Prediction of Length of Stay Using Linear Regression

Manali Phadke(map2039), Gaeun Kim (gk2501), Zhuyu Qiu(zq2157), Junting Ren(jr3755)

Abstract: *Background:* Length of stay is an important factor that influences medical cost. *Methods:* We used linear regression to find a model that would predict length of stay using automated and criterion based selection methods. *Results:* The overall model we created is significant and has an adjusted R^2 of 0.1216. The predicted value of days of stay is 2.315 times smaller or larger than the true days of stay when using the test dataset. *Conclusions:* While we were able to create a model, it is not an ideal one. Other methods should be used to create a better fitting model which can give more accurate prediction. Including additional covariates may also help in creating a better model.

1. Introduction

Length of stay in is an important variable to consider when a hospital is trying to reduce costs. If all necessary treatments have been completed, an extended hospital stay may not be beneficial [1]. In addition to the cost associated with a longer stay in the hospital, the patient is also at an increased risk of infection or other complications. However, if the length of stay is reduced too much, inadequate treatment or evaluation of the patient may result in a premature discharge of the patient [1]. Optimizing the length of stay would, therefore, be beneficial for both patients and hospitals. For this reason, in this analysis, we seek to understand the variables that can influence the length of stay and create a model that can be used to predict the length of stay.

2. Methods

The Data Analytics group from Good Health Corporation collected a total of 3682 records from 3612 patients. The data was divided into two equal training and testing sets to allow for cross validation of the model. Multiple linear regression was utilized to develop a model to predict length of stay using a variety of predictors. We selected a subset of predictors using automatic and criterion based methods. The final model was assessed through cross validation and bootstrap.

3. Results

Prior to any modeling, the data were examined using descriptive statistics and plots to identify potential errors in the data and necessary variable transformations. For categorical variables, the proportion of subjects in each category was examined and variables were condensed into a smaller number of categories that were more evenly distributed when appropriate. The predictors retained in the model are: is30dayreadmit, cindex (normal as the reference category), evisit, ageyear, heartrate, insurance type (not private insurance as the reference category), bpsystolic, temperature, and respiration rate.

The descriptive statistics showed that there were four problematic values in the data. For example, there was an individual with a body temperature of 52.3°C, which is biologically impossible. These values were removed from the dataset before proceeding with any further analysis. What's more, a natural log transformation was conducted on the outcome, length of stay (days) because it is a highly skewed variable (Appx. Figure 1). Additionally, because 98% of individuals didn't visit the ICU and people visited ICU will definitely have a long length of stay, we restricted our sample to those who didn't visit the ICU, as this would greatly increase the length of stay.

After examining the data, we proceeded to use forward, backward, and stepwise automated selection in SAS to obtain a preliminary model. All three selection methods gave the same model with 9 predictors. We then proceeded to criterion based selection. From the graph (Appx. Figure 2), we can see that cp is optimal as the number of parameters approaches 10. From the adjusted R^2 figure (Appx. Figure 2), we can see that after 10 parameters, the adjusted R^2 does not increase significantly. According to parsimony, we decided to choose the best model with 10 parameters. The model includes the following variables: is30dayreadmit, cindex, evisit, ageyear, heartrate, insurance type, bpsystolic, temperature, and respiration rate. The variables included in the criterion based model are the same as the automatic procedures.

Variable	n	mean	min	max	sd
loshours	3612	131.14	1.00	2111.00	142.11
losdays2	3612	5.46	.0042	87.96	5.92
mews	3449	4.11	1.00	14.00	1.70
ageyear	3612	65.67	18.00	105.00	18.69
bmi	2927	28.35	3.10	122.65	7.99
bpsystolic	3607	130.55	88.78	193.96	16.72
o2sat	3609	97.86	80.00	236.53	4.91
temperature	3610	36.73	11.85	52.28	8.99
heartrate	3607	80.07	37.58	242.58	13.00
respirationrate	3609	18.20	12.00	67.72	2.63
bpdiastolic	3611	72.52	29.56	154.40	9.80

Table 1: Descriptive Statistics of some key variables from the dataset including the number of observations, the mean, minimum, and maximum. Particularly, we note strange values for BMI and temperature.

