Waiting Time in the Emergency Room
A Case Study

Authors:

Chris Dong,
Department of Statistics,
University of California Los Angeles

Ye Zhi,
Department of Statistics
University of California Los Angeles

Zihao Zhou,
Department of Mathematics,
Department of Statistics,
University of California Los Angeles

Advisor:

Prof Vivian Lew
Department of Statistics,
University of California Los Angeles

This report is for the final project of Course Statistics 130.
The document is last modified on 16$^{\text{th}}$ December, 2014.
Any question about this project may be directed to Zihao Zhou ⟨zihao2011@g.ucla.edu⟩.

Abstract

We have tried to look for factors that significantly affect waiting time in a given emergency room. We find that net of the effects of the patient's age, the type of ailment, the acuity of the ailment, and the doctor in charge, an additional visitor to the emergency room is associated with about 2 minutes longer waiting time. Our analysis also finds that the efficiency of the doctor's work in the emergency room appears to have an important impact on the waiting time.

1 Introduction

The waiting time has long been a crucial factor in determining whether a hospital is good or not. This is particularly true for an emergency room, where every second counts. In fact, this issue has wide political resonance. In the UK, the waiting time at the National Health Service has been constantly raised and debated by the incumbent coalition government and the opposition party.[1]

Our study has to do with an emergency room of an unnamed hospital. According to the instructions of this project, there are two questions we would like to answer in this project. Firstly, we interested in knowing how the number of visitors in an emergency room affect a visitor's waiting time, net of the effects of other factors such as the visitor's age, type of ailment, the doctor in charge and acuity of the visitor's situation. Secondly, we are interested in knowing whether there is an important factor that affects a patient's waiting time.

In Section 2, we will give a detailed description of the two datasets we have and how we have processed them for the purpose of this study. In Section 3, we run analysis in direct response to the two questions in this project. Then we give the conclusion in Section 4.

2 Data

2.1 Emergency Room Data

We have two datasets at hand. The first one is the event report for each visit in a given time period in Year 2013 of an emergency room in an unnamed hospital. This dataset includes information such as identification number unique to each visit, the patient's date of birth and the arrival time and date of the patient. It also records the nature of each event in each visit. For example, in the field of event_name, "arrive" refers to the event that the visit has arrived and "MSEI" refers to the event that a doctor comes and attends a patient.

The core variable in this study is the waiting time of a visitor. However, in the original dataset, there is no such variable, and therefore it becomes our responsibility to define it. For this study, we define the waiting time to be the length of the period from when the patient arrives at the hospital to when the patient is attended by a doctor.

The second dataset is a emergency room census recording a routine procedure of the emergency room in which the number of visitors in the emergency room is recorded at the beginning of each hour. The key variable useful to our study in this dataset is the count variable, the number of visitors at the record time.

2.2 Data Processing

We have processed the data using **R** and the source codes are provided here. If the reader is interested in running the source codes in this document, please make sure that all required data files are properly named and their respective paths correctly specified. Also, before

[1] Here is an example this year reported by the Guardian: `http://www.theguardian.com/society/2014/jun/12/nhs-waiting-list-over-3-million`

running the codes, the reader should have a look at the library declaration to make sure that all necessary packages are installed and updated.

We first give the settings of **R** here.

```
# library declarations
library("knitr")
library("plyr")
library("stringr")
library("lubridate")
library("leaps")
library("lmtest")
library("ggplot2")
# ggplot2 theme
theme_set(theme_bw())
# digit display
options(digits = 3)
# chunk settings for knitr
opts_chunk$set(echo = TRUE, results = 'markup', fig.height = 4.5, comment = NA,
               fig.align = 'center', size = 'small')
```

