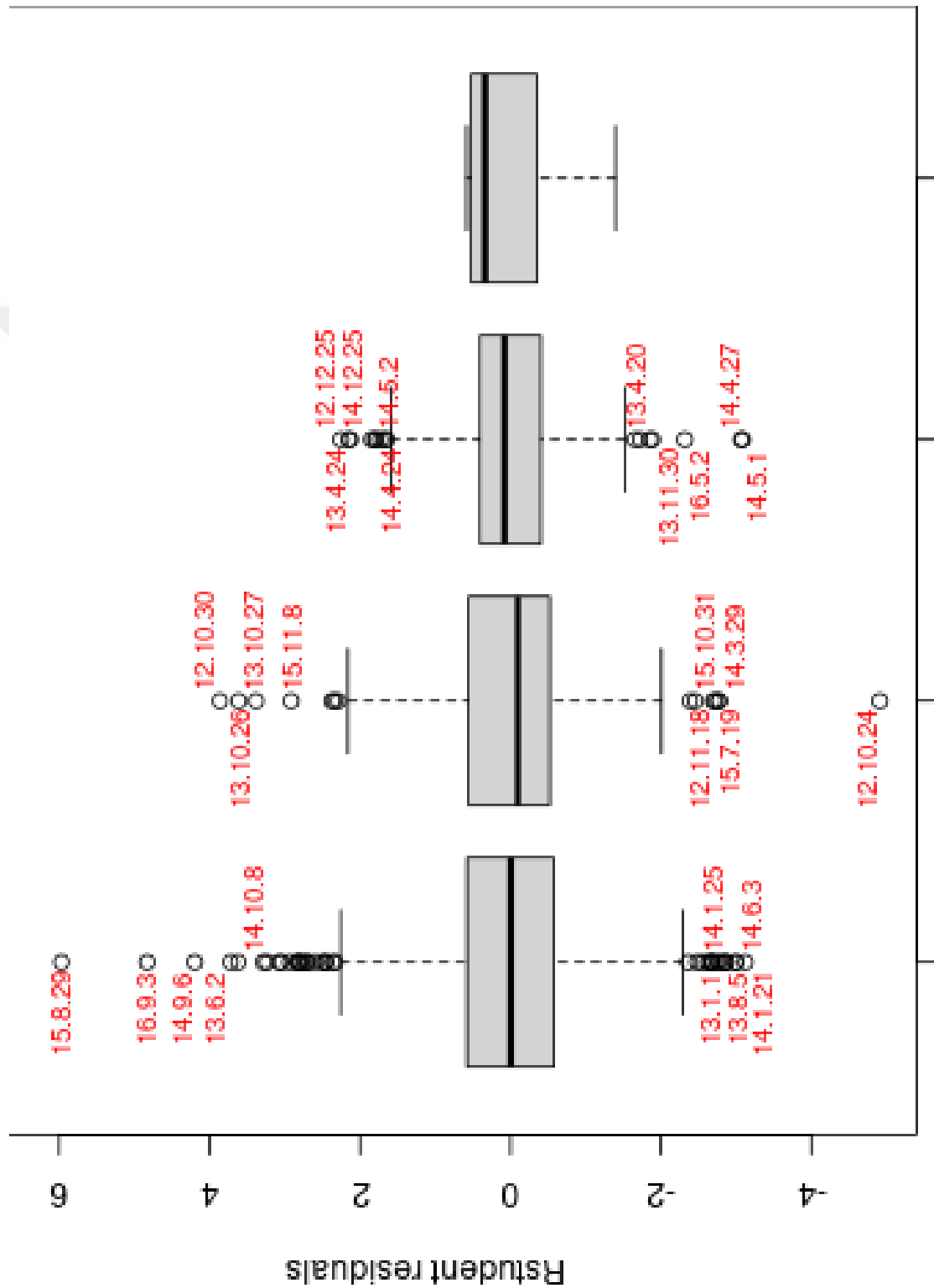


Figure 4.13: Studentized residual distributions conditional on week part

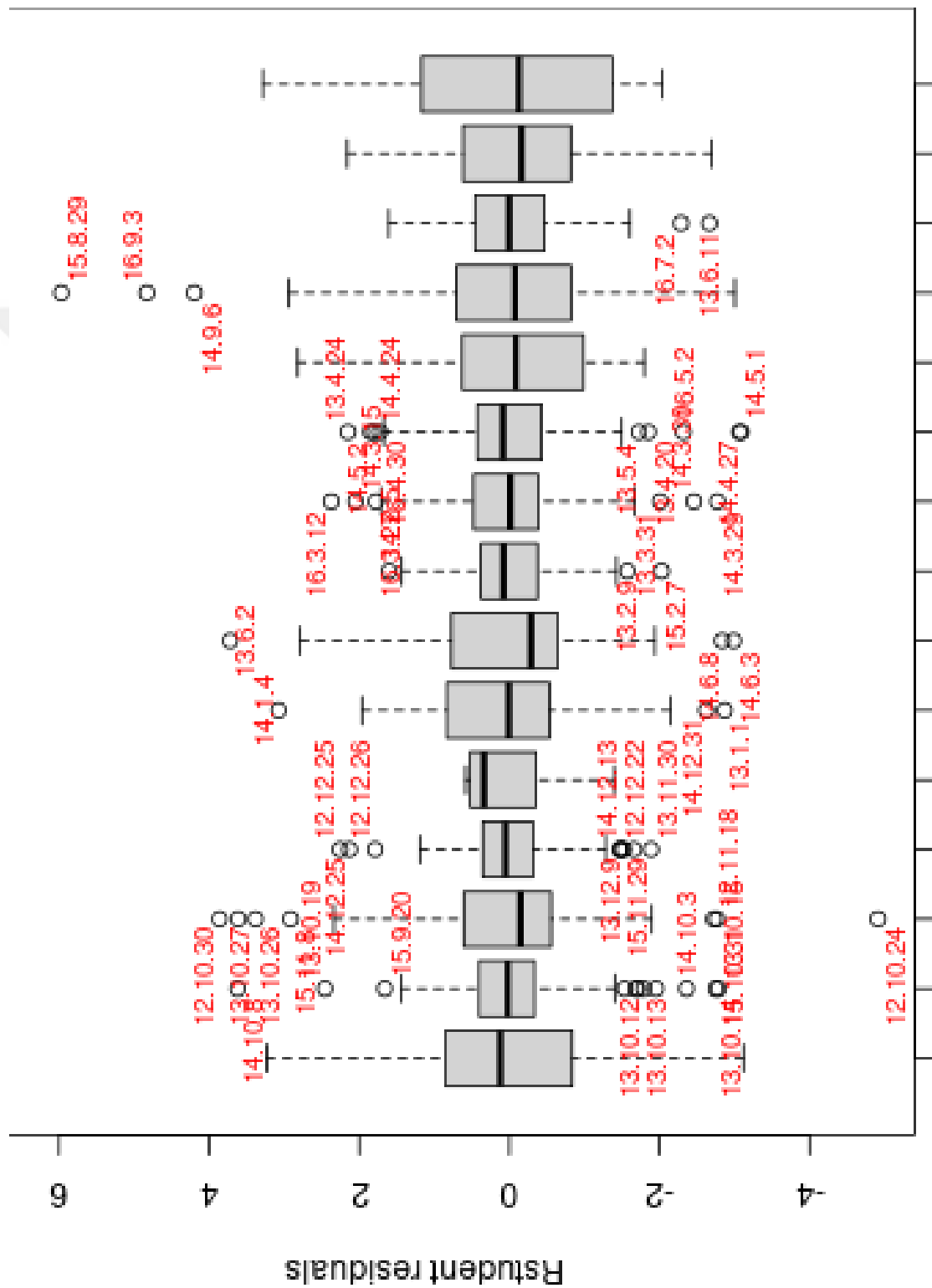


Figure 4.14: Studentized residual distributions conditional on semester part

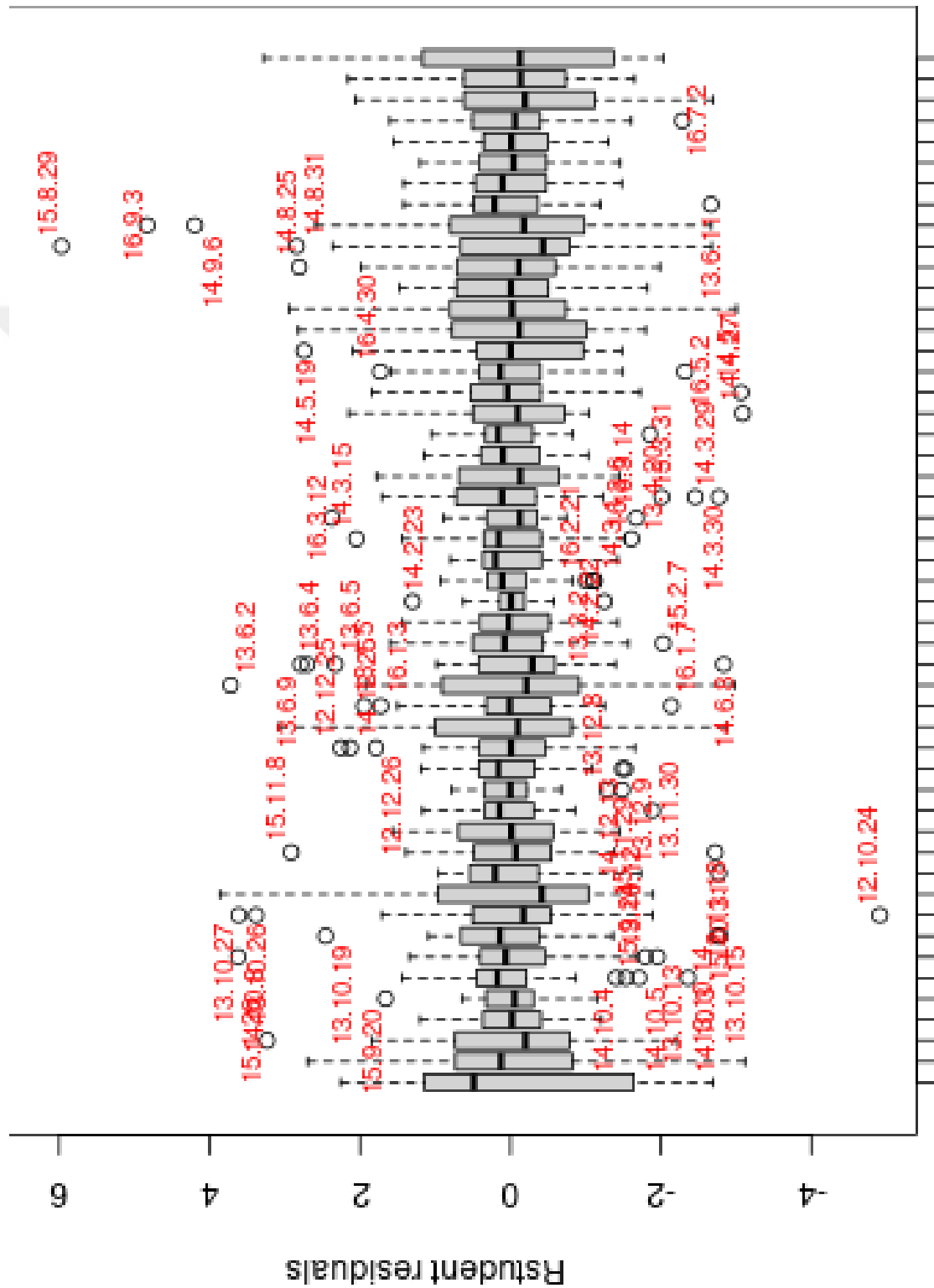


Figure 4.15: Studentized residual distributions conditional on semester week

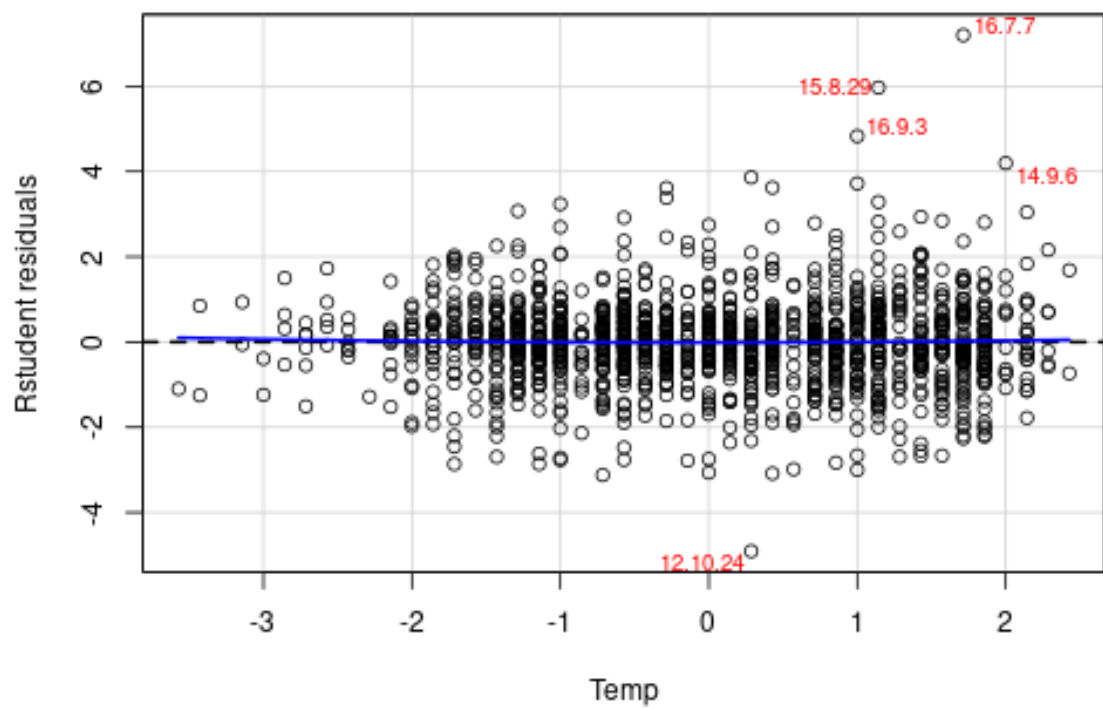


Figure 4.16: Studentized residual distributions conditional on temperature

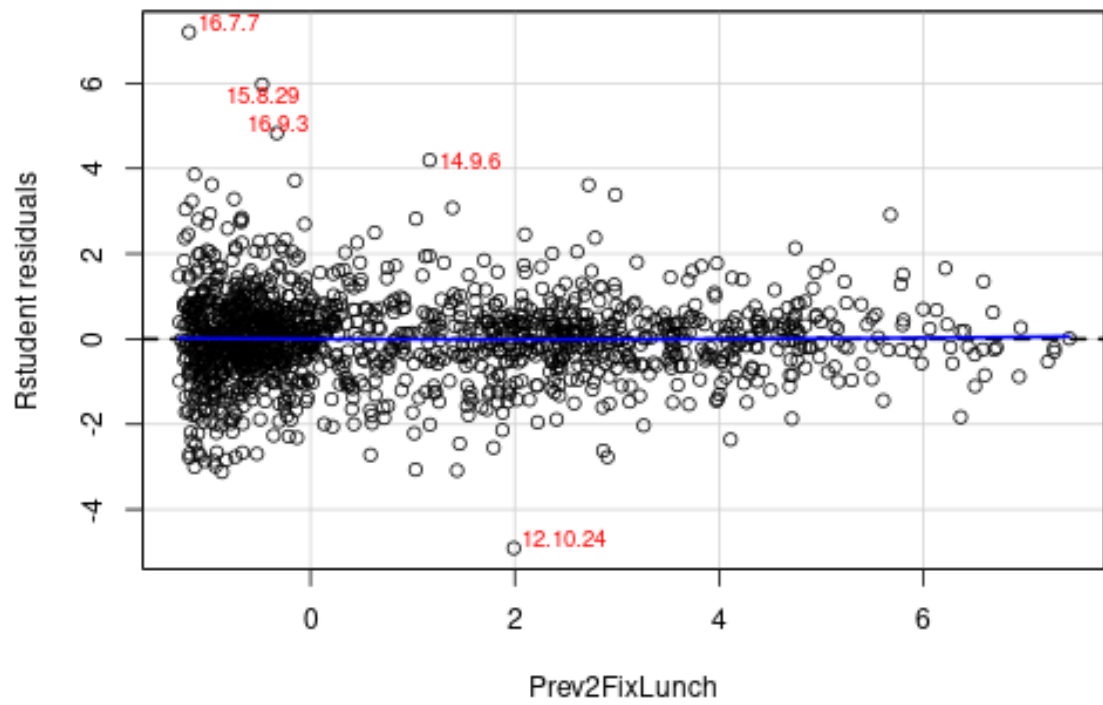
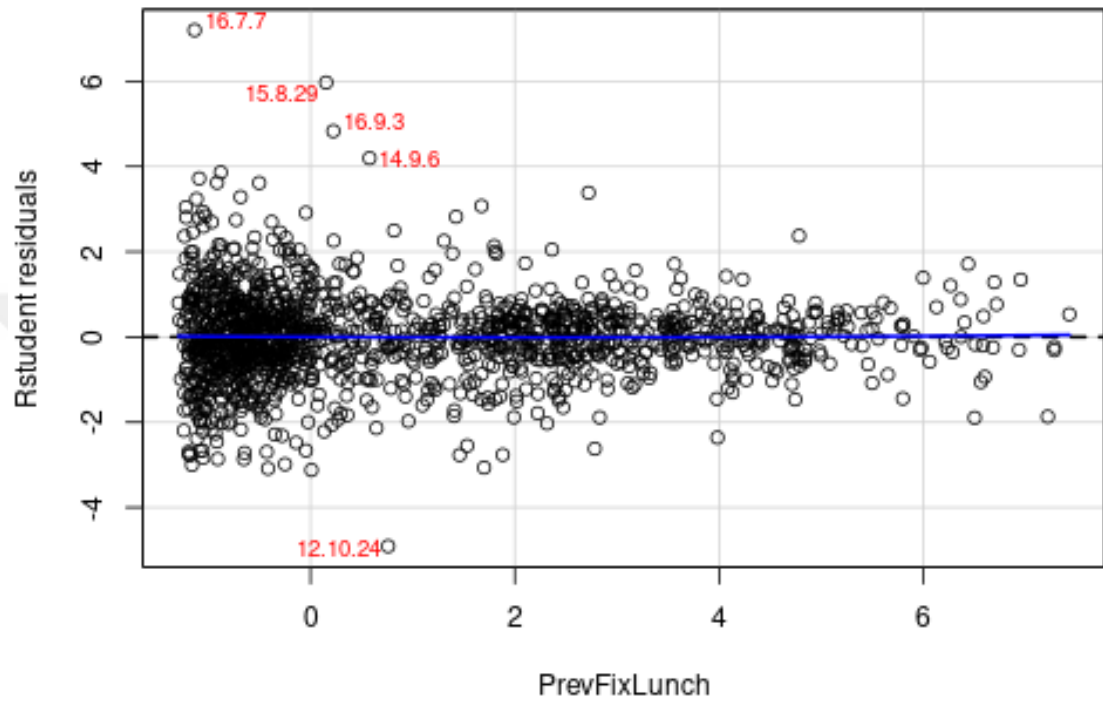


Figure 4.17: Studentized residuals vs fix lunch servings on previous two days

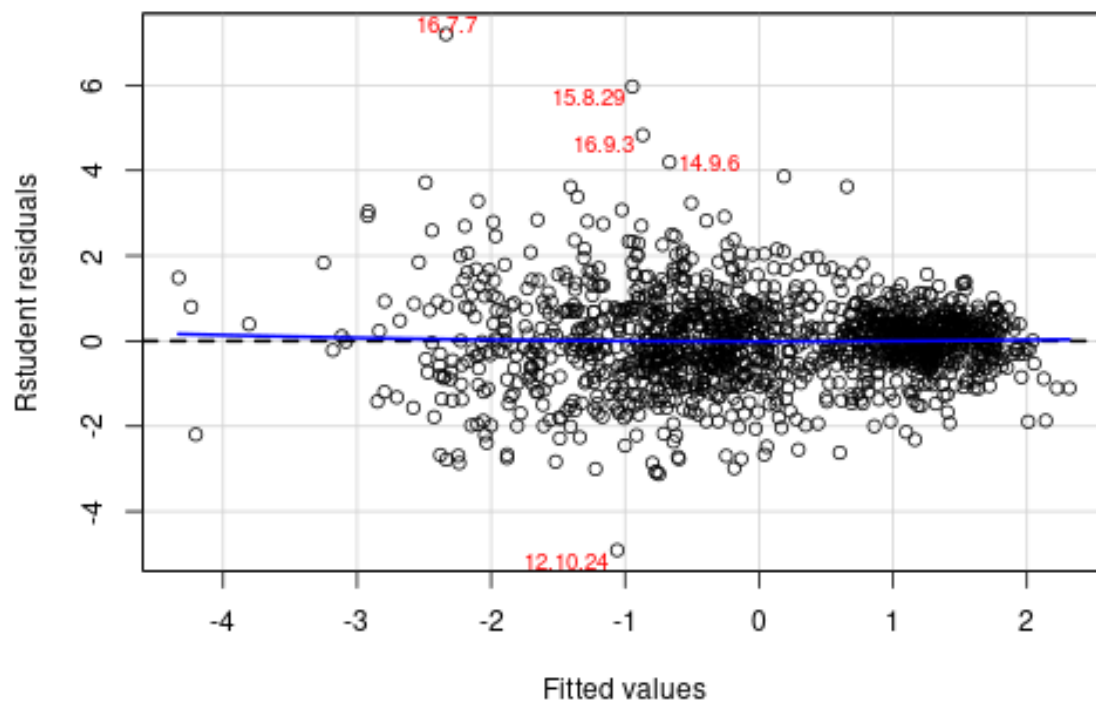
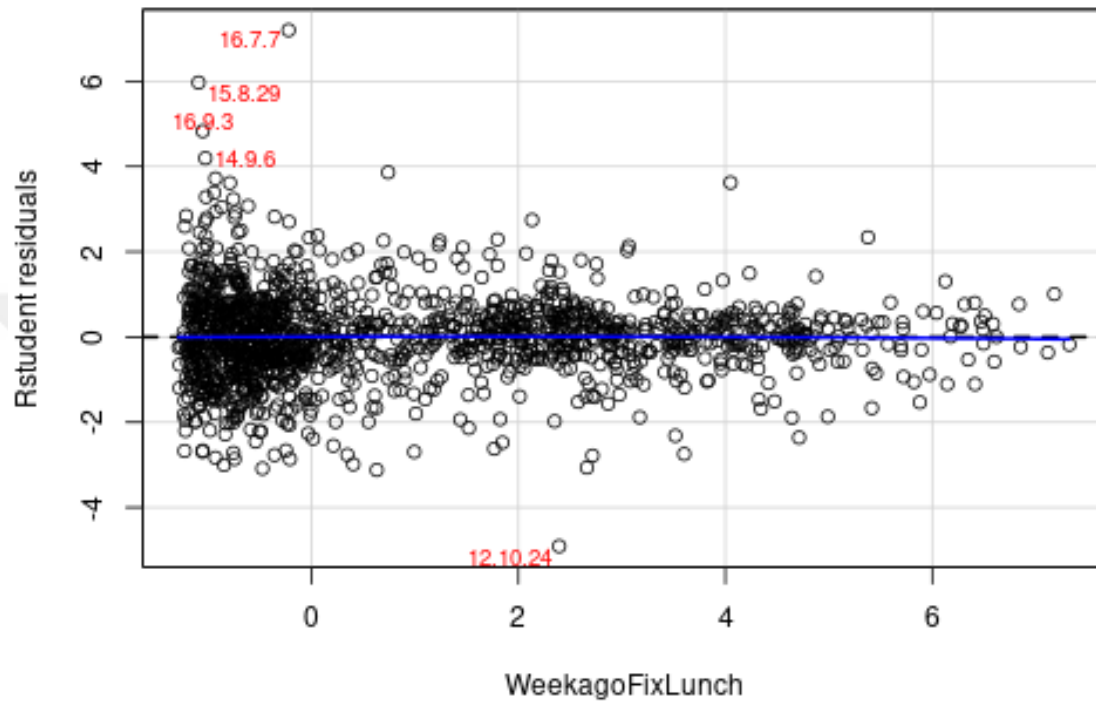


Figure 4.18: Studentized residuals vs temperature and fitted values

adding quadratic terms of the predictors in the model do not contribute to the model. Therefore, we did not perform any transformation for the predictors in the Table 4.2.

Table 4.2: T test results for quadratic effect

Predictor	Test stat	Pr(Test stat)
I(logPrevFixLunch ²)	1.772	0.0766
PrevFixLunch	1.3102	0.1904
logWeekagoFixLunch	-1.2265	0.2203
Temp	1.1143	0.2653
Prev2FixLunch	0.8561	0.3921
logPrev2FixLunch	0.7990	0.4244
WeekagoFixLunch	-0.702	0.4828
year	-0.4402	0.6598
logPrevFixLunch	-0.1018	0.9189
Tukey test	0.9677	0.3332

We exclude 5 points with leverage 1 to draw Q-Q plot with confidence envelope. Figure 4.19 shows that studentized residuals have a distribution with a fat tail and both the ends of the Q-Q plot to deviate from the straight line and its center follows a straight line.

Figure 4.20 shows the influential points by representing the observations proportional to the value Cook's distance. Vertical reference lines are drawn at twice and three times the average hat value, horizontal reference lines at -2, 0, and 2 on the studentized-residual scale. As one can see four data points whose dates are 2015-08-29, 2016-09-03, 2013-08-08 and 2013-10-15 appear as outliers but not influential since their Cook's distances in the plot are less than 0.5 and they fall outside the thresholds in terms of their residuals. However, their presence may inflate residual standard error (RSE). Moreover two points appear as influential