Therefore, we then proceeded to finalize a model, which is shown below along with statistics:
$Log(losdays2) = -2.379 + 0.144 \cdot is30dayreadmit + 0.013 \cdot I(cindex = mild) + 0.108 \cdot I(cindex = moderate) + 0.221 \cdot I(cindex = severe) + 0.078 \cdot evisit + 0.009 \cdot ageyear + 0.006 \cdot heartrate - 0.162 \cdot I(Insurancetype = Private) + 0.006 \cdot bpsystolic + 0.080 \cdot temperature + 0.014 \cdot respirationrate$

The results from the F-test of the final model indicates that the model is significant at a significance level of 1%. By the adjusted R-square, 12% the variation of the outcome is explained by the model. The values of the slopes are interpreted as to how much of a percent increase in Y will occur for a unit increase in a particular X predictor variable, given that all other variables are held constant. Based on the estimated coefficients, the variables is20dayreadmit, cindex, ervisit, ageyear, heartrate, temperature, respirationrate have a positive association with length of stay while insurancetype and bpsystolic have a negative association with the outcome.

	Estimate	Std. Error	95 % CI	t value	Pr(>|t|)
(Intercept)	**-2.379014**	**1.015241**	**(-4.37,-0.38)**	**-2.343**	**0.019226**
is30dayreadmit	**0.144362**	**0.060483**	**(0.03 ,0.26)**	**2.387**	**0.017101**
cindex mild	**0.013417**	**0.050527**	**(-0.09, 0.11)**	**0.266**	**0.790620**
cindex moderate	**0.108423**	**0.070190**	**(-0.03,0.25)**	**1.545**	**0.122602**
cindex severe	**0.221426**	**0.061746**	**(0.10,0.34)**	**3.586**	**0.000345**
evisit	**0.077858**	**0.013708**	**(0.05,0.10)**	**5.680**	**1.58e-08**
ageyear	**0.009426**	**0.001243**	**(0.007 0.01)**	**7.584**	**5.44e-14**
heartrate	**0.006037**	**0.001612**	**(0.003,0.009)**	**3.745**	**0.000186**
Insurancetype Private	**-0.161951**	**0.044044**	**(-0.25,-0.07)**	**-3.677**	**0.000243**
bpsystolic	**-0.005782**	**0.001266**	**(-0.008,-0.003)**	**-4.567**	**5.29e-06**
temperature	**0.080258**	**0.026900**	**(0.03,0.13)**	**2.984**	**0.002889**
respirationrate	**0.013961**	**0.007030**	**(0.00,0.03)**	**1.986**	**0.047204**

Multiple R-squared: 0.1271, Adjusted R-squared: 0.1216,
F-statistic: 23.05 on 11 and 1741 DF, p-value: < 2.2e-16

Table 2: Final Model. This table gives the regression coefficients, their corresponding standard errors, t value, p- value, and 95 % CI

We now check the model assumptions of the final model through diagnostic plots (Appx Figure 3). We see the residuals form a horizontal band around zero from the Residuals vs Fitted plot. The scale-location plot shows the assumption of equal variance is met. We see three observations above 2.0 indicating possible outliers in the data. From the Residuals vs Leverage plot, we do not see any observations beyond the Cook's distance lines. The Normal Q-Q plot shows the lower standard residuals are deviating from the quantiles of the normal distribution. Therefore, the normality assumption may be questionable.

We also calculated the hat leverage value. The 219th observation has a value of 0.255, indicating a moderate leverage. We might consider it as an outlier in X. Based on the Cook's distance no observation is influential. We also checked outliers in Y by calculating the studentized residuals and we detect 29 observations which might be considered as outliers in Y. Then we calculated the DFFITS values to see if there are any influential observations. Considering the dataset is large, we used the $2\sqrt{p/n}$ = 0.15 as a criterion. We detected 110 observations which can be considered as influential. By cross-checking the outliers in X, Y, and

influential points, we found that all outliers are influential points. Compared to estimated model results using all observations, the adjusted R-squared increased to 0.1286 by 6% and more than half of the parameters changed over 10% when influential points were removed. Therefore, we can conclude these observations are influential. However, since these are not due to measurement errors, we can keep these observations in the data set. Finally, we checked the final model for multicollinearity using VIF and found no indication of a problem.