Here we shows how we processed the data. For readability, we are not letting **R** run the codes in this chunk. The reader can set eval = TRUE for this chunk in the Rnw file if interested.

```
# import data
# make sure that both data sets are properly named
event_report <- read.csv("2013 event report.csv", na.strings = "",
                         sep=";", dec=",", stringsAsFactors=FALSE,)
svh_census <- read.csv("SVHCENSUS.csv", stringsAsFactors=FALSE)
# convert column names to lower caess for easier programming
names(event_report) <- tolower(names(event_report))
names(svh_census) <- tolower(names(svh_census))

# FIN_nbr is unique to visits.
# this is the id for each patient's visit
# we define the waiting time to be from
# when the patient arrives at the hospital to when he is attended by a doctor
# arr_time is included in every event record
# so the events we are interested in is MSEI
# also there are some orders cancelled
# so we would like to final_event_status to be complete
event_interest_index <- which(event_report$event_name == "MSEI" &
                                event_report$final_event_status == "Complete")
event_interest <- event_report[event_interest_index, ]
# we notice that there are multiple request for doctors for a same visit
# we assume that the patient get attended when the doctor responds to the first
# request
# therefore we get any duplicates after the first encounter
```

```r
fin_duplicates_index <- duplicated(event_interest$fin_nbr)
event_interest <- event_interest[!fin_duplicates_index, ]
# now there are no duplicates
anyDuplicated(event_interest$fin_nbr)
# define the arrival date time
arrive_date_time <- str_c(event_interest$arr_time, event_interest$arr_date,
                          sep = " ")
arrive_date_time <- as.POSIXct(strptime(arrive_date_time, "%R %m/%d/%y"))
# Check NA
anyNA(arrive_date_time)
# also define start date time
start_date_time <- str_c(event_interest$start_time, event_interest$start_date,
                         sep = " ")
start_date_time <- as.POSIXct(strptime(start_date_time, "%R %m/%d/%y"))
# check NA
anyNA(arrive_date_time)
# add arrival date time to dataset
event_interest$arr_date_time <- arrive_date_time
# add start_date_time to dataset
event_interest$start_date_time <- start_date_time
# compute waiting time
wait_time <- as.numeric(difftime(event_interest$start_date_time,
                                 event_interest$arr_date_time, units = "mins"))
# check NA
anyNA(wait_time)
# add wait time to dataset
event_interest$wait_time <- wait_time
# date_hour_event
date_hour_event <- format(event_interest$arr_date_time, format = "%H %m/%d/%y")
event_interest$date_hour <- date_hour_event
# date_hour_census
# check NA in census date hour
anyNA(svh_census$datetime)
# there is duplicate in svh_census
cencus_dup_index <- which(duplicated(svh_census$datetime, fromLast = T))
# take a look
svh_census[svh_census$datetime == svh_census$datetime[cencus_dup_index], ]
# we should preserve the latest update
svh_census <- svh_census[-cencus_dup_index, ]
date_time_census <- as.POSIXct(strptime(svh_census$datetime,
                                        format = "%Y-%m-%d %R:%S"))
# date_hour_census
date_hour_census <- format(date_time_census, format = "%H %m/%d/%y")
svh_census$date_hour <- date_hour_census

# merge two datasets
project_data <- join(event_interest, svh_census, "date_hour", "left")
```

```r
# check NA in count
anyNA(project_data$count)
# take a look
names(project_data)
# determine variables to delete
variables_delete <- c("arr_date", "arr_time", "start_time", "start_date",
                      "date_hour", "datetime", "dow", "hour", "day")
variables_delete_index <- which(names(project_data) %in% variables_delete)
project_data <- project_data[, -variables_delete_index]
# get date of birth
dob <- as.Date(project_data$dob, format = "%m/%d/%y")
yob <- as.numeric(format(dob, format = "%Y"))
# some years before 1968 are incorrectly imported as in 21st century
# correct those years
# first look for the latest year in the dataset
current_year <- max(year(project_data$start_date_time))
wrong_year_index <- which(yob > current_year)
# decrement incorrect years by 100 years
yob[wrong_year_index] <- yob[wrong_year_index] - 100
# age
age <- year(project_data$arr_date_time) - yob
# plug year to data
project_data$age <- age

# some pecularities of the dataset
# some waiting times were suspiciously long
head(sort(project_data$wait_time, decreasing = T), n = 100L)
# we take a look at those record in which the waiting time is more than 10 hours
project_data[which(project_data$wait_time >= 600),
             c("arr_date_time", "start_date_time", "wait_time")]
# we cannot be sure about the exceptionally long waiting time was due to
#   recording error or something else, so we are to keep those records
# but we are almost sure that the observation with waiting time more than 2 months
#  is due to recording error, so we delete this observation
record_del_index <- which(project_data$wait_time >= 60 * 24 * 60)
project_data <- project_data[-record_del_index, ]
# some waiting time are negative
negative_time_index <- which(project_data$wait_time < 0)
project_data$wait_time[negative_time_index]
# we believe that these were data errors and we reverse the sign
project_data$wait_time[negative_time_index] <-
  -project_data$wait_time[negative_time_index]

# add a indicator variable for acuity 2-Very Urgent for Part b
project_data$very_urgent <- NA
project_data$very_urgent[complete.cases(project_data$acuity)] <- 0
very_urgent_index <- which(project_data$acuity == "2-Very Urgent")
```

```
project_data$very_urgent[very_urgent_index] <- 1
table(project_data$very_urgent, useNA = 'ifany')

# save the data
saveRDS(object = project_data, file = "project_data.RDS")
```

We computed a new variable named wait_time for the emergency event report. Since we have defined the waiting time to be associated with arrival time and the time when a doctor comes, we have eliminated rows featuring event irrelevant for our study. Then we add the number of visitors when the patient makes the visit to the event record from the census record matched by the hour and date of an event. We have also computed each patient's age based on their date of birth and the date they made the visit.

Pecularities We have noticed something unusual in the dataset. Disucssions of these unusual findings are also offered in comments in the above chunk of codes, but we are going to summarise these findings here.

Firstly, sometimes when a patient was attended by a doctor, the doctor's name (md) was not recorded. The same problem happened to the seriousness of a patient's situation (acuity) and to the patient's chief complaint (chief_complaint). We dealt with this issue by treating these cases as missing values.

For the wait_time variable we have created, we notice that there are two observations with exceptionally long waiting time, each more than 2 months. There are also several cases where the waiting time is more than 10 hours, which should be considered as unusual for an emergency room. Because we are unsure about the reason for these long waiting times, we try to aliminate as few cases as possible from our data. Therefore we still keep those observations with unusually long waiting times but we do delete the two observations mentioned above, since we are almost sure that some recording error happened.

Also, we notice that some cases have a negative waiting time. Since we suspect that the recorder got the order wrong, we deal with these cases by converting these negative waiting time to positive.