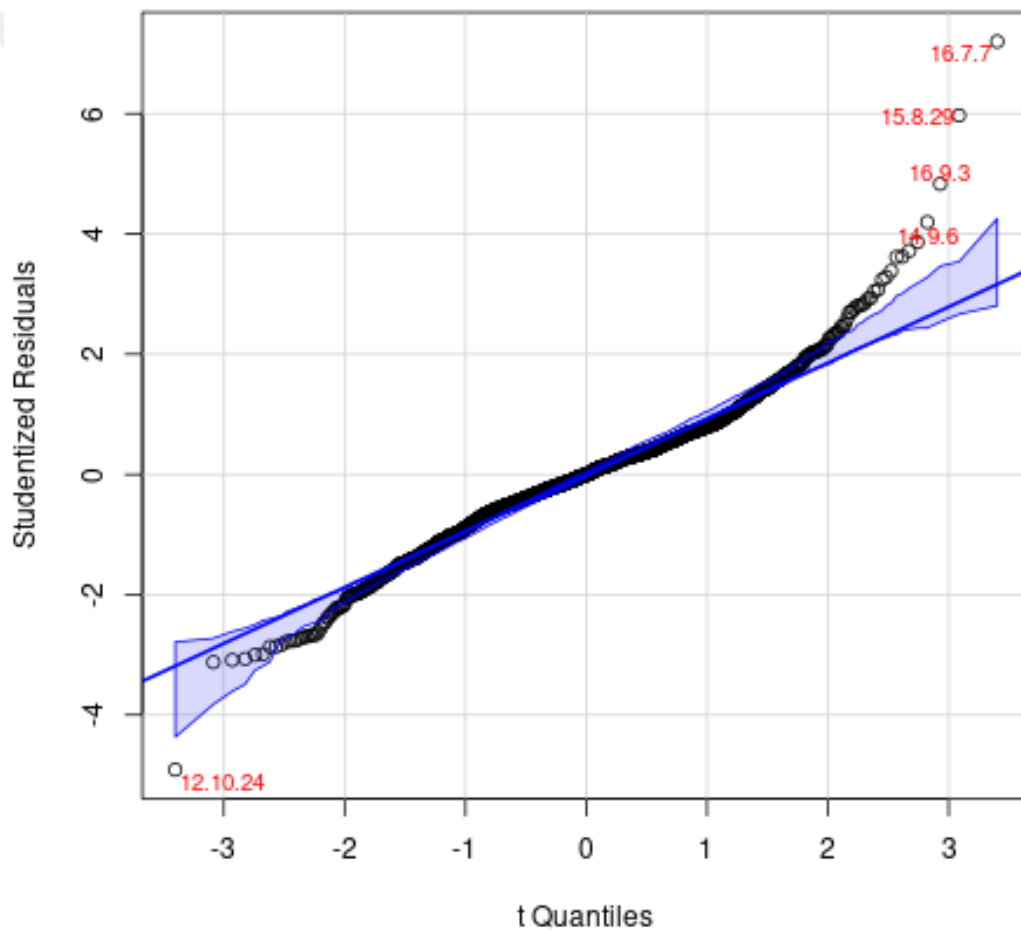


Figure 4.19: Q-Q plot with studentized residuals

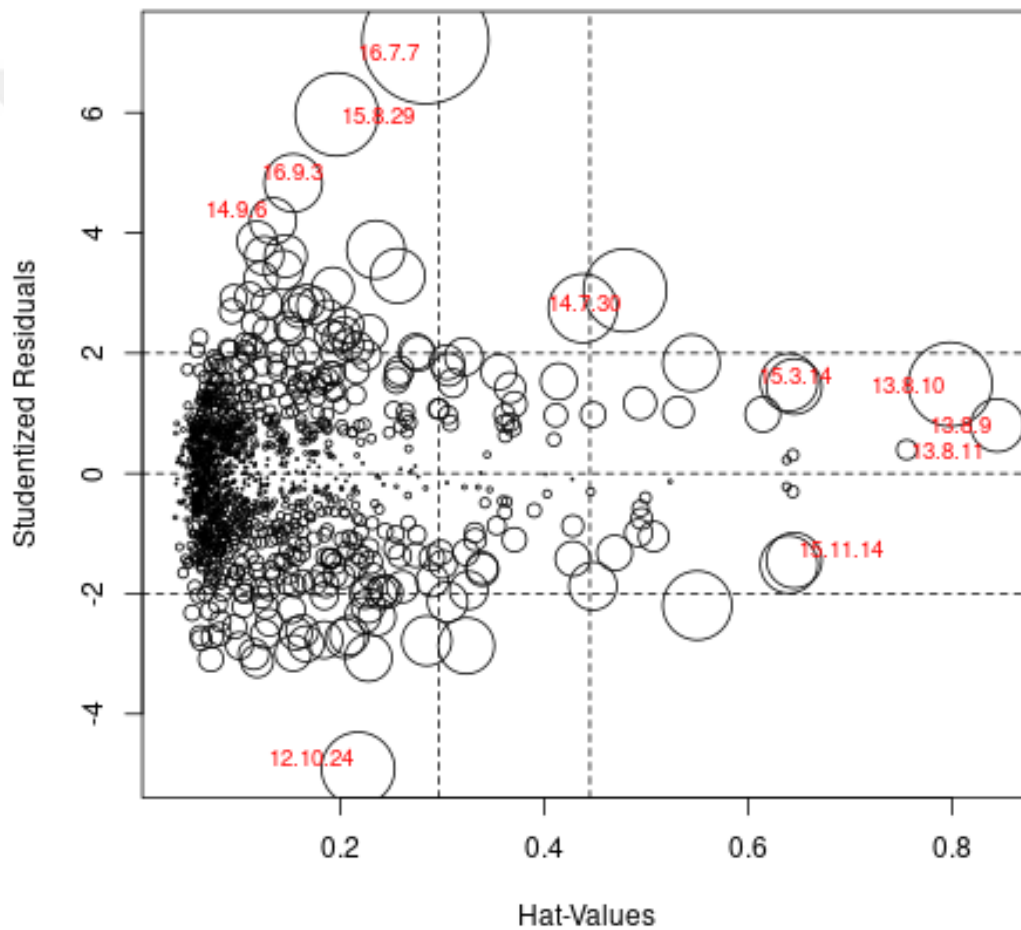


Figure 4.20: Influential points with dates. 7 cases (12.12.31, 13.8.7, 15.3.18, 15.7.16, 15.9.23, 16.5.19 and 16.7.4) have leverage equal to one. Plot has to be created without them because studentized residual is undefined when leverage equals one.

but not outliers. Therefore, we decided to exclude these points and compare the model's residual standard deviation with and without these points. The residual standard deviations of linear regression model with and without these points are 0.3090 and 0.3083 respectively. Since the difference between residual standard errors is insignificant, we decided to keep them in the data set.

Figure 4.21 shows the measure of the difference in each parameter estimate with and without the influential point. For every j , $DFBETA_j$ is the difference between betas calculated before and after the j^{th} data row is excluded. $DFBETAS_j$ is the standardized $DFBETA_j$. There exists a DFBETAS for each data point, if there are n observations and k variables, there will be $n \times k$ DFBETAS. In general, large values of DFBETAS indicate observations that are influential in estimating a given parameter. In general, it is recommended 2 as a general cutoff value to indicate influential observations and $\frac{2}{\sqrt{(n)}}$ as a size-adjusted cutoff.

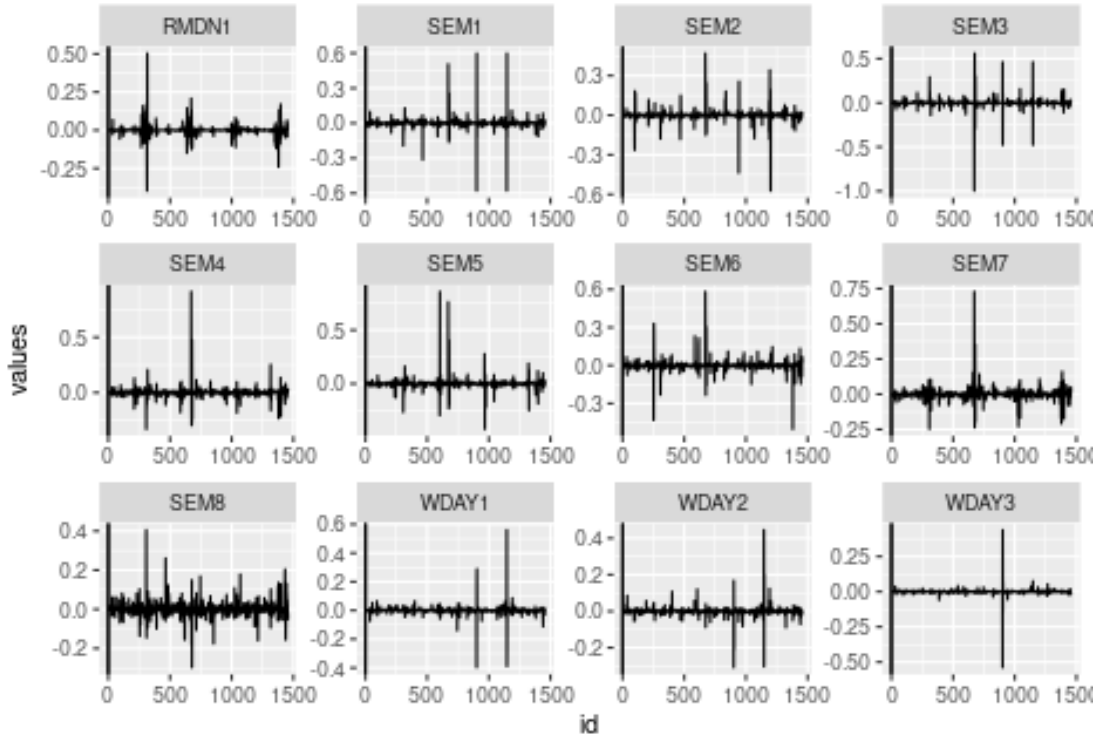


Figure 4.21: DFBETAS for different effects

To conclude, Figure 4.7 shows that the linear regression model's error assumption does not hold. One probable reason is that the data set contains outliers. On the other hand, when we closely analyze the outliers, we observed that they are not influential. Figure 4.21 shows that the coefficients of linear regression model are sensitive to data changes. To sum up, the errors do not have constant variance, therefore, the assumption, which is the errors have constant variance, for linear regression model does not hold which leads us to search for other methods such as Bayesian models.

4.3 Prediction

We fit the linear regression model to test data to evaluate model performance. The data set is divided into test and train and train set consists data about certain academic years from 2012-2013 to 2015-2016, represented as year 1,2,3 and 4 in the model.

Figure 4.22 shows the observed vs predicted values and residuals on test data respectively. We reported root-mean-square-error(RMSE), normalized RMSE (a method with the formula $RMSE/median(sales)$) and mean-absolute-percentage-error(MAPE).

Figure 4.22 shows that linear regression model performs good for lower sales(<1000) while it does not have the same performance with higher sales (>1000) and it overestimates. Residuals are fitted under zero meaning that linear regression model overestimates and the residuals are large during fall and spring semesters (October 2016 and April 2017). The residuals decrease when the academic semesters come to end especially in summer school and fall break. Since linear regression model tends to overestimate, this leads us to apply bayesian regression models.

The different measures for the linear regression model, namely RMSE, Normalized_RMSE, MAPE are 403, 0.6839, 33.5602% respectively.

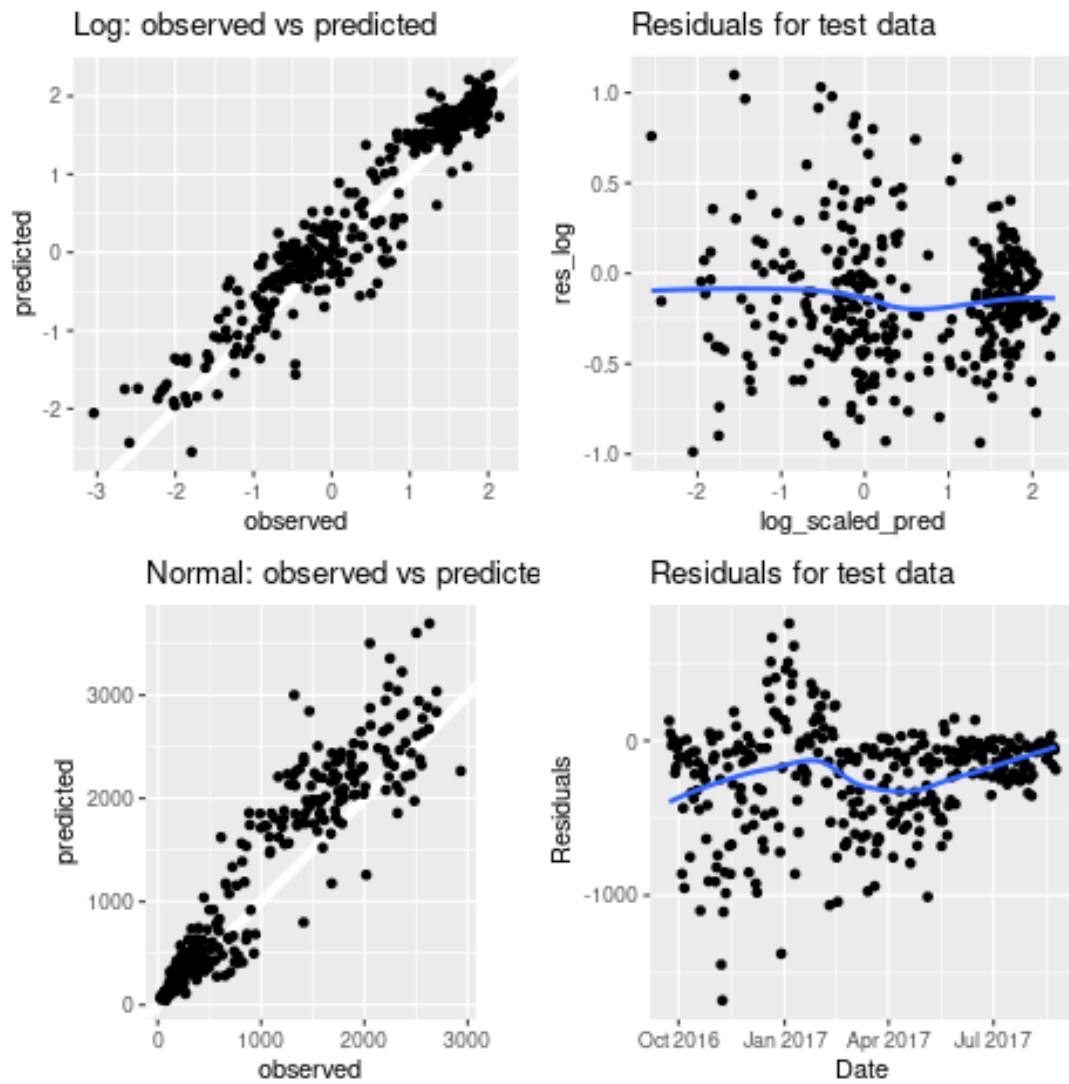


Figure 4.22: Log and normal scale estimated vs observed values

Chapter 5

Single Layer Hierarchical Model

In this model, we assumed that the sales on log-scale follow a hierarchical Bayesian linear regression model. The linear model can be defined as

$$\begin{aligned}\mathbb{Z} &= \log(y), \\ \mathbb{Z} &= \mathbb{X}\beta + \epsilon, \\ \epsilon &\sim \mathcal{N}(0, \sigma^2 \mathbb{I}_n),\end{aligned}\tag{5.1}$$

Figure 5.1 shows the Directed Acyclic Graph (DAG) model for the hierarchical Bayesian linear regression model. The joint distribution for this model is

$$p(\bar{\Sigma}, \bar{\beta}, \sigma^2, \beta, \mathbb{Z}) = p(\bar{\Sigma})p(\bar{\beta})p(\beta|\bar{\beta}, \bar{\Sigma})p(\mathbb{Z}|\mathbb{X}\beta, \sigma^2 \mathbb{I}_n)\tag{5.2}$$

with prior distributions

$$\begin{aligned}\beta|\bar{\beta}, \bar{\Sigma} &\propto \mathcal{N}(\bar{\beta}, \bar{\Sigma}), \\ \bar{\beta}|\bar{\bar{\beta}}, \bar{\bar{\Sigma}} &\propto \mathcal{N}(\bar{\bar{\beta}}, \bar{\bar{\Sigma}}), \\ \bar{\Sigma}|\nu, \mathbb{V} &\propto \mathcal{IW}(\nu, \mathbb{V}), \\ \sigma|\frac{\alpha}{2}, \frac{\lambda}{2} &\propto \mathcal{IG}(\frac{\alpha}{2}, \frac{\lambda}{2}).\end{aligned}\tag{5.3}$$

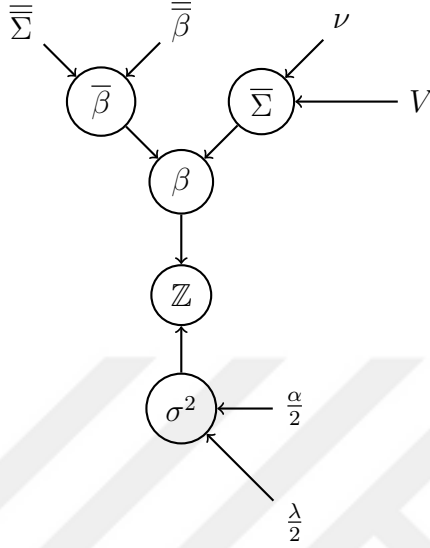


Figure 5.1: DAG for hierarchical linear model

The full conditional posterior distributions are

$$\begin{aligned}
p(\beta|\text{everything else}) &\propto p(\beta|\bar{\beta}, \bar{\Sigma}) \times p(\mathbb{Z}|X\beta, \sigma^2\mathbb{I}_n), \\
p(\bar{\beta}|\text{everything else}) &\propto p(\beta|\bar{\beta}, \bar{\Sigma}) \times p(\bar{\beta}|\bar{\beta}, \bar{\Sigma}), \\
p(\bar{\Sigma}|\text{everything else}) &\propto p(\bar{\Sigma}|\nu, \mathbb{V}) \times p(\beta|\bar{\beta}, \bar{\Sigma}), \\
p(\sigma^2|\text{everything else}) &\propto p(\sigma^2|\frac{\alpha}{2}, \frac{\lambda}{2}) \times p(\mathbb{Z}|X\beta, \sigma^2\mathbb{I}_n).
\end{aligned} \tag{5.4}$$

Then full conditional distributions become

$$\begin{aligned}
\beta|\text{everything else} &\propto \mathcal{N}(\tilde{\beta}, \sigma^2(\bar{\Sigma}^{-1} + X^T X)^{-1}), \\
\bar{\beta}|\text{everything else} &\propto \mathcal{N}(\hat{\beta}, (\bar{\Sigma}^{-1} + \sigma^2\bar{\Sigma}^{-1})^{-1}), \\
\bar{\Sigma}|\text{everything else} &\propto \mathcal{IW}(\nu + 1, \mathbb{V} + \frac{1}{\sigma^2}(\beta - \bar{\beta})(\beta - \bar{\beta})^T), \\
\sigma^2|\text{everything else} &\propto \mathcal{IG}\left(\frac{\alpha + n + p}{2}, \frac{\lambda + (\beta - \bar{\beta})^T(\bar{\Sigma})^{-1}(\beta - \bar{\beta}) + \|z - X\beta\|^2}{2}\right).
\end{aligned}$$

where

$$\begin{aligned}\tilde{\beta} &= (\bar{\Sigma}^{-1} + X^T X)^{-1} (X^T X \beta_{MLE} + \bar{\Sigma}^{-1} \bar{\beta}), \\ \hat{\beta} &= (\bar{\Sigma}^{-1} + \bar{\bar{\Sigma}}^{-1})^{-1} (\bar{\Sigma}^{-1} \beta + \bar{\bar{\Sigma}}^{-1} \bar{\beta}),\end{aligned}$$

and $\beta_{MLE} = (X^T X)^{-1} X^T z$ is the maximum likelihood estimator of β . When we carefully examine the parameters for posterior distributions of β and $\bar{\beta}$, the mean vectors of full conditional distributions are a convex combination of prior beliefs and *pure data* based least squares solution where the *weights* are proportional to prior and *pure data* based precision matrices.