Prediction capability of our model was checked by two methods. We use bootstrap to calculate the variability of the coefficient and adjusted R^2. Although we see that the bias for coefficients is small, their standard errors are extremely large. This is understandable since our model had a low adjusted R^2. The variability of adjusted R^2 is relatively small, but the bootstrap mean value of adjusted R^2 is just 0.12, same as the original data.

The other 50% of the data that we left out to validate our prediction. We obtained a mean square prediction error (MSPE) of 0.70. This was transformed to the original units: log(y/y_fit)=exp{sqr(MSPE)}=2.315. Therefore, we can interpret it as: the predicted value of length of stay is 2.315 times smaller or larger than the true length of stay, which is inaccurate.

4. Conclusion

We were able to generate a linear regression model for length of stay using several predictors. However, it is clear that our model could be improved upon. The goodness of fit is quite low, and cross validation shows that prediction using the model deviates from the true value quite a lot. We recommend using other methods such as polynomial regression, piecewise and nonparametric models to produce a better model. Including additional covariates may also help create a better model.

References

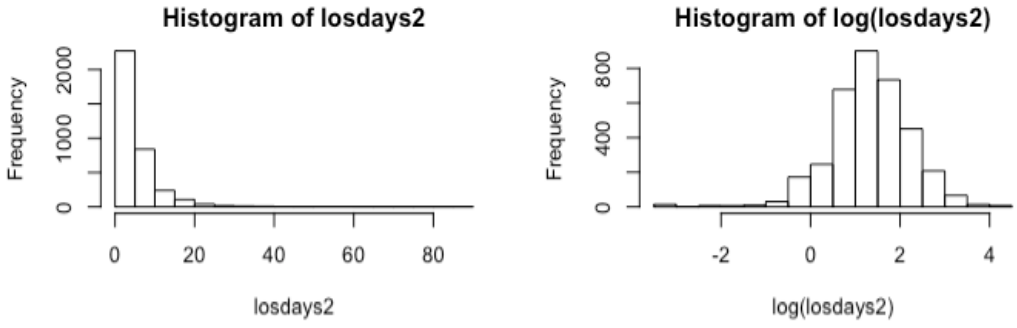
[1] Kossovsky MP, Sarasin FP, Chopard P*, et al.* Relationship between hospital length of stay and quality of care in patients with congestive heart failure. *BMJ Quality & Safety* 2002;**11:**219-223.

Appendix:

Table 1: Model without Outliers

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	**-2.502284**	**1.031749**	**--2.425**	**0.015406**
is30dayreadmit	**0.159017**	**0.062685**	**2.537**	**0.011282**
cindex mild	**0.012930**	**0.051976**	**0.249**	**0.803576**
cindex moderate	**0.157723**	**0.073712**	**2.140**	**0.032528**
cindex severe	**0.224033**	**0.063415**	**3.533**	**0.000423**
evisit	**0.084037**	**0.014189**	**5.923**	**3.86e-09**
ageyear	**0.009203**	**0.001290**	**7.135**	**1.45e-12**
heartrate	**0.005541**	**0.001652**	**3.354**	**0.000814**
Insurancetype Private	**-0.148106**	**0.045554**	**-3.251**	**0.001173**
bpsystolic	**-0.006082**	**0.001303**	**-4.666**	**3.33e-06**
temperature	**0.083171**	**0.027356**	**3.040**	**0.002401**
respirationrate	**0.018648**	**0.007437**	**2.507**	**0.012259**

Figure 1: Histogram of length of stay (days)



Histogram of losdays2 shows that the distribution of the outcome variable is highly skewed to the right. When applying log transformation on losdays2, we can see that the outlying data points from the right tail are brought towards the rest of the data. However, compared to a normal distribution, the log(losdays2) is slightly skewed to the left.