3 Analysis

To answer the first question of this project, we regress the waiting time on the patient's age, the type of ailment, the acuity, the doctor and the count. Due to large number of variables created in this model, we are not showing the entire result of this model.

```
# load dataset
project_data <- readRDS("project_data.RDS")
# run the regression of wait time on age, type of ailment, doctor and acuity
reg_a <- lm(wait_time ~ count + age + md + chief_complaint + acuity,
            data = project_data)
# we do not show the summary output of the model here because there are too
#   many indicator variables
```

We show the estimation for variable count here.

```
reg_a_coef_matrix <- summary(reg_a)$coefficients
# look at the estimate for count
count_a_report <- reg_a_coef_matrix["count", ]
count_a_report

  Estimate Std. Error    t value   Pr(>|t|)
  1.85e+00   6.70e-02   2.76e+01  6.38e-165
```

So the model estimates that with the patient's age, md, chief_complaint and acuity are accounted for, an additional visitor in the mergency room lengthens the waiting time by about 1.849 minutes. Note that the R^2 for this model is 0.071, which is extremely low.

To have an idea whether the magnitude of this coefficient is big, we look at the five-number summary statistics of waiting time in our data.

```
# have a sense of the distribution of the waiting time
fivenum(project_data$wait_time)

[1]    0    6   12   23 6825
```

So based on the median waiting time, lengthening the waiting time by about 1.849 minutes is quite significant.

Now we go to the second question of this project. Note that the previous model is extremely unhelpful because that model has over 1000 predictors. We have decided to eliminate the chief_complaint variable from our model. The justificaition of this elimination comes from our finding that a really small proportion of indicator variables for the chief_complaint is statistically significant in the previous model.

```
# we should get rid of the indicator variables for complaints
# the reason is that only a very small proportion of these indicator variables
# are statistically significant
complaint_index <- str_detect(rownames(reg_a_coef_matrix),
                              pattern = "chief_complaint")
chief_complaint_p <- reg_a_coef_matrix[complaint_index, 4]
chief_com_significant <- which(chief_complaint_p <= 0.2)
complaint_sig_prop <- length(chief_com_significant) / length(complaint_index)
complaint_sig_prop

[1] 0.00427
```

Since the proportion of statistically significant predictors in the chief_complaint is only 0.427%, we get rid of this variable.

We then search for a good model with the help from the *regsubsets* function from package *leaps*.

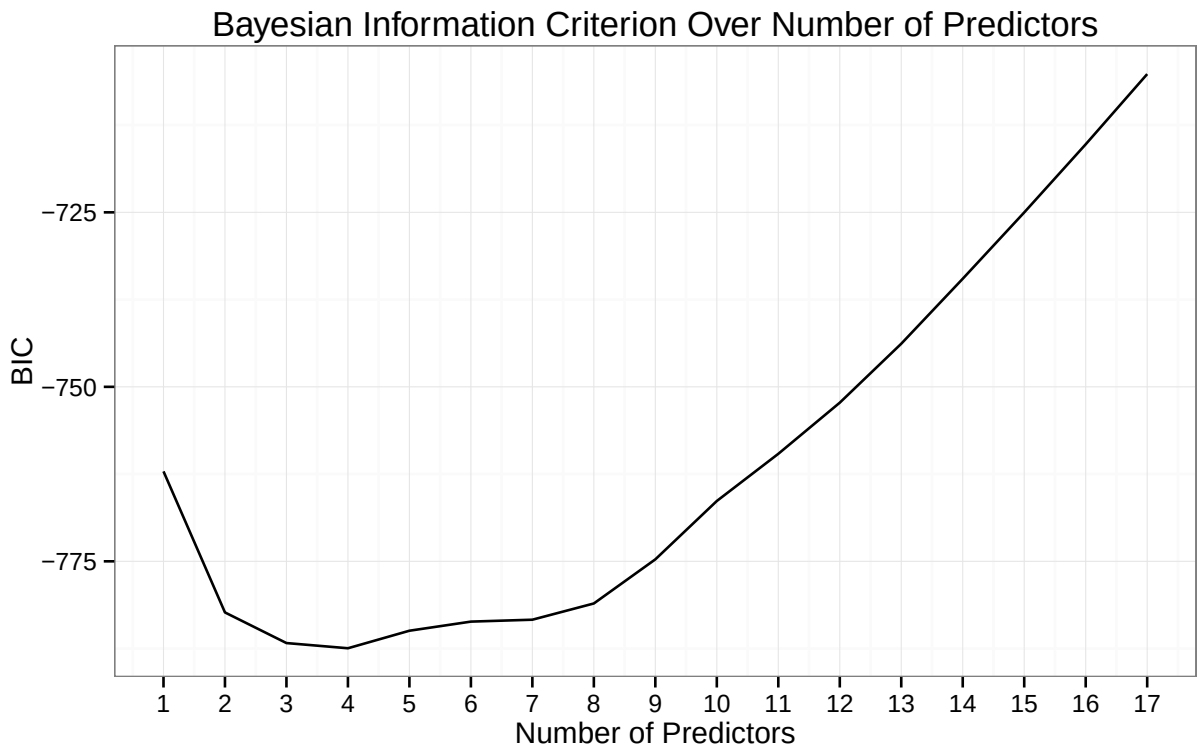

Figure 1: BIC of First Model Search

```
# get rid of the complaint indicator variables
# find a best model with regsubsets from the leaps package
model_find <- regsubsets(wait_time ~ count + age + md + acuity,
                         data = project_data, nvmax = 30,
                         method = "exhaustive")
```

We determine the optimal number of predictors by looking for the minimiser of Bayesian Information Criterion (Schwarz et al. 1978), short for BIC, as in Figure 1 on Page 7.