Details of the calculations can be found in Appendix.

5.1 Convergence Analysis

We run single layer bayesian regression model with 5 chains and 6000 draws sequentially. The initial points for these chains are selected randomly and they are selected from full conditional distributions for covariates. Moreover, the mean and variance of the covariates are selected randomly with the hyper parameters as well.

Figure 5.2 shows the trace plots of coefficients including *year*, *SEM1*, *FixLunchhas_wmeat1*, *WDAY1*, *COPEcope11*. One can deduce that *year* coefficient has smoother line in positive values while *FixLunchhas_wmeat1* and *COPEcope11* have smoother in negative area. Moreover, *SEM1*, *WDAY1* have means in zero considering the smoother. Then, we can conclude that *year* has positive effect on sales meaning as year increases the sales increase while *FixLunchhas_wmeat1* (white meat is not offered) and *COPEcope11* (first stage of COPE does not occur) negatively affect the sales.

Figure 5.3 shows the trace plots of variance coefficients. First plot shows all draws including warm-up period while second plot shows the instances after 200 draws. As an interpretation of Figure 5.3, we can conclude that the chain for

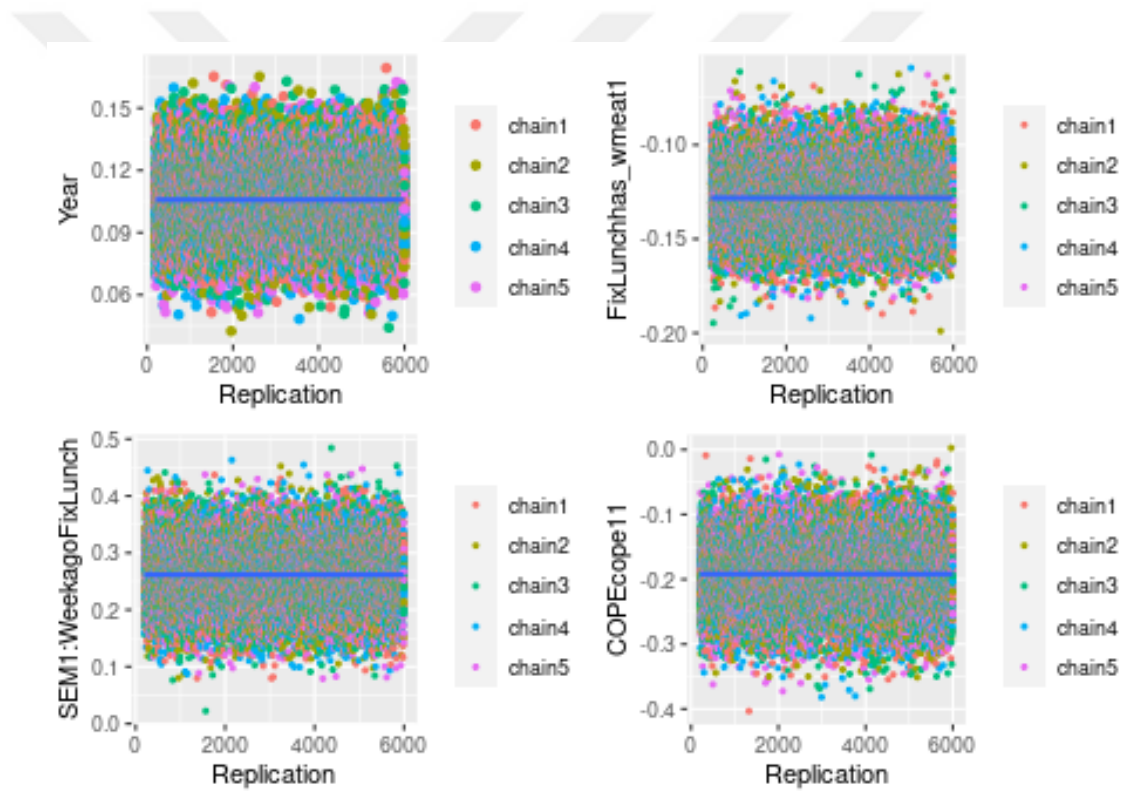


Figure 5.2: Trace plots for coefficients of year

variance converges after certain number of draws. On the other hand, calculating warm-up period and ensure convergence theoretically, we calculate potential scale reduction factor which will be referred as Gelman-Rubin statistics or $\hat{R}$ interchangeably.

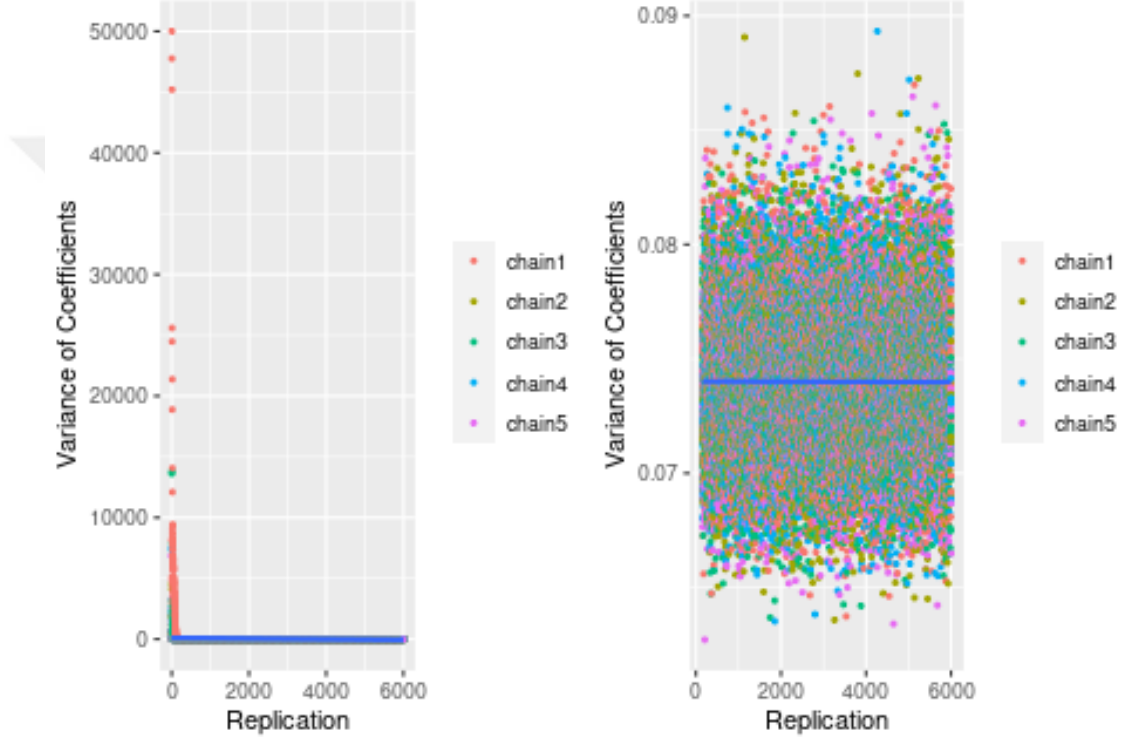


Figure 5.3: Trace plot for variance of coefficients

Figure 5.4 shows the $\hat{R}$ values for all covariates and their variances. Even though Gelman and Rubin recommends the threshold 1.1 for $\hat{R}$, we decided to level down it to 1.02 to ensure convergence. After $\hat{R}$ calculation, we decided our burn-in period. For each chain, we decided to form a sequence of starting points with increment of 100 and this sequence starts from 100. We extracted the first point among these starting points in which $\hat{R}$ value fell below our threshold. We excluded the draws between zero to this point. To exemplify, first element of this sequence calculates $\hat{R}$ of all covariates consisting draws from 100 to 6000 and second element calculates $\hat{R}$ consisting draws from 200 to 6000 etc. Figure 5.4 shows that $\hat{R}$ for all covariates with all intervals is below 1.02. Therefore, one can conclude that first element of above-mentioned sequence creates $\hat{R}$ below

1.02. Since $\hat{R}$ is already below 1.02 for all the covariates and their variances, the chains converged before second element of the sequence. In other words, the chains converge after 100 and before 200 draws with fast mixing rates.

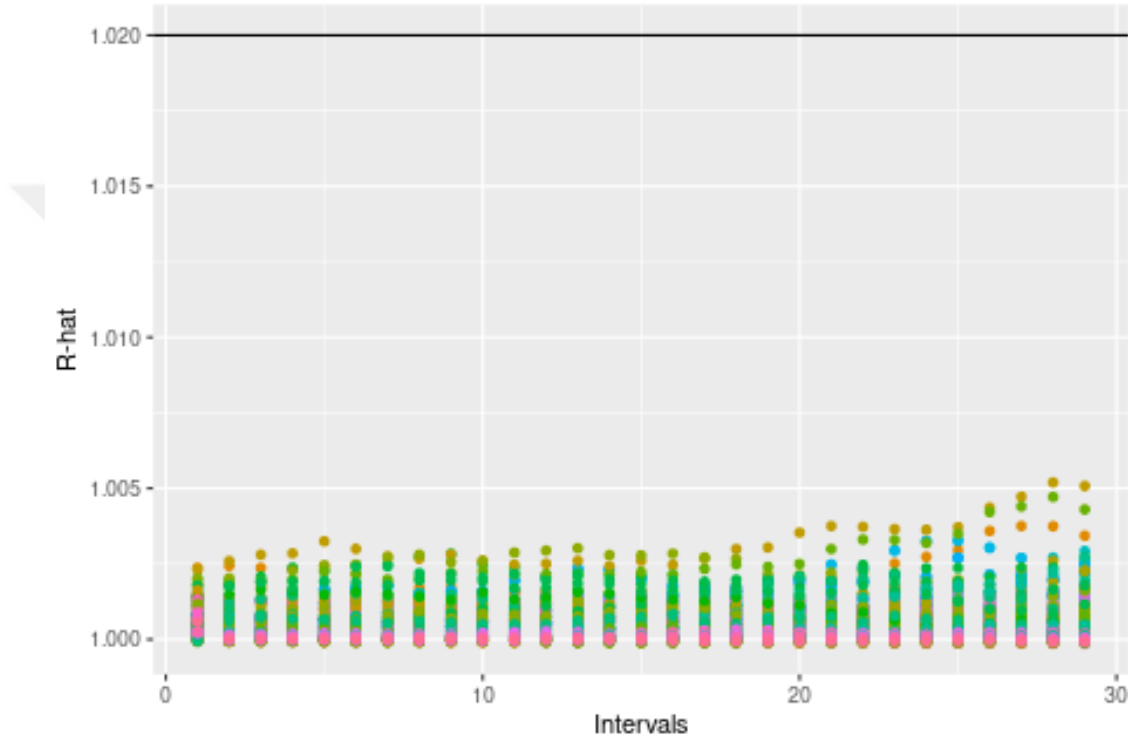


Figure 5.4: R-hat statistics for variance and all covariates with different interval lengths

5.2 Diagnostics

Figure 5.5 shows predicted values on train data for different chains. Predicted and observed values are close to each other for each chain. Therefore, one can conclude that the chains are likely to converge.

Figure 5.6 shows the bar chart of t-statistics of significant effects whose t-statistic is greater than two in absolute value for first chain of the single layer model. The variables such as *FixLunchhas_wmeat1*, *year*, *FixLunchhas_kof1*,

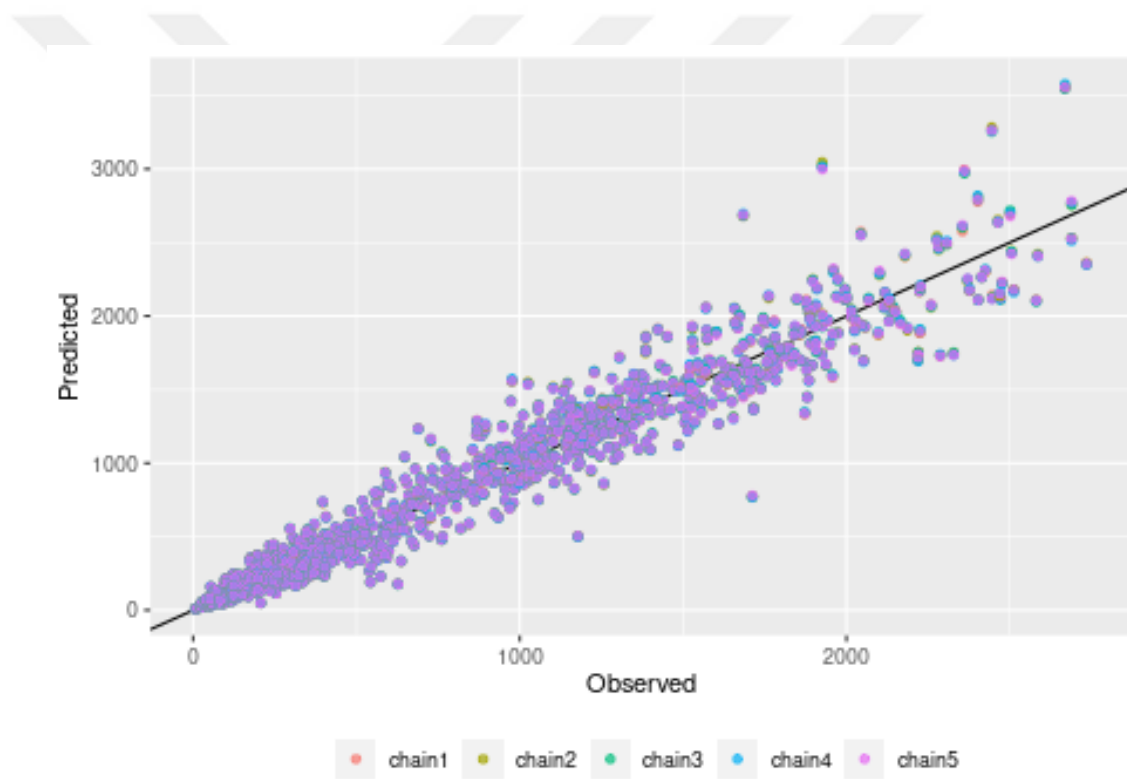


Figure 5.5: Observed vs predicted on train data

SEM1: WeekagoFixLunch, *SEM2:logPrevFixLunch* have high absolute value of t-statistic. *FixLunchhas_wmeat1* has negative t statistics meaning that the lunch without white meat such as chicken has negative impact on sales. Similarly, *FixLunchhas_kof1* corresponding to no meatball on menu is not preferred.