Figure 2: Graphs of Cp and R^2 vs. number of parameters

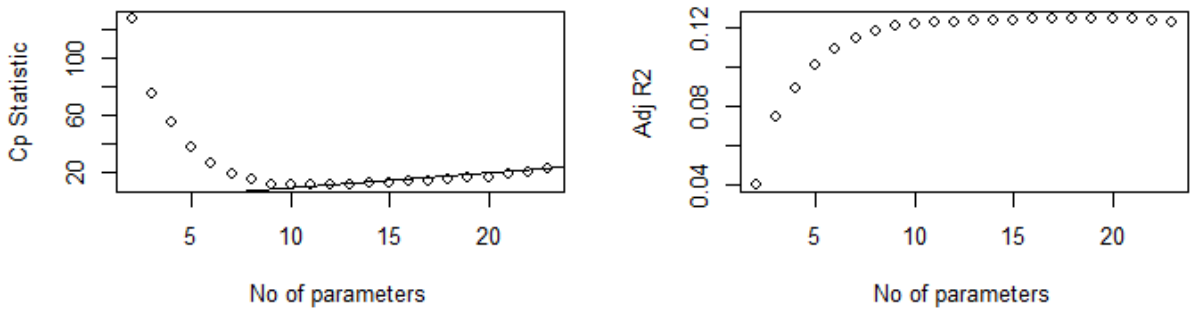
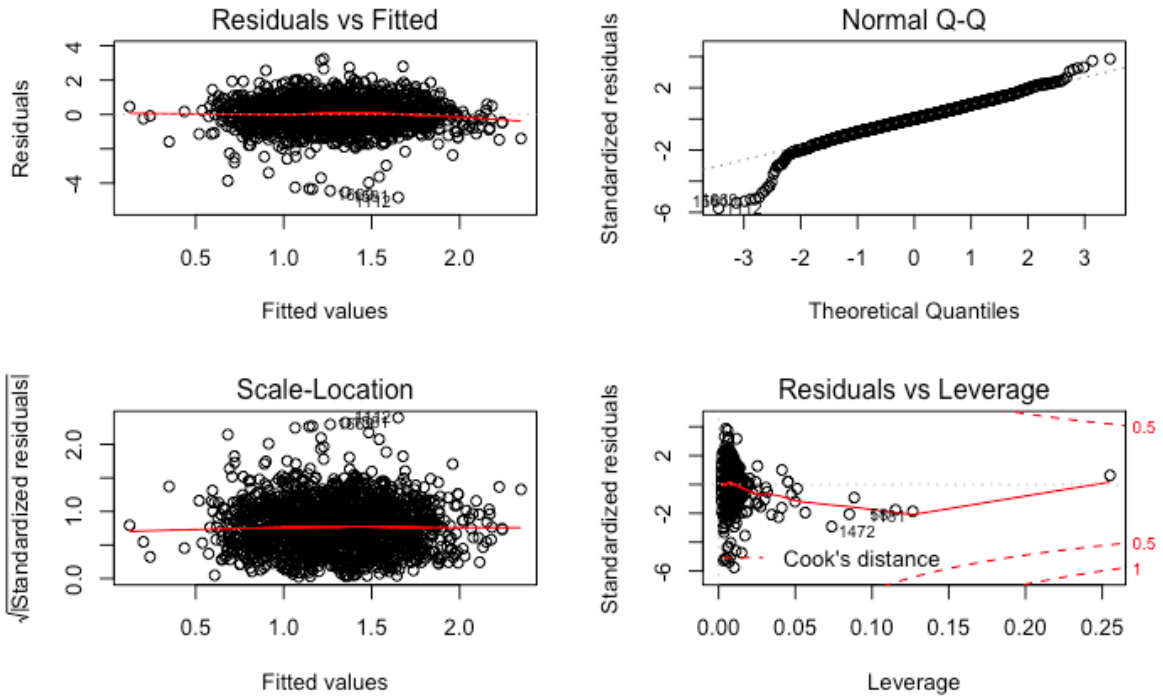


Figure 3: Diagnostic Plots for Final Model



R Code:

```
---
title: "BMI_final_project"
author: "Junting Ren Uni:jr3755"
date: "December 9, 2017"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
library(leaps)
library(readxl)
library(janitor)
library(psych)
library(HH)
library(purrr)
library(tidyr)
library(ggplot2)
library(boot)
library(tidyverse)
```