The BIC recommends a four-variable model and we look at the recommendation from the previous search function for four-variable model.

```
# model says that we should choose four variables
# take a look at the model search
summary(model_find)

Subset selection object
Call: regsubsets.formula(wait_time ~ count + age + md + acuity, data = project_data,
    nvmax = 30, method = "exhaustive")
17 Variables  (and intercept)
                      Forced in Forced out
count                     FALSE      FALSE
age                       FALSE      FALSE
```

```
mdDanescu, MD, Adrian      FALSE      FALSE
mdFriend, MD, David I      FALSE      FALSE
mdGharakhani, DO, Hoss     FALSE      FALSE
mdKuban, MD, Alan L        FALSE      FALSE
mdMatts , MD, Christin     FALSE      FALSE
mdNorman, MD, Joyce        FALSE      FALSE
mdPollack, MD, Barry H     FALSE      FALSE
mdSun, Nancy N             FALSE      FALSE
mdTilles, MD, Ira H        FALSE      FALSE
mdVan Zyl, MD, Carin       FALSE      FALSE
mdYu, MD, Alfred           FALSE      FALSE
acuity2-Very Urgent        FALSE      FALSE
acuity3-Urgent             FALSE      FALSE
acuity4-Less Urgent        FALSE      FALSE
acuity5-Non Urgent         FALSE      FALSE
1 subsets of each size up to 17
Selection Algorithm: exhaustive
          count age mdDanescu, MD, Adrian mdFriend, MD, David I
1  ( 1 )  "*"   " " " "                   " "
2  ( 1 )  "*"   " " " "                   " "
3  ( 1 )  "*"   "*" " "                   " "
4  ( 1 )  "*"   "*" " "                   " "
5  ( 1 )  "*"   "*" " "                   " "
6  ( 1 )  "*"   "*" " "                   " "
7  ( 1 )  "*"   "*" " "                   " "
8  ( 1 )  "*"   "*" " "                   " "
9  ( 1 )  "*"   "*" "*"                   " "
10  ( 1 ) "*"   "*" "*"                   "*"
11  ( 1 ) "*"   "*" "*"                   " "
12  ( 1 ) "*"   "*" "*"                   " "
13  ( 1 ) "*"   "*" "*"                   "*"
14  ( 1 ) "*"   "*" "*"                   "*"
15  ( 1 ) "*"   "*" "*"                   "*"
16  ( 1 ) "*"   "*" "*"                   "*"
17  ( 1 ) "*"   "*" "*"                   "*"
          mdGharakhani, DO, Hoss mdKuban, MD, Alan L
1  ( 1 )  " "                    " "
2  ( 1 )  " "                    " "
3  ( 1 )  " "                    " "
4  ( 1 )  " "                    " "
5  ( 1 )  " "                    " "
6  ( 1 )  " "                    "*"
7  ( 1 )  " "                    "*"
8  ( 1 )  " "                    "*"
9  ( 1 )  " "                    "*"
10  ( 1 ) " "                    "*"
11  ( 1 ) " "                    "*"
```

```
12  ( 1 ) " "                    "*"
13  ( 1 ) " "                    "*"
14  ( 1 ) " "                    "*"
15  ( 1 ) " "                    "*"
16  ( 1 ) "*"                    "*"
17  ( 1 ) "*"                    "*"
          mdMatts , MD, Christin mdNorman, MD, Joyce
1   ( 1 )  " "                   " "
2   ( 1 )  " "                   " "
3   ( 1 )  " "                   " "
4   ( 1 )  " "                   "*"
5   ( 1 )  " "                   "*"
6   ( 1 )  " "                   "*"
7   ( 1 )  " "                   "*"
8   ( 1 )  " "                   "*"
9   ( 1 )  " "                   "*"
10  ( 1 ) " "                    "*"
11  ( 1 ) " "                    "*"
12  ( 1 ) " "                    "*"
13  ( 1 ) " "                    "*"
14  ( 1 ) "*"                    "*"
15  ( 1 ) "*"                    "*"
16  ( 1 ) "*"                    "*"
17  ( 1 ) "*"                    "*"
          mdPollack, MD, Barry H mdSun, Nancy N mdTilles, MD, Ira H
1   ( 1 )  " "                   " "            " "
2   ( 1 )  " "                   " "            " "
3   ( 1 )  " "                   " "            " "
4   ( 1 )  " "                   " "            " "
5   ( 1 )  " "                   " "            "*"
6   ( 1 )  " "                   " "            "*"
7   ( 1 )  " "                   " "            "*"
8   ( 1 )  "*"                   " "            "*"
9   ( 1 )  "*"                   " "            "*"
10  ( 1 ) "*"                    " "            "*"
11  ( 1 ) "*"                    " "            "*"
12  ( 1 ) "*"                    " "            "*"
13  ( 1 ) "*"                    " "            "*"
14  ( 1 ) "*"                    " "            "*"
15  ( 1 ) "*"                    "*"            "*"
16  ( 1 ) "*"                    "*"            "*"
17  ( 1 ) "*"                    "*"            "*"
          mdVan Zyl, MD, Carin mdYu, MD, Alfred acuity2-Very Urgent
1   ( 1 )  " "                  " "             " "
2   ( 1 )  " "                  " "             "*"
3   ( 1 )  " "                  " "             "*"
4   ( 1 )  " "                  " "             "*"
```

```
5  ( 1 )  " "                  " "              "*"
6  ( 1 )  " "                  " "              "*"
7  ( 1 )  " "                  "*"              "*"
8  ( 1 )  " "                  "*"              "*"
9  ( 1 )  " "                  "*"              "*"
10  ( 1 ) " "                  "*"              "*"
11  ( 1 ) " "                  "*"              " "
12  ( 1 ) " "                  "*"              "*"
13  ( 1 ) " "                  "*"              "*"
14  ( 1 ) " "                  "*"              "*"
15  ( 1 ) " "                  "*"              "*"
16  ( 1 ) " "                  "*"              "*"
17  ( 1 ) "*"                  "*"              "*"
          acuity3-Urgent acuity4-Less Urgent acuity5-Non Urgent
1  ( 1 )  " "             " "                 " "
2  ( 1 )  " "             " "                 " "
3  ( 1 )  " "             " "                 " "
4  ( 1 )  " "             " "                 " "
5  ( 1 )  " "             " "                 " "
6  ( 1 )  " "             " "                 " "
7  ( 1 )  " "             " "                 " "
8  ( 1 )  " "             " "                 " "
9  ( 1 )  " "             " "                 " "
10  ( 1 ) " "             " "                 " "
11  ( 1 ) "*"             "*"                 "*"
12  ( 1 ) "*"             "*"                 "*"
13  ( 1 ) "*"             "*"                 "*"
14  ( 1 ) "*"             "*"                 "*"
15  ( 1 ) "*"             "*"                 "*"
16  ( 1 ) "*"             "*"                 "*"
17  ( 1 ) "*"             "*"                 "*"
```