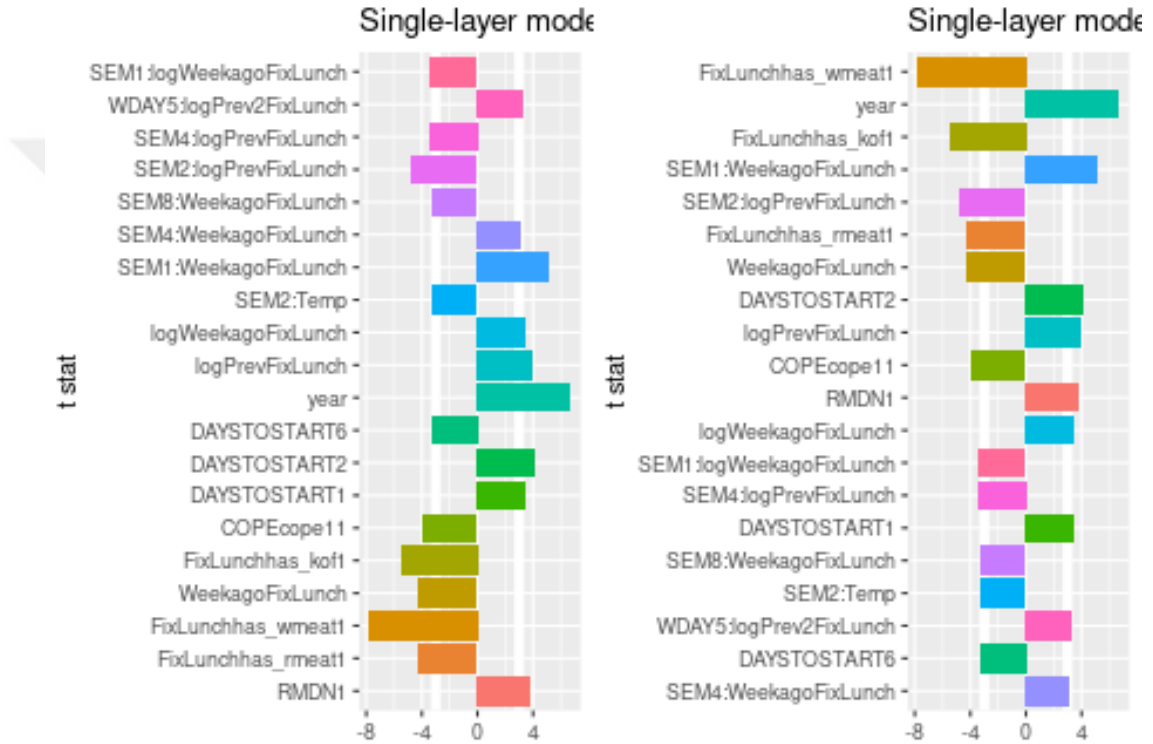


Figure 5.6: t-statistics for single layer model. The white thick vertical lines mark -2 and +2 values. The bars show the t-statistics for only those predictors whose t-statistics are greater than two in absolute value.

Chapter 6

Two Layers Bayesian Regression Model

In this model, we assume that the number of lunch meals being sold follows two layers Bayesian regression model.

$$\begin{aligned} z_a &= \log(y_a), a = 2012, \dots, 2016 \\ z_a &= X_a \beta_a + \epsilon_a, a = 2012, \dots, 2016 \\ \epsilon_a &\sim \mathcal{N}(0, \sigma_a^2 \mathbb{I}_{n_a}), a = 2012, \dots, 2016 \end{aligned} \tag{6.1}$$

Here, y_a stands for the number of fix meal being sold during lunch time and, β_a is p -dimensional vector representing covariates' effects for a^{th} academic year. Our assumption for this model is that covariates' effects β_a have apriori same distribution with parameters which are distributed with same hyper-parameters, in this case, $\mathcal{N}(\bar{\beta}, \bar{\Sigma})$. The random variables $\bar{\beta}$ and $\bar{\Sigma}$ represent the mean vector and covariance matrix of β respectively. Early 'empirical Bayes' approaches simply estimate these parameters, and then conditional on these estimates, perform an approximate Bayesian analysis of regression. A full Bayesian solution can be obtained by specifying a further 'level' of priors on $\bar{\beta}$ and $\bar{\Sigma}$.

Prior to modern simulation methods, a premium was placed on models that

would allow analytical expressions for various posterior summaries. Typically, this means that we choose models for which we can compute posterior moments analytically. We desire that the posterior be a distributional form for which the posterior moments are available in analytical expressions. This requirement imposes constraints on both the choice of likelihood and prior distributions [11].

In this model, we defined the prior for β_a draws is specified through process with two levels. First, we define normal prior for β_a and second level prior for the parameters of this distribution for different academic years.

This model can be expressed as a directed acyclic graph as in Figure 6.1:

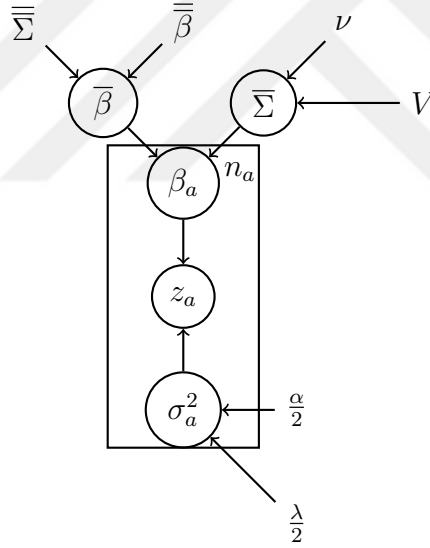


Figure 6.1: DAG for multilevel hierarchical bayesian regression model

The prior distributions are

$$\begin{aligned}
 \beta_a | \bar{\beta}, \bar{\Sigma}, \mathbb{Z} &\propto \mathcal{N}(\bar{\beta}, \bar{\Sigma}), a = 2012, \dots, 2016 \\
 \bar{\beta} | \bar{\bar{\beta}}, \bar{\bar{\Sigma}} &\propto \mathcal{N}(\bar{\bar{\beta}}, \bar{\bar{\Sigma}}) \\
 \bar{\Sigma} | \nu, \mathbb{V} &\propto \mathcal{IW}(\nu, \mathbb{V}) \\
 (\sigma_a)^2 | \frac{\alpha}{2}, \frac{\lambda}{2} &\propto \mathcal{IG}(\frac{\alpha}{2}, \frac{\lambda}{2}), a = 2012, \dots, 2016
 \end{aligned} \tag{6.2}$$

The joint distribution for hierarchical Bayesian linear regression is

$$f\left((z_a, \beta_a, \sigma_a^2)_{a=2012 \dots 2016}, \bar{\beta}, \bar{\Sigma} | \bar{\beta}, \bar{\Sigma}, \nu, \mathbb{V}, \frac{\alpha}{2}, \frac{\lambda}{2}\right) =$$

$$f(\bar{\beta} | \bar{\beta}, \bar{\Sigma}) f(\bar{\Sigma} | \nu, \mathbb{V}) \prod_{a=2012}^{2016} f\left(\sigma_a^2 | \frac{\alpha}{2}, \frac{\lambda}{2}\right) f(\beta_a | \bar{\beta}, \bar{\Sigma}) f(z_a | X_a \beta_a, \sigma_a^2 \mathbb{I})$$
(6.3)

Full conditional distributions are as follows:

$$f(\beta_a | \text{everything else}) \propto f(\beta_a | \bar{\beta}, \bar{\Sigma}) \times f(z_a | X_a \beta_a, \sigma_a^2 \mathbb{I})$$

$$f(\bar{\beta} | \text{everything else}) \propto f(\bar{\beta} | \bar{\beta}, \bar{\Sigma}) \prod_{a=2012}^{2016} f(\beta_a | \bar{\beta}, \bar{\Sigma})$$

$$f(\bar{\Sigma} | \text{everything else}) \propto f(\bar{\Sigma} | \nu, \mathbb{V}) \prod_{a=2012}^{2016} f(\beta_a | \bar{\beta}, \bar{\Sigma})$$

$$f(\sigma_a^2 | \text{everything else}) \propto f\left(\sigma_a^2 | \frac{\alpha}{2}, \frac{\lambda}{2}\right) f(z_a | \sigma_a^2 \mathbb{I}_{\kappa_{\mathcal{D}} \times \kappa_{\mathcal{D}}}, X_a \beta_a)$$
(6.4)

The descriptions for full conditional distributions are

$$p(\beta_a | \text{everything else}) \propto \mathcal{N}(\tilde{\beta}', \sigma^2((\bar{\Sigma})^{-1}(\sigma_a)^2 + X_a^T X_a)^{-1})$$

$$p(\bar{\beta} | \text{everything else}) \propto \mathcal{N}(\hat{\beta}', ((\bar{\Sigma})^{-1} + \sum_{a=2012}^{2016} (\bar{\Sigma})^{-1})^{-1})$$

$$p(\bar{\Sigma} | \text{everything else}) \propto \mathcal{IW}(\nu + p + 1, \mathbb{V} + \sum_{a=2012}^{2016} (\beta_a - \bar{\beta})(\beta_a - \bar{\beta})^T)$$

$$p(\sigma_a^2 | \text{everything else}) \propto \mathcal{IG}\left(\frac{\alpha + n_a}{2}, \frac{\|z_a - X_a \beta_a\|^2}{2}\right)$$
(6.5)

Given the conditional structure of the model of interest, the Gibbs sampler has been used to generate a N-sample from the posterior distribution. Computation using Gibbs sampler [18] has made Bayesian estimation in these models feasible. The Gibbs sampler becomes a candidate for evaluating the joint posterior distribution since it relies on the availability of all full conditional distributions with well-known forms. To explain more, the applications of the Gibbs sampler requires two conditions to be met. First, we require that the marginal probability density functions for each of the component distributions

are not available in closed form; without this condition there would be no need to make any numerical approximation to the posterior. Second, we require that each of the full conditional distributions comprising the joint posterior have well-known forms. A consequence of using Gibbs sampling is that an experimenter is responsible of defining values for hyper parameters that are in deep levels of a hierarchy. In our model, we select loose and uninformative hyper parameters to ensure $\bar{\beta}$ and $\bar{\Sigma}$ to move around the state space.

We run MCMC in R with openblas. We draw 25,000 and 50,000 samples with five MCMC chains for hierarchical model and multilevel hierarchical bayesian regression model respectively. We take parallel runs with computation time 13 hours. In the multilevel hierarchical bayesian regression model, we prefer not to use parallel computation since it does not offer considerable advantage. The computation times for this model are 376, 394, 382, 367 and 352 minutes for different chains. The machine has 8 cores with multiple users and others users on the machine have an impact on the variability of computation times.

6.1 Convergence Analysis

We run two layers bayesian regression model with 5 chains and 25000 draws sequentially. Since this model has two layers, we assume that the warm-up period will be longer compared to single layer model. The initial points for these chains are selected randomly and they are selected from full conditional distributions for covariates. Since covariates for different academic years have different mean vectors and covariance matrices, the initial points are likely to be far away from each other for each chain and year.

Figure 6.4 represents $\hat{R}$ values for covariates' effects. Similar to single layer model, the chains converge after 100th draw and before 200nd draw. Even though we expect the warm-up period longer, this results are logical considering selection of diffuse hyperparameters.

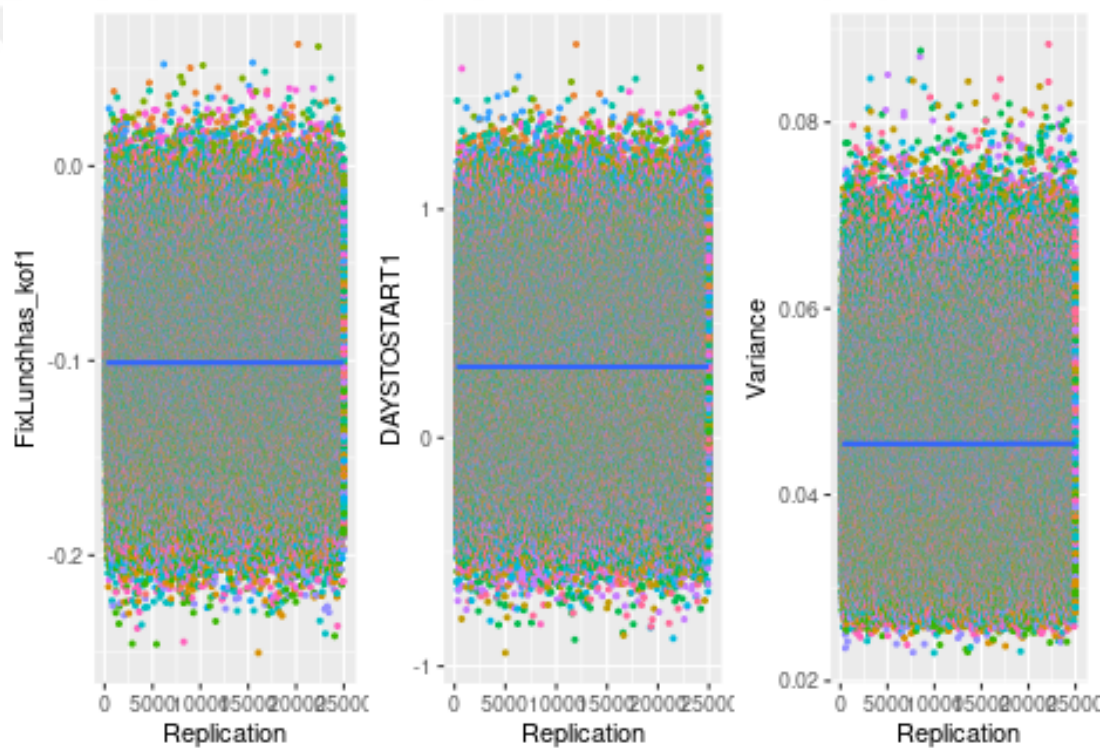


Figure 6.2: Trace plots of `FixLunchhas_kof1`(meatball is not offered), `DAYSTOSTART1`(1 day is left to beginning of the academic year) and variance of coefficients