##reading and cleaning the data
```{r reading and cleaning the data}
gh_project_raw <- read_excel("GHProject_Dataset.xlsx", sheet = 1) %>%
  clean_names() %>%
  mutate(mews = as.factor(mews), cindex = as.factor(cindex),
         icu_flag = as.factor(icu_flag), gender = as.factor(gender), race =
           as.factor(race), religion = as.factor(religion), maritalstatus =
           as.factor(maritalstatus), insurancetype = as.factor(insurancetype))
```
```

```
gh_duplicate <- gh_project_raw[gh_project_raw$patientid %in%
gh_project_raw$patientid[duplicated(gh_project_raw$patientid)],] %>%
  group_by(patientid) %>%
  count()

##We got 69 patients with more than 1 visit, 68 had 2 visits, 1 had 3 visits.
dim(gh_duplicate)

gh_project <- gh_project_raw %>%
  arrange(patientid, visitid) %>%
  filter(!duplicated(patientid))
```

```{r descriptive statistics}
des_stat = gh_project %>%
  select(-gender, -race, -maritalstatus, -facilityname, -insurancetype, -facilityzip, -admitdtm, -religion, -postalcode, -is30dayreadmit, -patientid, -visitid -mews, -cindex, -icu_flag) %>%
  describe(na.rm = TRUE)

#there is a BMI of 3.1 (min) and 122.7.  Both of these are impossible
#there is a min temp of 11.85 52.275.  Again, both of these are incorrect.
# funny bp values as well??

gender = table(gh_project$gender) %>% prop.table()
race = table(gh_project$race) %>% prop.table()
marital_status = table(gh_project$maritalstatus)  %>% prop.table()
insurance = table(gh_project$insurancetype) %>% prop.table()
religion = table(gh_project$religion) %>% prop.table()
readmit = table(gh_project$is30dayreadmit) %>% prop.table()
mews = table(gh_project$mews) %>% prop.table()
cindex = table(gh_project$cindex) %>% prop.table()
icu_flag = table(gh_project$icu_flag) %>% prop.table()
```

```
gh_project = gh_project %>%
  filter(bmi < 100 | is.na(bmi)) %>%
  filter(temperature < 45 | is.na(temperature)) %>%
  filter(temperature > 15 | is.na(temperature))
```

```{r recoding categorical var}
library(forcats)
gh_project$race = fct_collapse(gh_project$race, White = "White", Non_White = c("African Amer/Black","Asian","Native Amer/Alaskan","Natv Hawaii/Pacf Isl", "Other/Multiracial"))

gh_project$maritalstatus = fct_collapse(gh_project$maritalstatus, Married = "Married", Not_Married = c("Civil Union","Divorced","Separated","Single", "Widowed"))

gh_project$insurancetype = fct_collapse(gh_project$insurancetype, Private = "Private", Not_Private = c("Medicaid","Medicare"))

gh_project$religion = fct_collapse(gh_project$religion, Catholic = "Catholic", Christian = "Christian", Other = c("Angelican","Hebrew","Hindu","Islam", "Jewish", "Mormon", "No Affiliation", "Non Denominational", "Other"))

gh_project$insurancetype = fct_collapse(gh_project$insurancetype, Private = "Private", Not_Private = c("Medicaid","Medicare"))

gh_project$mews = fct_collapse(gh_project$mews, Normal = c("0", "1"), Increase_caution = c("2", "3"), Further_deterioration = c("4", "5"), Immediate_action_required = c("6", "7", "8", "9", "10", "11", "12", "14"))

gh_project$cindex = fct_collapse(gh_project$cindex, normal = "0", mild = c("1", "2"), moderate = "3", severe = "5")

gh_project = gh_project %>%
  filter(icu_flag == "0")
# model only applies to patients who did not go to ICU
```

```

```{r plots}
gh_project %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_histogram()

attach(gh_project)

hist(losdays2)
hist(log(losdays2))
```

```{r transformations}
gh_project =
  gh_project %>%
  mutate(log_losdays2 = log(losdays2))

#write dataset as csv
#write_csv(sas_train_data, path = "C:/Users/mphad/Documents/1st semester/Intro to Biostats/sas_train_data.csv")
```
```{r model selection}
set.seed(1)
train <- sample(3543,1772)

train_data <- gh_project[train,]
test_data <- gh_project[-train,]