However, we see that the serach function recommends we use indicator variables for two doctors in our model. As their performance may be statistically significantly different from their peers, for the purpose of study, including a small number of doctors in our model is not helpful, therefore we exclude md in our search function and redo the search.

```r
# the model suggests we include an indicator variable for a single doctor
# this is not useful for our study for the emergency room as a whole
# so we get rid of the doctor variable as well
model_find_2 <- regsubsets(wait_time ~ count + age + acuity,
                           data = project_data, nvmax = 30,
                           method = "exhaustive")
```

Similarly, we use the BIC as in Figure 2 on Page 11. The plot shows that we should use a three-variable model.

We take a look at the results of our second model search.

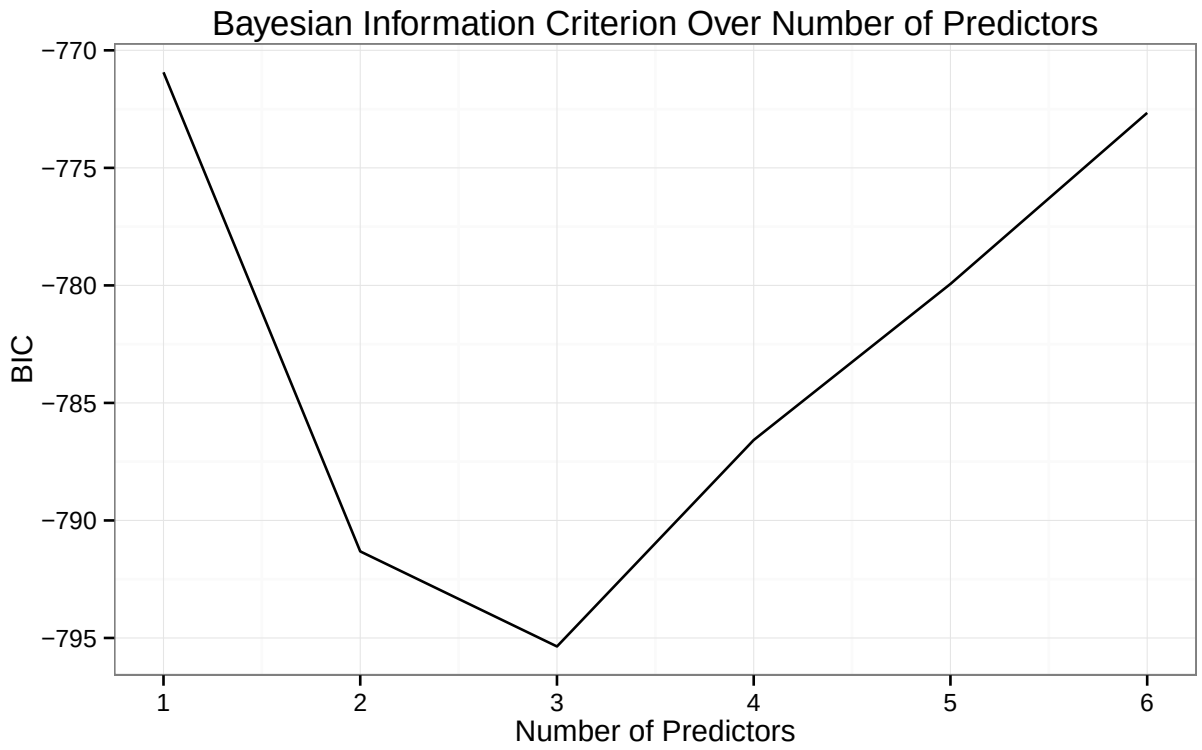

Figure 2: BIC for Second Model Search

```
# take a look at the search algorithm
summary(model_find_2)

Subset selection object
Call: regsubsets.formula(wait_time ~ count + age + acuity, data = project_data,
    nvmax = 30, method = "exhaustive")
6 Variables  (and intercept)
                     Forced in Forced out
count                    FALSE      FALSE
age                      FALSE      FALSE
acuity2-Very Urgent      FALSE      FALSE
acuity3-Urgent           FALSE      FALSE
acuity4-Less Urgent      FALSE      FALSE
acuity5-Non Urgent       FALSE      FALSE
1 subsets of each size up to 6
Selection Algorithm: exhaustive
         count age acuity2-Very Urgent acuity3-Urgent acuity4-Less Urgent
1  ( 1 ) "*"   " " " "                 " "             " "
2  ( 1 ) "*"   " " "*"                 " "             " "
3  ( 1 ) "*"   "*" "*"                 " "             " "
4  ( 1 ) "*"   "*" "*"                 " "             " "
5  ( 1 ) "*"   "*" " "                 "*"             "*"
6  ( 1 ) "*"   "*" "*"                 "*"             "*"
         acuity5-Non Urgent
1  ( 1 ) " "
2  ( 1 ) " "
3  ( 1 ) " "
4  ( 1 ) "*"
5  ( 1 ) "*"
6  ( 1 ) "*"

# software recommends use count, age and acuity level very urgent
# to run this model, create an indicator variable for the very urgent
```

So the search function recommends that we should use count, age and an indicator variable for "Very Urgent" level of acuity in our model. We create this variable called very_urgent as shown in our data processin step. Now we run the model here.

```
reg_b_search <- lm(wait_time ~ count + age + very_urgent, data = project_data)
summary(reg_b_search)


Call:
lm(formula = wait_time ~ count + age + very_urgent, data = project_data)

Residuals:
   Min     1Q Median     3Q    Max
```

```
 -43.5  -10.5   -4.1    4.1 1732.6

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.8141     0.7458    7.80  6.7e-15 ***
count         1.8085     0.0641   28.21  < 2e-16 ***
age          -0.0400     0.0106   -3.76  0.00017 ***
very_urgent  -3.5585     0.7670   -4.64  3.5e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 41.2 on 24658 degrees of freedom
  (146 observations deleted due to missingness)
Multiple R-squared:  0.0333,Adjusted R-squared:  0.0332
F-statistic:  283 on 3 and 24658 DF,  p-value: <2e-16
```

The model implies that if a patient has very urgent conditions, his waiting time will be greatly shortened, which is inline with our expectations.