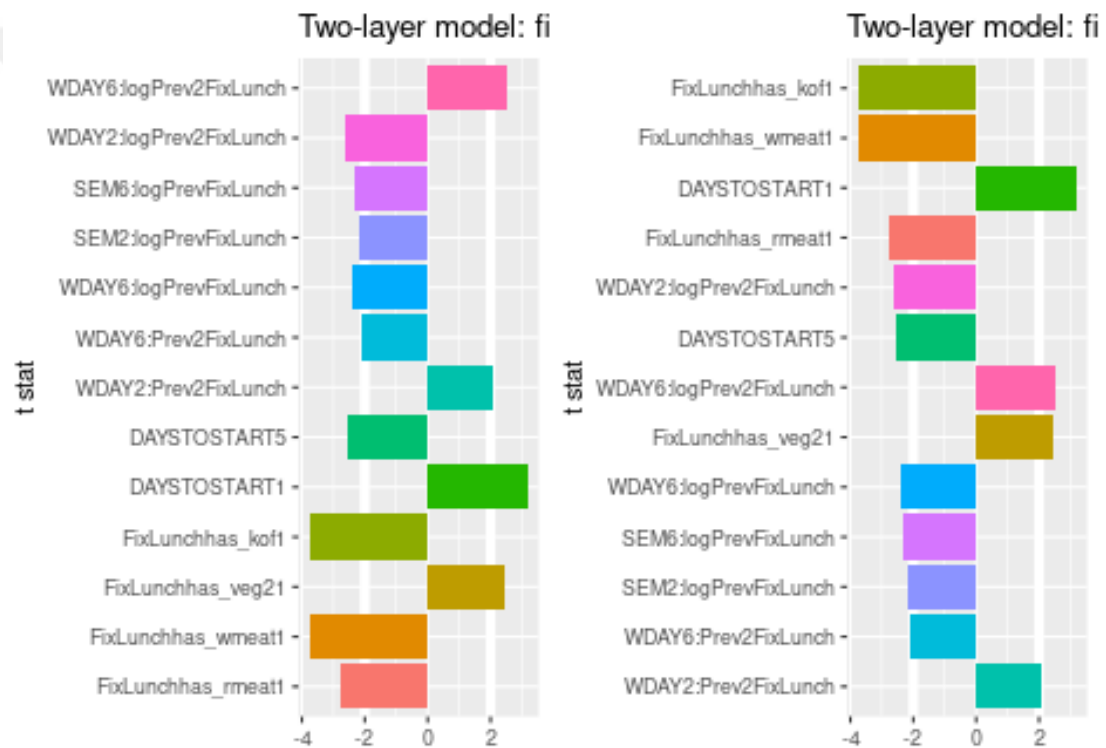


Figure 6.3: t-statistics for two layers model for first chain and first academic year. The white thick vertical lines mark -2 and +2 values. The bars show the t-statistics for only those predictors whose t-statistics are greater than two in absolute value.

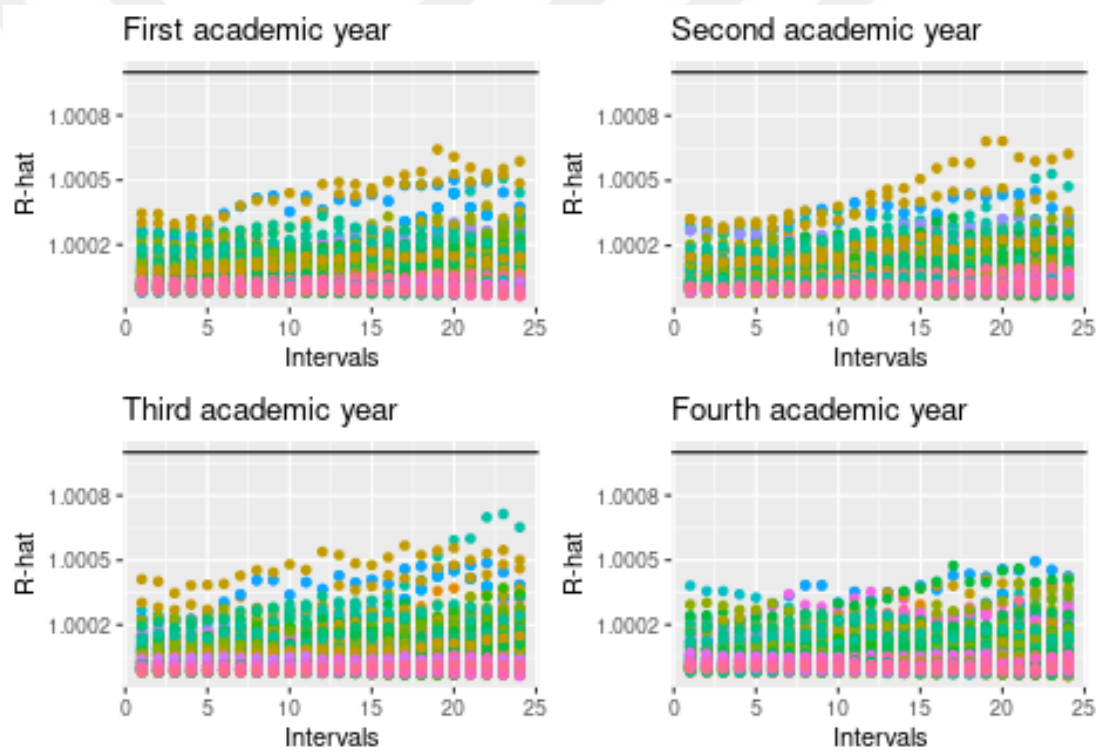


Figure 6.4: Gelman-Rubin statistic for different intervals for two layer bayesian regression

6.2 Posterior Distributions

Posterior distributions deliver the effect of prior and data information. From our models, the posterior is plotted as in Figure 6.5. Combining the chains after burn-in, we observe that density of *DAYSTOSTART5*. For this β , two layer model provides tighter distribution with lower variance. However, means change through years for two layers model.

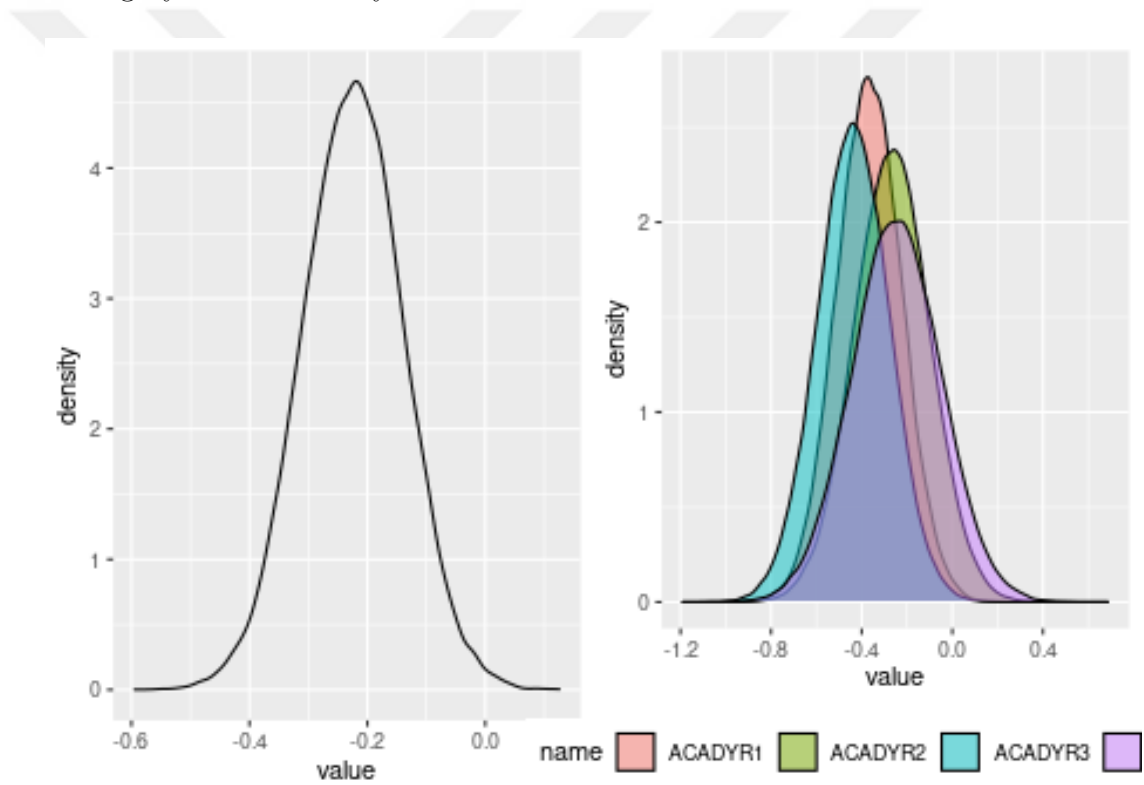


Figure 6.5: *DAYSTOSTART5* density for through years under single layer model and two layers model respectively

Chapter 7

Forecasting

We divide the data set into train and test set by academic years. The last academic year, 2016-2017, forms the test set and train set is composed of the remaining academic years' data. We perform our analysis on train set and using results from the analysis, we forecast the sales on 2016-2017. As it is previously mentioned, we perform pre-processing steps in data. Before starting forecasting, we perform scale-back steps. For the model without academic years' effect differentiation, let $p(\beta, \sigma^2 | X)$ be the joint posterior distribution of β and σ^2 given past sales data. Then,

$$\begin{aligned}\mathbb{E}[y|Past Data] &= \mathbb{E}\left[\mathbb{E}(y|\mathbb{X}, \beta, \sigma^2, Past Data)|Past Data\right] \\ &= \mathbb{E}\left[e^{\frac{\sigma^2}{2} + \mathbb{X}^T \beta} | Past Data\right] \\ &= \int e^{\frac{\sigma^2}{2} + \mathbb{X}^T \beta} \{(\beta, \sigma^2 | Past Data)\} d\beta d\sigma^2 \\ &\simeq \frac{1}{N} \sum_{i=1}^N e^{\frac{\sigma_i^2}{2} + \mathbb{X}^T \beta_i}\end{aligned}\tag{7.1}$$

where i represents the number of draws and $N = 6000$. Moreover, a robust predictor will be the sample median of $e^{\frac{\sigma_i^2}{2} + \mathbb{X}^T \beta_i}$. We calculate the $(1 - \alpha)100\%$ interval with lower and upper bounds. Observed vs predicted values are plotted

in Figure 7.1 for single layer model.

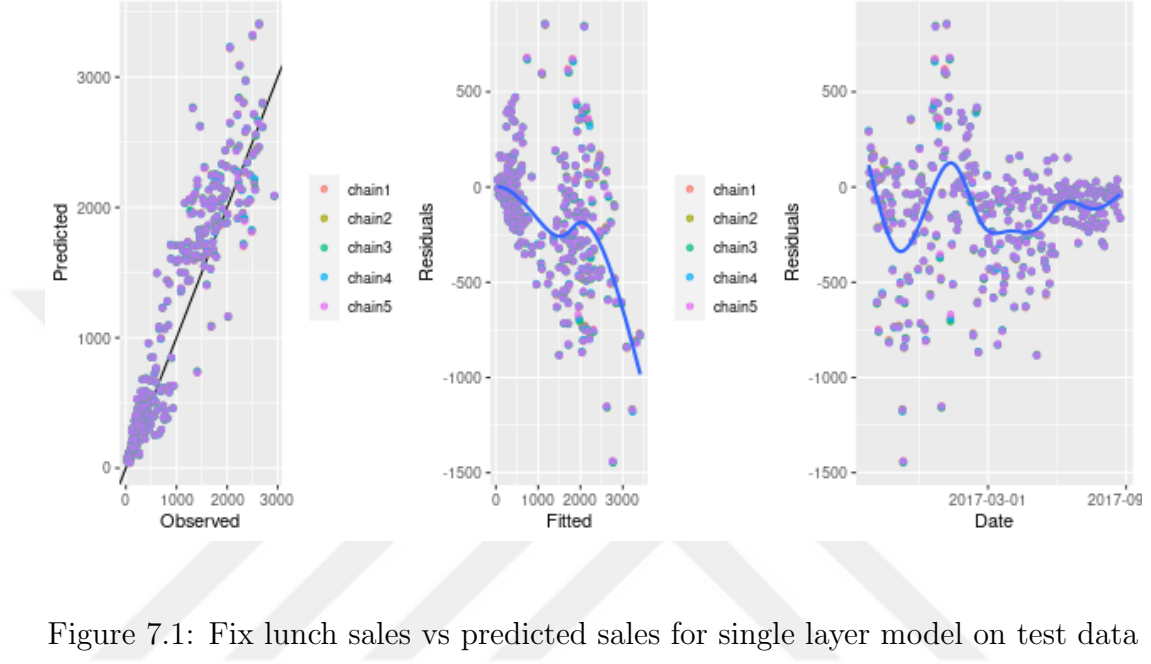


Figure 7.1: Fix lunch sales vs predicted sales for single layer model on test data

We fitted a line with slope of 1 to examine whether observed and predicted values are close to each other. Instances are generally accumulated around this line. However, one can deduce that for lower values the instances are closer to this line while larger observed predicted values are far away. Moreover, predicted values appear to be above the line meaning that the model overestimates when the observed values are higher than 1000. These findings are similar to linear regression model results which can be seen from Figure 4.22.

Then, we examined the results of two layer bayesian regression model. Since two layer bayesian regression model has different covariates' effects for each year, we adjusted this procedure. Firstly, we extracted σ_a^2 values from MCMC chains.

Figure 7.2 shows σ_a^2 values fluctuates among academic years. In case there is a pattern, we expect σ_a^2 to decrease again and therefore, $\sigma_{a=2014-2015}^2$ will be a fine assumption for forecasting. Moreover, we need to decide $\beta_{a=2016-2017}$ and we develop two methods and compare the results. First method is to use the last year's covariates' variance directly. Secondly, the mean of last two years' covariates' variance will be used to form MCMC chain for the test year.