```
best <- function(model, ...)
{
  subsets <- regsubsets(formula(model), model.frame(model), ...)
  subsets <- with(summary(subsets),
                  cbind(p = as.numeric(rownames(which)), which, rss, rsq, adjr2, cp, bic))

  return(subsets)
}

##Final datset for model
train_data <- train_data %>%
  select(-patientid, -visitid, -admitdtm, -icu_flag, -postalcode,
         -facilityname, -facilityzip, -loshours, -losdays2)

test_data <- test_data %>%
  select(-patientid, -visitid, -admitdtm, -icu_flag, -postalcode,
         -facilityname, -facilityzip, -loshours, -losdays2)

multi_fit <- lm(log_losdays2 ~ ., data = train_data)
View(best(multi_fit, nvmax = 30, method = "exhaustive"))
final_model <- lm(log_losdays2 ~ is30dayreadmit + cindex +
                    evisit + ageyear + heartrate +
                    insurancetype + bpsystolic + temperature +
                    respirationrate, data = train_data)

vif(final_model)
gh_project %>%
  dplyr::select(bpdiastolic, bpsystolic) %>%
  na.omit() %>%
  cor(x = .)
```

```{r final model}
plot_crit <- regsubsets(log_losdays2 ~ ., data = train_data, nvmax = 30)
rs<-summary(plot_crit)
par(mfrow=c(2,2))
plot(2:23, rs$cp, xlab="No of parameters", ylab="Cp Statistic")
abline(0,1)
plot(2:23, rs$adjr2, xlab="No of parameters", ylab="Adj R2")
```

```{r check model assumption}
par(mfrow = c(2,2))
plot(final_model)
# normality assumption - skewed to the left
# obs 1112, 1581, 1669

gh_project_raw %>%
  arrange(losdays2)
# all 3 subjects lowest losdays2
# lets see after deleting

#outliers in y  (29)
stu_res<-rstandard(final_model)
outliers_y = as.data.frame(stu_res)
n = c(1:1753)
outliers_y = cbind(n,outliers_y)

outliers_y = outliers_y %>%
  filter(abs(stu_res)>2.5) %>%
  mutate(stu = stu_res[abs(stu_res)])

#outliers in x, hat value >0.5 very high, >2p/n = 0.0087 high, 0.2~0.5 moderate (1)
#15 parameters
```

```
hat =  hatvalues(final_model)
hat = round(hat,4)
n = c(1:1753)
hat = cbind(n,hat)
hat_inf = as.data.frame(hat)%>%
  clean_names()%>%
  filter(hat > 0.2)

#influential observation (parameters = 16)
#cook's distance >0.5, >0.2
cook = cooks.distance(final_model)
cook = as.data.frame(cook)
cook%>%
filter(cook>0.2)
#no cook's distance > 0.2

#dffits>2sqrt(p/n) = 0.1323 or 1
#influence = influence.measures(final_model)
#summary(influence)
#plot(rstudent(final_model) ~ hatvalues(final_model))
dffit = dffits(final_model)
dffit = cbind(n,dffit)
dffit = as.data.frame(dffit)
inf_dffit = dffit %>%
  filter(abs(dffit)>0.15)
num_dffit = inf_dffit$n#110

#influential x outliers
x_inf = left_join(hat_inf,dffit,by = "n")%>%
  filter(abs(dffit)>0.15)
num_inf = x_inf$n #1
```

```
#influential y outliers
y_inf = left_join(outliers_y,dffit,by = "n")%>%
  filter(abs(dffit)>0.15)
num_inf = c(num_inf,y_inf$n)%>%
  unique() #30

#omit NAs in the dataset
gh_project_remove = train_data %>%
  dplyr::select(is30dayreadmit, cindex, evisit, ageyear, insurancetype, heartrate, respirationrate, log_losdays2,
temperature,bpsystolic) %>%
  na.omit()

#remove influential points
gh_project_remove1 = gh_project_remove
for(i in 1:131){
gh_project_remove1 = gh_project_remove1[-num_dffit[i],]}

final_model_remove1 <- lm(log_losdays2 ~ is30dayreadmit + cindex +
                    evisit + ageyear + heartrate +
                    insurancetype + bpsystolic + temperature +
                    respirationrate, data = gh_project_remove1)

summary(final_model_remove1)