We now test the specifications of our model with the Ramsey Reset (Ramsey 1969) to see whether there are any omitted variable bias in our model. The null hypothesis of this test is that the model has no omitted variable bias.

```
# run a Ramsey reset test to see whether we have omitted variable bias
resettest(reg_b_search)
```

```
RESET test

data:  reg_b_search
RESET = 39.7, df1 = 2, df2 = 24656, p-value < 2.2e-16
```

From the test result, we see that our model is misspecified. Here we have two guesses. The first is that perhaps the waiting time shortens with respect to age at an increasing rate. The second is that for very urgent situations, the effect of age may become more conspicuous. In terms of our model specification, we now look for a model including the first-order, second order and interaction terms of age, count and very_urgent terms.

First, we test whether addding an interaction term for very_urgent is statistically significant with a Chow Test (Chow 1960), after we have added the quadratic terms to the model.

```
# quadratic model
reg_b_quadratic <- lm(wait_time ~ count + I(count^2) + age + I(age^2),
                      data = project_data)
# Chow test
chow_model <- update(reg_b_quadratic, formula. = . ~ . * very_urgent)
chow_anova <- anova(chow_model)
chow_anova
```

```
Analysis of Variance Table

Response: wait_time
                          Df   Sum Sq Mean Sq F value  Pr(>F)
count                      1  1363966 1363966  807.80 < 2e-16 ***
I(count^2)                 1   118897  118897   70.42 < 2e-16 ***
age                        1    40197   40197   23.81 1.1e-06 ***
I(age^2)                   1     1779    1779    1.05   0.305
very_urgent                1    35710   35710   21.15 4.3e-06 ***
count:very_urgent          1     9756    9756    5.78   0.016 *
I(count^2):very_urgent     1     6331    6331    3.75   0.053 .
age:very_urgent            1      665     665    0.39   0.530
I(age^2):very_urgent       1     1242    1242    0.74   0.391
Residuals              24652 41624579    1688
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
chow <- lm(wait_time ~ (age), project_data)
chow_test_stat <- (sum(chow_anova[5:9, 2]) / 5) / chow_anova[10, 3]
chow_test_stat
```

```
[1] 6.36
```

```r
# the chow_test_stat is quite large, so we should add the indicator term
```

The test-statistic for Chow test is 6.361. Given that this test statistic follows an F-distribution, we see that adding the interaction term is statistically significant and we therefore include these terms in seraching for the model.

```r
# the chow_test_stat is quite large, so we should add the indicator term
# re run the algorithm to find our model
model_find_3 <- regsubsets(
  wait_time ~ (count + I(count^2) + age + I(age^2)) * very_urgent,
  data = project_data, nvmax = 20,
  method = "exhaustive"
)
```

As usual, we look at the BIC plot as in Figure 3 on Page **??**. The plot shows that we should use a three-variable model.

```r
# bic over # of predictors
qplot(x = 1:9, y = summary(model_find_3)$bic, geom = "line") +
  scale_x_continuous(breaks = 1:9) +
  labs(x = "Number of Predictors", y = "BIC",
       title = "Bayesian Information Criterion Over Number of Predictors")
```

We look to the output from the search function to see which three variables we should choose.