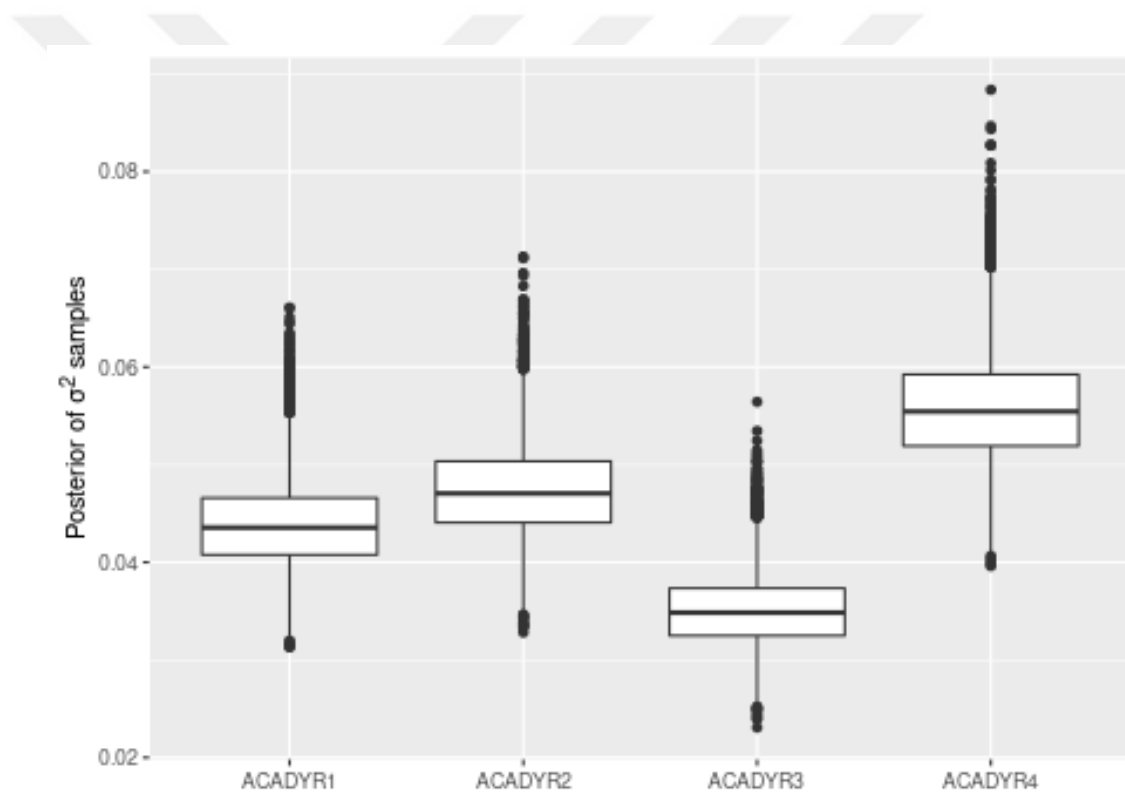


Figure 7.2: Variance values through Academic Years

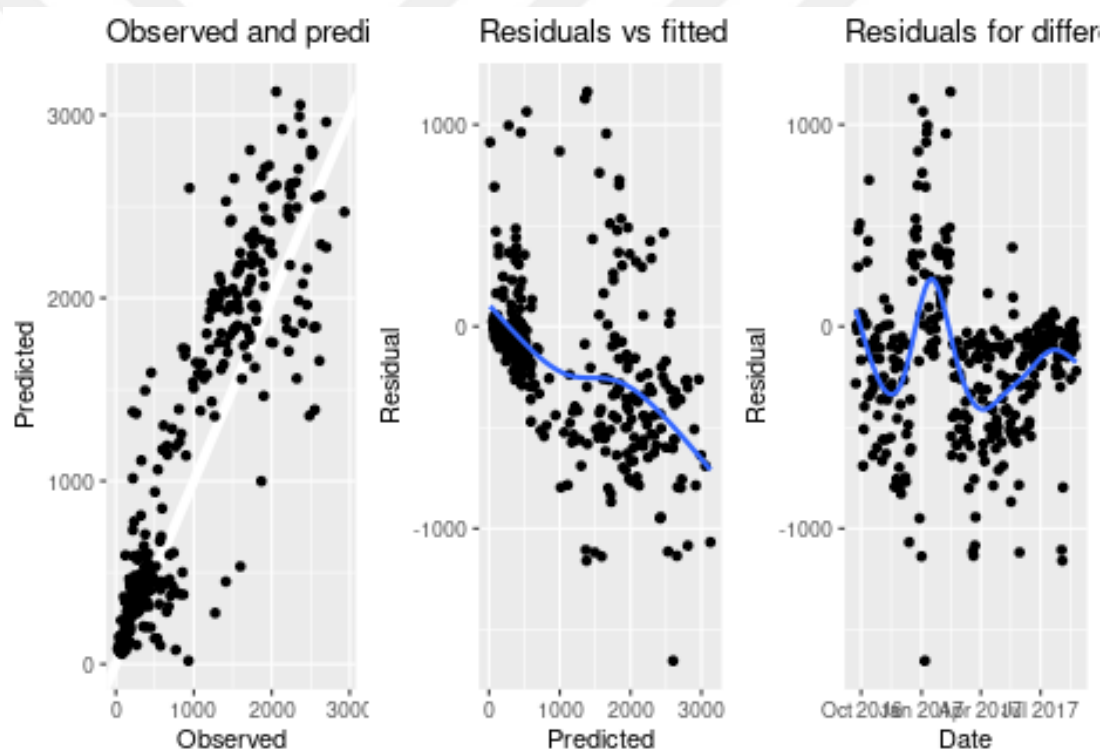


Figure 7.3: Sampling covariates of with last year variance

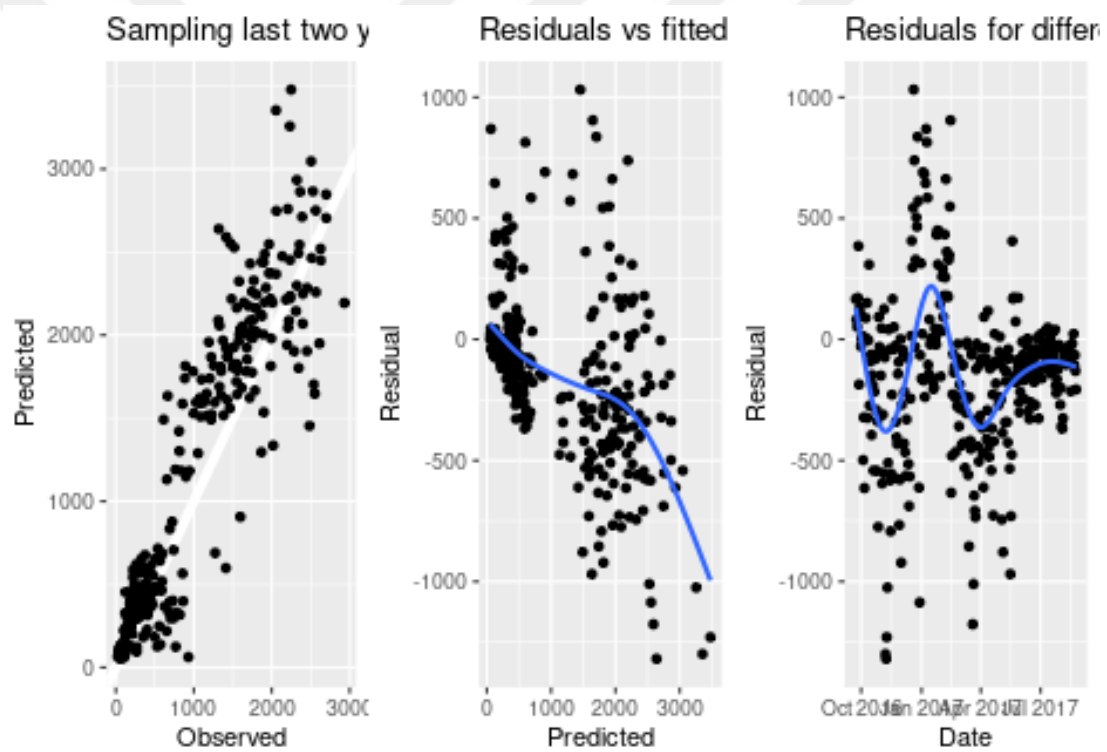


Figure 7.4: Sampling covariates of with mean of last two years variance

Table 7.1: Measures of accuracy comparison for different prediction techniques

Technique	RMSE	NormalizedRMSE	MAPE
Sampling covariates of with mean of last two years variance	370.1331	0.6273	29.8487
Sampling covariates of with last year variance	429.4903	0.7280	35.1883
Single layer model	325.0843	0.5510	26.3203

Table 7.1 shows the accuracy measures of two layer model with different selection choices for variance of variances. One can conclude that we have more accurate prediction when we use last two years' variances than last year's variance. Moreover, Figure 7.2 shows that the variances are likely to have a pattern which is an increase followed by decrease. Therefore, we suspected the variances are likely to be close to variance of 3rd year(2014-2015). Measures validate that using 3rd and 4th year(2015-2016) leads more accurate results.

We also decide to divide test data into eleven four weeks blocks and update forecast models after every block. Since the holiday intervals and/or beginning of the academic years alters, each block does not necessarily contain the same number of instances. On the other hand, one can assume that the number of instances in each block is similar. To exemplify, first block of train data represents first four weeks of first academic year. We calculated accuracy measures of forecasts for the next four weeks for linear regression, single layer and two layers model with common variance as a variation to two layer model. These results are displayed in Table 7.2 and show that linear regression have better scores in performance measures, namely RMSE, MAE, MAPE on test data. One can notice that the scores fluctuate. One reason for this fluctuation is date. Especially, these scores decrease for semester breaks considering the low number of instances in these dates.

Table 7.2: The comparisons of linear regression, single layer and two layer models updated every four week block of 2016-2017 and tested on the next four week block

Accuracy of Measures	Blocks										
	01	02	03	04	05	06	07	08	09	10	11
σ^2	0.07	0.071	0.071	0.073	0.075	0.076	0.076	0.076	0.078	0.078	0.078
$\frac{100 \times \sigma}{median(y)} \%$	3.62	3.67	3.66	4.28	3.76	3.83	3.96	4.53	5.36	4.94	5.52
log train RMSE	0.25	0.25	0.25	0.25	0.25	0.26	0.26	0.26	0.26	0.26	0.26
train RMSE	142.25	147.9	150.15	159.75	161.41	163.1	164.56	165.38	164.79	163.03	162.12
train MAE	89.84	92.96	94.86	99.81	101.05	102.51	103.79	104.73	104.76	103.46	102.75
train MAPE	12.26	12.34	12.50	12.48	12.81	12.62	12.61	12.69	12.65	12.6	12.8
log test RMSE	0.30	0.22	0.52	0.49	0.24	0.24	0.3	0.38	0.43	0.3	0.42
test RMSE	530.60	307.68	502.18	210.57	318.82	314.06	277.95	218.8	90.43	136.27	77.17
test MAE	386.62	229.62	428.62	163.16	247.19	234.48	214.49	167.98	64.43	110.96	51.88
test MAPE	20.40	12.67	40.91	35.69	15.92	16.24	17.98	25.55	23.31	26.1	24.62

Accuracy of measures	Blocks										
	01	02	03	04	05	06	07	08	09	10	11
$median(\sigma^2)$	0.07	0.071	0.071	0.073	0.075	0.076	0.076	0.076	0.078	0.078	0.078
$\frac{100 \times \sigma}{median(y)} \%$	4.4	4.39	4.38	4.42	4.48	4.49	4.47	4.47	4.53	4.54	4.57
log train RMSE	0.24	0.25	0.25	0.25	0.25	0.26	0.26	0.26	0.26	0.26	0.26
train RMSE	142.3	147.77	150.3	159.6	161.3	163.2	164.5	165.31	164.62	163.15	162.18
train MAE	89.84	92.98	94.9	99.76	100.98	102.58	103.78	104.72	104.66	103.53	102.77
train MAPE	12.3	12.38	12.55	12.38	12.81	12.63	12.71	12.62	12.55	12.57	12.84
log test RMSE	0.30	0.22	0.52	0.49	0.24	0.23	0.30	0.38	0.43	0.39	0.42
test RMSE	530.47	308.09	502.49	209.14	316.18	313.47	279.62	220.74	89.81	136.87	77.79
test MAE	386.83	229.86	429.20	162.81	245.39	233.99	215.89	168.99	63.97	111.63	52.27
test MAPE	20.50	12.95	41.23	35.57	15.90	16.25	17.94	25.34	23.28	26.14	24.78

	Blocks										
Accuracy of measures	01	02	03	04	05	06	07	08	09	10	11
median ssq	0.04	0.04	0.04	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
2012-2013 log train RMSE	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.16	0.16
2013-2014 log train RMSE	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16
2014-2015 log train RMSE	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.14	0.13
2015-2016 log train RMSE	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.18	0.18
2016-2017 log train RMSE	0.004	0.05	0.08	0.12	0.12	0.12	0.13	0.14	0.14	0.15	0.15
2012-2013 train RMSE	72.2	72.3	72.16	72.14	72.05	72.26	72.17	72.26	72.2	72.08	71.93
2013-2014 train RMSE	92.99	92.9	92.62	92.7	92.31	92.64	93.02	92.75	92.42	92.66	92.66
2014-2015 train RMSE	93.84	93.93	94.06	94	94.13	93.82	94.12	94.14	93.49	93.83	93.93
2015-2016 train RMSE	131.48	131.28	131.36	131.87	131.81	131.84	132.11	131.97	131.96	131.65	131.85
2016-2017 train RMSE	37.39	105.77	122.61	182.9	165.1	173.37	170.97	170.13	166.95	159.57	155.04
2012-2013 train MAE	47.63	47.7	47.59	47.45	47.48	47.63	47.53	47.61	47.51	47.50	47.47
2013-2014 train MAE	56.52	56.31	56.16	56.31	56.16	56.34	56.5	56.31	56.27	56.38	56.29
2014-2015 train MAE	61.24	61.54	61.44	61.5	61.56	61.51	61.53	61.59	61.41	61.38	61.45
2015-2016 train MAE	85.35	85.26	85.13	85.48	85.47	85.28	85.54	85.48	85.41	85.29	85.42

Blocks

(continued on the next page)

		Blocks										
(continued from the previous page)		01	02	03	04	05	06	07	08	09	10	11
2016-2017		32.88	68.47	90.05	125.12	106.42	113.72	110.25	111.06	106.98	99.17	92.78
train MAE												
2012-2013		7.88	7.84	7.69	7.7	7.85	7.74	7.87	7.74	7.7	7.95	7.72
train MAPE												
2013-2014		8.14	8.11	8.08	8.05	8.09	8.17	8.22	8.15	8.13	8.08	8.01
train MAPE												
2014-2015		6.50	6.42	6.44	6.5	6.51	6.66	6.54	6.54	6.58	6.57	6.5
train MAPE												
2015-2016		9.12	9.05	9.18	9.14	9.18	9.14	9.16	9.09	9.23	9.14	9.34
train MAPE												
2016-2017		1.87	2.88	4.56	6.23	5.32	6.36	6.32	7.3	7.79	8.27	8.3
train MAPE												
log test		0.47	0.35	0.9	0.88	0.46	0.34	0.44	0.62	0.91	0.56	0.53
RMSE												
test RMSE		942.64	419.71	1624.96	317.74	506.39	468.39	566.71	378.89	241.7	254.44	77.82
test MAE		679.69	321.61	826.76	241.85	388.79	373.48	441.64	302.02	143.02	167.78	63.3
test MAPE		34.86	20.87	51.61	57.5	24.79	23.22	32.4	43.39	41.87	31.87	27.68

Chapter 8

Conclusion

In the catering industry, forecasting sales is a significant issue regarding the amount of waste produced and the business loss and consequently, predicting sales amount precisely is benefited by authorities. However, predicting sales depends on different parameters. We introduced two different Bayesian regression models along with linear regression model to predict sales with minimum error. One model has single layer without considering differentiation in effect of academic years and two layer model use year differentiation in estimations. The results of the analysis also demonstrated that two layer Bayesian method can be applied to model this change and the parameters of the posterior distributions can be calculated with Markov Chain Monte Carlo methods. We start applying linear regression model and observed that it constantly overestimates on test data. Then, we started building Bayesian models. After the analysis of three models, we divided the data set into consecutive four weeks blocks and calculated the performance measures in each model. We can conclude that linear regression and single layer bayesian model have similar scores and lower error related scores compared to two layer model. This study showed that modeling the estimation problem considering changing effects through years does not contribute to performance measures and relatively simple models perform better.