#remove influential x,y outliers
gh_project_remove2 = gh_project_remove
for(i in 1:length(num_inf)){
gh_project_remove2 = gh_project_remove2[-num_inf[i],]}

final_model_remove2 <- lm(log_losdays2 ~ is30dayreadmit + cindex +
```

```
                evisit + ageyear + heartrate +
                insurancetype + bpsystolic + temperature +
                respirationrate,, data = gh_project_remove2)

summary(final_model_remove2)

summary(final_model)

par(mfrow=c(2,2))
plot(final_model_remove1)

par(mfrow=c(2,2))
plot(final_model_remove2)

#remove 3 observations based on residuals
gh_project_remove3 = gh_project_remove
x = c(1112, 1581, 1669)
gh_project_remove3 = gh_project_remove3[-x,]
final_model_remove3 <- lm(log_losdays2 ~ is30dayreadmit + cindex +
                evisit + ageyear + heartrate +
                insurancetype + bpsystolic + temperature +
                respirationrate, data = gh_project_remove3)

summary(final_model_remove3)

par(mfrow=c(2,2))
plot(final_model_remove3)
```

```{r model validation}
#bootstrap to access the model coefficient varibility
```

```
boot.fn <- function(data, index){
        return(coef(lm(log_losdays2 ~ is30dayreadmit + cindex +
            evisit + ageyear + temperature +
            insurancetype + bpsystolic + heartrate +
            respirationrate, subset = index,
                data = train_data)))
}

boot.adj.r <- function(data, index){
        return(summary(lm(log_losdays2 ~ is30dayreadmit + cindex +
            evisit + ageyear + temperature +
            insurancetype + bpsystolic + heartrate +
            respirationrate, subset = index,
                data = train_data))$adj.r.squared)
}
boot_coef <- boot(train_data, boot.fn, 100)
boot_adj_r <- boot(train_data, boot.adj.r, 100)

# How does it compare to the original (non-bootstrap) estimates?
summary(final_model)

# Use predict() and mean to calculate the MSPE for the 1317 obs used in validation.
# MSPE mean square prediction error.
test_data <- na.omit(test_data)
MSPE <- mean((test_data$log_losdays2 - predict(final_model, test_data))^2)
##getting the length of stay error in days
exp(sqrt(MSPE))
##2.31 days of prediction error
```

SAS Code:

```
PROC Import datafile = "sas_train_data.csv" DBMS = CSV out = GH_data;
GETNAMES = YES;
```

```
run;
proc contents data = GH_data;
run;

*forward automated selection;

PROC glm data = GH_data;
class is30dayreadmit mews cindex gender religion race maritalstatus religion
insurancetype;
model log_losdays2 = is30dayreadmit mews cindex evisit ageyear gender race
religion maritalstatus insurancetype bmi bpdiastolic bpsystolic heartrate
o2sat respirationrate temperature;
run;


PROC glmselect data = GH_data;
class is30dayreadmit mews cindex gender religion race maritalstatus religion
insurancetype;
model log_losdays2 = is30dayreadmit mews cindex is30dayreadmit evisit ageyear
gender race religion maritalstatus insurancetype bmi bpdiastolic bpsystolic
heartrate o2sat respirationrate temperature/ selection = forward(select = sl
sle=0.15) stats = (adjrsq aic cp);
run;

*backward automated selection;
PROC glmselect data = GH_data;
class is30dayreadmit mews cindex gender religion race maritalstatus religion
insurancetype;
model log_losdays2 = mews is30dayreadmit cindex evisit ageyear gender race
religion maritalstatus insurancetype bmi bpdiastolic bpsystolic heartrate
o2sat respirationrate temperature/ selection = backward(select = sl sls=0.15)
stats = (adjrsq aic cp);
run;

*stepwaise automated selection;
PROC glmselect data = GH_data;
class is30dayreadmit mews cindex gender religion race maritalstatus religion
insurancetype;
model log_losdays2 = mews is30dayreadmit cindex evisit ageyear gender race
religion maritalstatus insurancetype bmi bpdiastolic bpsystolic heartrate
o2sat respirationrate temperature/ selection = stepwise(select = sl sle = .15
sls=0.15) stats = (adjrsq aic cp);
run;
```