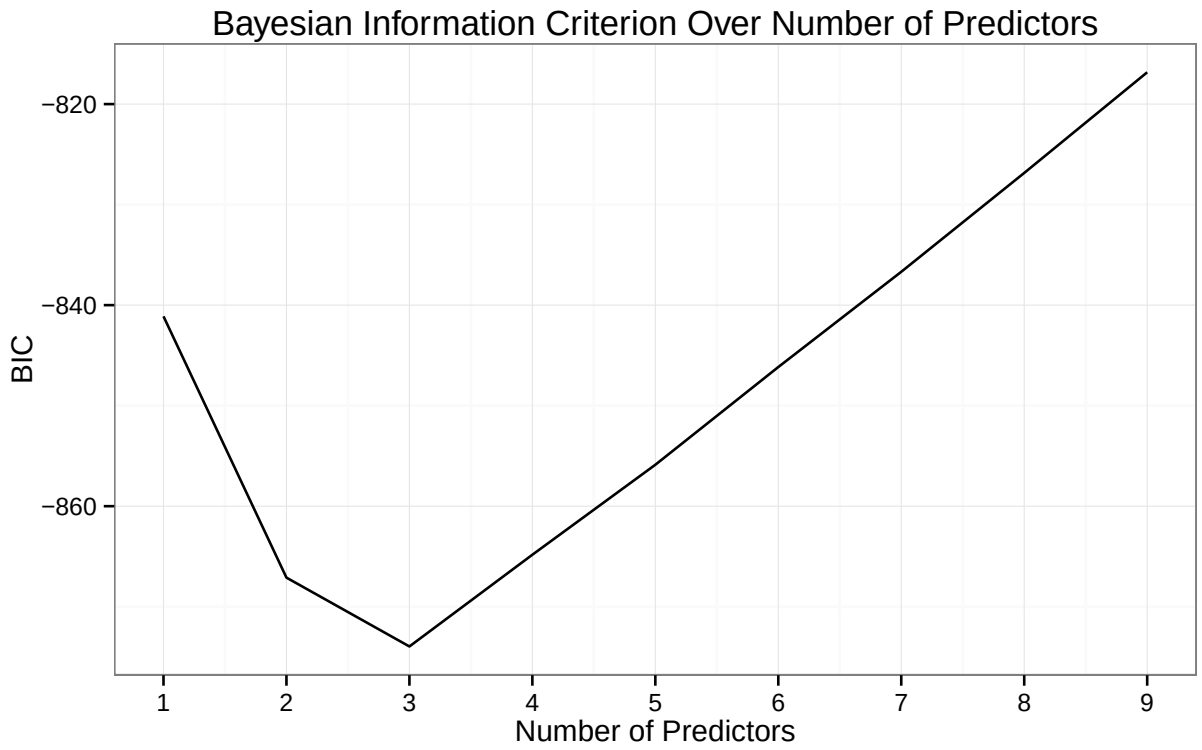

Figure 3: BIC for Third Serach Attempt

```
# take a look
summary(model_find_3)

Subset selection object
Call: regsubsets.formula(wait_time ~ (count + I(count^2) + age + I(age^2)) *
    very_urgent, data = project_data, nvmax = 20, method = "exhaustive")
9 Variables  (and intercept)
                      Forced in Forced out
count                     FALSE      FALSE
I(count^2)                FALSE      FALSE
age                       FALSE      FALSE
I(age^2)                  FALSE      FALSE
very_urgent               FALSE      FALSE
count:very_urgent         FALSE      FALSE
I(count^2):very_urgent    FALSE      FALSE
age:very_urgent           FALSE      FALSE
I(age^2):very_urgent      FALSE      FALSE
1 subsets of each size up to 9
Selection Algorithm: exhaustive
         count I(count^2) age I(age^2) very_urgent count:very_urgent
1  ( 1 ) " "   "*"        " " " "      " "         " "
2  ( 1 ) " "   "*"        " " " "      " "         " "
3  ( 1 ) " "   "*"        " " "*"      " "         " "
4  ( 1 ) " "   "*"        " " "*"      "*"         " "
5  ( 1 ) " "   "*"        " " "*"      "*"         "*"
6  ( 1 ) " "   "*"        " " "*"      "*"         "*"
7  ( 1 ) " "   "*"        " " "*"      "*"         "*"
8  ( 1 ) "*"   "*"        " " "*"      "*"         "*"
9  ( 1 ) "*"   "*"        "*" "*"      "*"         "*"
         I(count^2):very_urgent age:very_urgent I(age^2):very_urgent
1  ( 1 ) " "                    " "             " "
2  ( 1 ) "*"                    " "             " "
3  ( 1 ) "*"                    " "             " "
4  ( 1 ) "*"                    " "             " "
5  ( 1 ) "*"                    " "             " "
6  ( 1 ) "*"                    "*"             " "
7  ( 1 ) "*"                    "*"             "*"
8  ( 1 ) "*"                    "*"             "*"
9  ( 1 ) "*"                    "*"             "*"
```