Appendix A

Derivations of Full Conditional Distributions

A.1 Variances and Covariance Matrices for Single Layer Model

Posterior for $\bar{\Sigma}$

$$\begin{aligned}
 f(\bar{\Sigma}, .) &\propto f(\bar{\Sigma}|\nu, V)f(\beta|\bar{\beta}, \bar{\Sigma}) \\
 &\propto |\bar{\Sigma}|^{-(\frac{\nu+p+1}{2})}etr\left(-\frac{1}{2}V\bar{\Sigma}^{-1}\right) - |\bar{\Sigma}|^{-\frac{1}{2}}\underbrace{exp\left(-\frac{1}{2}(\beta - \bar{\beta})^T\bar{\Sigma}^{-1}(\beta - \bar{\beta})\right)}_{=tr((\beta - \bar{\beta})^T\bar{\Sigma}^{-1}(\beta - \bar{\beta}))} \\
 &\propto |\bar{\Sigma}|^{-(\frac{\nu+1+p+1}{2})}etr\left(-\frac{1}{2}V\bar{\Sigma}^{-1}\right) - |\bar{\Sigma}|^{-(\frac{1}{2})}exp\left(-\frac{1}{2}tr\left((\beta - \bar{\beta})(\beta - \bar{\beta})^T\bar{\Sigma}^{-1}\right)\right) \\
 &\propto |\bar{\Sigma}|^{-(\frac{\nu+1+p+1}{2})}etr\left(-\frac{1}{2}\left(V + \frac{(\beta - \bar{\beta})(\beta - \bar{\beta})^T}{\sigma^2}\right)\bar{\Sigma}^{-1}\right)
 \end{aligned} \tag{A.1}$$

Then, $f(\bar{\Sigma}|.) \sim IW(\nu + 1, V + (\beta - \bar{\beta})(\beta - \bar{\beta})^T)$

Posterior for σ^2

$$\begin{aligned}
f(\sigma^2, \cdot) &\propto f(\sigma^2) f(Y|X, \beta, \sigma^2) \\
&\propto \left(\frac{1}{2\pi|\sigma^2 I_n|^{\frac{1}{2}}} \right) \exp\left(-\frac{1}{2\sigma^2} (Z - X\beta)^T (Z - X\beta) \right) \\
&\propto \left(\frac{1}{\sigma^2} \right)^{\frac{n}{2}} \left(\frac{1}{\sigma^2} \right)^{\frac{\alpha}{2}+1} \exp\left(-\frac{1}{\sigma^2} \left[\frac{\|Z - X\beta\|^2 + \lambda}{2} \right] \right) \quad (\text{A.2}) \\
&\propto \left(\frac{1}{\sigma^2} \right)^{\frac{n+\alpha}{2}+1} \exp\left(-\frac{1}{\sigma^2} \left[\frac{\|Z - X\beta\|^2 + \lambda}{2} \right] \right)
\end{aligned}$$

Then $f(\sigma^2|\cdot) \sim IG\left(\frac{n+\alpha}{2}, \frac{\|Z-X\beta\|^2+\lambda}{2}\right)$

A.2 Covariates' Effects for Single Layer Model

Posterior for β

$$\begin{aligned}
f(\beta, \cdot) &\propto f(\beta|\bar{\beta}, \bar{\Sigma}) \times f(\mathbb{Z}|X\beta, \sigma^2) \\
&\propto \exp\left(-\frac{1}{2}(\beta - \bar{\beta})^T \bar{\Sigma}^{-1}(\beta - \bar{\beta}) - \frac{1}{2\sigma^2} (\mathbb{Z} - X\beta)^T (\mathbb{Z} - X\beta) \right) \quad (\text{A.3})
\end{aligned}$$

By using Cholesky decomposition for precision matrix $\bar{\Sigma}^{-1} = U^T U$, we can reorganize the matrices as follows:

$$\begin{aligned}
\left(\begin{bmatrix} \mathbb{Z} - X\beta \\ U\bar{\beta} - U\beta \end{bmatrix}^T \begin{bmatrix} \mathbb{Z} - X\beta \\ U\bar{\beta} - U\beta \end{bmatrix} = \begin{bmatrix} \mathbb{Z} \\ U\bar{\beta} \end{bmatrix} - \begin{bmatrix} X \\ U \end{bmatrix} \beta \right)^T \left(\begin{bmatrix} \mathbb{Z} \\ U\bar{\beta} \end{bmatrix} - \begin{bmatrix} X \\ U \end{bmatrix} \beta \right) \quad (\text{A.4}) \\
\Rightarrow (\tilde{\mathbb{Z}} - \tilde{X}\beta)^T (\tilde{\mathbb{Z}} - \tilde{X}\beta)
\end{aligned}$$

$$\begin{aligned} &\propto \exp\left(-\frac{1}{2\sigma^2}(\tilde{\mathbb{Z}} - \tilde{X}\beta)^T(\tilde{\mathbb{Z}} - \tilde{X}\beta)\right) = \exp\left(-\frac{1}{2\sigma^2}\|\tilde{\mathbb{Z}} - \tilde{X}\beta\|^2\right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}\left([\beta - \tilde{\beta}_{MLE}]^T \tilde{X}^T \tilde{X} [\beta - \tilde{\beta}_{MLE}] + \|\tilde{\mathbb{Z}} - \tilde{X}\tilde{\beta}_{MLE}\|^2\right)\right) \end{aligned}$$

where $\tilde{\beta}_{MLE} = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{\mathbb{Z}}$

$$\begin{aligned} &= \left(\begin{bmatrix} X^T & U^T \end{bmatrix} \begin{bmatrix} X \\ U \end{bmatrix} \right)^{-1} \begin{bmatrix} X^T & U^T \end{bmatrix} \begin{bmatrix} \mathbb{Z} \\ U\bar{\beta} \end{bmatrix} \\ &= \left(X^T X + \underbrace{U^T U}_{\bar{\Sigma}^{-1}} \right)^{-1} (X^T \mathbb{Z} + U^T U \bar{\beta}) \\ &= \left(X^T X + \bar{\Sigma}^{-1} \right)^{-1} (X^T X \underbrace{(X^T X)^{-1} X^T \mathbb{Z}}_{\beta_{MLE}} + \bar{\Sigma}^{-1} \bar{\beta}) \\ &= \left(X^T X + \bar{\Sigma}^{-1} \right)^{-1} (X^T X \beta_{MLE} + \bar{\Sigma}^{-1} \bar{\beta}) \end{aligned} \tag{A.5}$$

Then, the mean and covariance matrix can be written as follows

mean;

$$\beta_{MLE} \equiv \left(X^T X + \bar{\Sigma}^{-1} \right)^{-1} (X^T X \beta_{MLE} + \bar{\Sigma}^{-1} \bar{\beta}) \tag{A.6}$$

covariance matrix :

$$\left(\frac{\tilde{X}^T \tilde{X}}{\sigma^2} \right)^{-1} \equiv \left(\bar{\Sigma}^{-1} \sigma^2 + X^T X \right)^{-1}$$

A.3 Mean Covariates' Effects for Single Layer Model

Posterior for $\bar{\beta}$

$$\begin{aligned}
f(\bar{\beta}, \cdot) &\propto f(\beta|\bar{\beta}, \bar{\Sigma}) \times f(\bar{\beta}|\bar{\beta}, \bar{\Sigma}) \\
&\propto \exp\left(-\frac{1}{2}(\beta - \bar{\beta})^T \underbrace{\bar{\Sigma}^{-1}}_{H^T H} (\beta - \bar{\beta}) - \frac{1}{2}(\bar{\beta} - \bar{\bar{\beta}})^T \underbrace{\bar{\Sigma}^{-1}}_{G^T G} (\bar{\beta} - \bar{\bar{\beta}})\right) \\
&\propto \exp\left(-\frac{1}{2\sigma^2}\left([H\beta - H\bar{\beta}]^T [H\beta - H\bar{\beta}] + [G\bar{\beta} - G\bar{\bar{\beta}}]^T [G\bar{\beta} - G\bar{\bar{\beta}}]\right)\right) \\
&\Rightarrow \begin{bmatrix} H\beta - H\bar{\beta} \\ G\bar{\beta} - G\bar{\bar{\beta}} \end{bmatrix}^T \begin{bmatrix} H\beta - H\bar{\beta} \\ G\bar{\beta} - G\bar{\bar{\beta}} \end{bmatrix} \\
&\Rightarrow \left(\begin{bmatrix} H\beta \\ G\bar{\beta} \end{bmatrix} - \begin{bmatrix} H \\ G \end{bmatrix} \bar{\beta}\right)^T \left(\begin{bmatrix} H\beta \\ G\bar{\beta} \end{bmatrix} - \begin{bmatrix} H \\ G \end{bmatrix} \bar{\beta}\right) \\
&\Rightarrow (\tilde{G} - \tilde{H}\bar{\beta})^T (\tilde{G} - \tilde{H}\bar{\beta}) \\
&\propto \exp\left(-\frac{1}{2\sigma^2}(\tilde{G} - \tilde{H}\bar{\beta})^T (\tilde{G} - \tilde{H}\bar{\beta})\right) = \exp\left(-\frac{1}{2\sigma^2} \|\tilde{G} - \tilde{H}\bar{\beta}\|^2\right) \\
&\propto \exp\left(-\frac{1}{2\sigma^2}\left([\tilde{H}\bar{\beta} - \tilde{H}\beta'_{MLE}]^T [\tilde{H}\bar{\beta} - \tilde{H}\beta'_{MLE}] + [\tilde{G} - \tilde{H}\beta'_{MLE}]^T [\tilde{G} - \tilde{H}\beta'_{MLE}]\right)\right) \\
&\propto \exp\left(-\frac{1}{2\sigma^2}\left([\bar{\beta} - \beta'_{MLE}]^T \tilde{H}^T \tilde{H} [\bar{\beta} - \beta'_{MLE}]\right)\right)
\end{aligned} \tag{A.7}$$

Then, the mean and covariance matrix can be written as follows

$$\begin{aligned}
\text{mean; } \beta'_{MLE} &= \left(\tilde{H}\tilde{H}\right)^{-1} \tilde{H}^T \tilde{G} \\
&= \left(\begin{bmatrix} H^T & G^T \end{bmatrix} \begin{bmatrix} H \\ G \end{bmatrix}\right)^{-1} \begin{bmatrix} H^T & G^T \end{bmatrix} \begin{bmatrix} H\beta \\ G\bar{\beta} \end{bmatrix} \\
&= \left(\underbrace{H^T H}_{\bar{\Sigma}^{-1}} + \underbrace{G^T G}_{\bar{\bar{\Sigma}}^{-1}}\right)^{-1} (H^T H\beta + G^T G\bar{\beta}) \\
&= (\bar{\Sigma}^{-1} + \bar{\bar{\Sigma}}^{-1})^{-1} (\bar{\Sigma}^{-1}\beta + \bar{\bar{\Sigma}}^{-1}\bar{\beta})
\end{aligned} \tag{A.8}$$

$$\text{covariance matrix : } \left(\begin{bmatrix} H^T & G^T \end{bmatrix} \begin{bmatrix} H \\ G \end{bmatrix}\right)^{-1} = [H^T H + G^T G] = \bar{\Sigma}^{-1} + \bar{\bar{\Sigma}}^{-1} \tag{A.9}$$

Bibliography

- [1] J. Jarrett, *Business forecasting methods*. B. Blackwell, 1991.
- [2] V. H. JE, D. Wichern, and A. Reitsch, “Business forecasting,” *Int J Forecast*, vol. 22, no. 4, pp. 823–824, 2005.
- [3] J. K. Shim, *Strategic business forecasting: the complete guide to forecasting real world company performance*. CRC Press, 2000.
- [4] K. Ryu and A. Sanchez, “The evaluation of forecasting methods at an institutional foodservice dining facility,” *The journal of hospitality financial management*, vol. 11, no. 1, pp. 27–45, 2003.
- [5] L. Blecher and R. J. Yeh, “Forecasting meal participation in university residential dining facilities,” *Journal of Foodservice Business Research*, vol. 11, no. 4, pp. 352–362, 2008.
- [6] H. Theil, “Principles of econometrics,” tech. rep., 1971.
- [7] R. S. Pindyck and D. L. Rubinfeld, *Econometric models and business forecasts*. McGraw-Hill, 1976.
- [8] G. S. Maddala and K. Lahiri, *Introduction to econometrics*, vol. 2. Macmillan New York, 1992.
- [9] C. Malefors, I. Strid, P.-A. Hansson, and M. Eriksson, “Potential for using guest attendance forecasting in swedish public catering to reduce overcatering,” *Sustainable Production and Consumption*, vol. 25, pp. 162–172, 2021.

- [10] C. Malefors, L. Secondi, S. Marchetti, and M. Eriksson, “Food waste reduction and economic savings in times of crisis: The potential of machine learning methods to plan guest attendance in swedish public catering during the covid-19 pandemic,” *Socio-Economic Planning Sciences*, p. 101041, 2021.
- [11] P. Rossi, G. Allenby, and R. McCulloch, *Bayesian Statistics and Marketing*. Wiley Series in Probability and Statistics, Wiley, 2012.
- [12] S. Radchenko *et al.*, “The long-run forecasting of energy prices using the model of shifting trend,” *University of North Carolina at Charlotte, Working Paper*, 2005.
- [13] M. Geda and C. Kwong, “An mcmc based bayesian inference approach to parameter estimation of distributed lag models for forecasting used product returns for remanufacturing,” *Journal of Remanufacturing*, pp. 1–20, 2021.
- [14] “Stan reference 2.28.” <https://mc-stan.org/docs/2.28/reference-manual/convergence.html>. Accessed: 2021-09-30.
- [15] “Stan reference 2.28.” <https://mc-stan.org/docs/2.28/reference-manual/notation-for-samples-chains-and-draws.html>. Accessed: 2021-09-30.
- [16] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.
- [17] J. Fox and S. Weisberg, *An R Companion to Applied Regression*. Thousand Oaks CA: Sage, third ed., 2019.
- [18] A. E. Gelfand, S. E. Hills, A. Racine-Poon, and A. F. Smith, “Illustration of bayesian inference in normal data models using gibbs sampling,” *Journal of the American Statistical Association*, vol. 85, no. 412, pp. 972–985, 1990.