The search function recommends that we should use age^2, count^2 and the interaction term between very_urgent and count in our three-variable model. We run this model.

```
# model from third search
reg_b_3 <- lm(wait_time ~ count : very_urgent + I(age^2) + I(count^2),
              project_data)
# get coef of this model
```

```
reg_b_3_coef <- coef(reg_b_3)
summary(reg_b_3)


Call:
lm(formula = wait_time ~ count:very_urgent + I(age^2) + I(count^2),
    data = project_data)

Residuals:
   Min     1Q Median     3Q    Max
 -58.1   -9.8   -4.5    4.0 1734.9

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       11.714310   0.474190   24.70  < 2e-16 ***
I(age^2)          -0.000465   0.000117   -3.97  7.2e-05 ***
I(count^2)         0.100018   0.003350   29.85  < 2e-16 ***
count:very_urgent -0.418246   0.080387   -5.20  2.0e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 41.1 on 24658 degrees of freedom
  (146 observations deleted due to missingness)
Multiple R-squared:  0.0363,Adjusted R-squared:  0.0362
F-statistic:  310 on 3 and 24658 DF,  p-value: <2e-16
```

We run a Ramsey Rest for this model to detect omitted variable bias.

```
resettest(reg_b_3)


RESET test

data:  reg_b_3
RESET = 0.518, df1 = 2, df2 = 24656, p-value = 0.5956
```

The test shows no statistically significant sign of omitted variable bias.

In this model, we see that there is not a single factor that seems to greatly influences the waiting time. Indeed, for an old patient aged at 80, an addtional year of age is only associated with 0.074 fewer minutes in wait time. Also, we notice that the explaining power of our model is very weak, smaller than 5%.

This result indicates that having adjusted for the patient's age, the number of visitors in the emergency room and the level of acuity, the intercept of the model still plays a huge part in explaining the variation in waiting time. Indeed, the estimated intercept, at 11.714 minutes, is very close to the median waiting time which is 12 minutes, another indicator that the variables in our model is not practically important in influencing the waiting time. This

could mean that there are some other unmeasured factors not included in our model or not even in the datasets.

Let us add the doctors (md) to see whether different doctors have different impact on the waiting time.

```
reg_b_final <- update(reg_b_3, formula. = . ~ . + md)
summary(reg_b_final)


Call:
lm(formula = wait_time ~ I(age^2) + I(count^2) + md + count:very_urgent,
    data = project_data)

Residuals:
   Min     1Q Median     3Q    Max
 -57.9   -9.7   -4.4    3.8 1733.2

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            8.959811   0.872060   10.27  < 2e-16 ***
I(age^2)              -0.000474   0.000117   -4.05  5.2e-05 ***
I(count^2)             0.099675   0.003363   29.64  < 2e-16 ***
mdDanescu, MD, Adrian  2.374930   1.052140    2.26  0.02400 *
mdFriend, MD, David I  3.731404   2.115494    1.76  0.07777 .
mdGharakhani, DO, Hoss 0.767513   1.054444    0.73  0.46669
mdKuban, MD, Alan L    4.020883   1.064088    3.78  0.00016 ***
mdMatts , MD, Christin -2.379071   3.631694   -0.66  0.51242
mdNorman, MD, Joyce    5.084523   1.046110    4.86  1.2e-06 ***
mdPollack, MD, Barry H 6.291992   1.971757    3.19  0.00142 **
mdSun, Nancy N         1.136242   1.104249    1.03  0.30350
mdTilles, MD, Ira H    4.487925   1.128022    3.98  7.0e-05 ***
mdVan Zyl, MD, Carin   0.976093   3.071290    0.32  0.75063
mdYu, MD, Alfred       4.343865   1.177093    3.69  0.00022 ***
count:very_urgent     -0.414604   0.080438   -5.15  2.6e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 41.1 on 24589 degrees of freedom
  (204 observations deleted due to missingness)
Multiple R-squared:  0.038,Adjusted R-squared:  0.0375
F-statistic: 69.4 on 14 and 24589 DF,  p-value: <2e-16
```

From the model's summary output we can tell that there is great variation amongst different doctors. So we can see that which doctor is in charge does play an important role in influecing the waiting time. However, the model's explaining power is still rather weak.

4 Conclusion

Our analysis shows net of patient's age, type of ailment, acuity and the doctor in charge, an additional visitor to the emergency room is associated with 110.969 seoncd longer waiting time.

For the second part, our analysis shows that doctors appear to play an important role in influencing the waiting time in the emergency room. Given that our analysis shows no strong effect of number of visitors in the emergency room on waiting time, we do not think that a shortage of doctors is a problem. Instead, given the variations amongst the doctors, the quality of each doctor's work seems to an essential consideration. Therefore for the managers at this hospital, we recommend that the hospital should invest in training its doctors to become more efficient at dealing with emergency room visit, rather than in recruiting more doctors.

However, our model is not perfect, since none of the models in this study has an R^2 greater than 0.1. we suspect that there are other unmeasured factors not included in our model or not even in the dataset that greatly influence the waiting time.

References

Chow, Gregory C (1960). "Tests of equality between sets of coefficients in two linear regressions". In: *Econometrica: Journal of the Econometric Society*, pp. 591–605.

Ramsey, James Bernard (1969). "Tests for specification errors in classical linear least-squares regression analysis". In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 350–371.

Schwarz, Gideon et al. (1978). "Estimating the dimension of a model". In: *The annals of statistics* 6.2, pp. 461–464.