FinalProject

Hyewon Choi

12/8/2017

Loading the data, observing duplicates

```
library(readr)
library(readxl)
library(janitor)

library(pander)
library(leaps)
library(boot)
library(tidyverse)
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
```

```
## Conflicts with tidy packages ----------------------------------------------
```

```
## filter(): dplyr, stats
## lag():    dplyr, stats
```

```
hos = read_excel("./data/GHProject_Dataset.xlsx")
```

```
## Warning in strptime(x, format, tz = tz): unknown timezone 'default/America/
## New_York'
```

```
duplicates = data.frame(table(hos$PatientID))

duplicates = duplicates[duplicates$Freq > 1,] %>%
  rename(PatientID = Var1)
#69 patients visited the hospital more than once, 68 visited twice and 1 visited thre
e times

hos_duplicates = merge(duplicates, hos)
```

Tidying the data

```
hos_tidy <- hos %>%
  clean_names() %>%
  dplyr::select(-loshours, -postalcode, -facilityname, -facilityzip) %>%
  dplyr::group_by(patientid) %>%
  dplyr::filter(!duplicated(patientid)) %>%
  ungroup(patientid)
```

Data observations & building dummies

```
histogram <- hos_tidy %>%
  ggplot(aes(x = losdays2)) +
    geom_histogram() +
    labs(title = "Figure 1: Length of Stay",
         x = "Length of Stay (Days)",
         y = "Count")
```

```
fill_na = function(x) {
  if (is.numeric(x)){
   mean = mean(x, na.rm = TRUE)
   x = replace(x, is.na(x), mean)
  } else {x = x}

  return(x)
}

hos_tidy_omitna = map_df(hos_tidy, fill_na)
```

```
outliersZ <- function(data, zCutOff = 3.291, replace = NA, values = FALSE, digits = 2
) {
    #compute standard deviation (sample version n = n [not n-1])
    stdev <- sqrt(sum((data - mean(data, na.rm = T))^2, na.rm = T) / sum(!is.na(data)
))
    #compute absolute z values for each value
    absZ <- abs(data - mean(data, na.rm = T)) / stdev
    #subset data that has absZ greater than the zCutOff and replace them with replace
    #can also replace with other values (such as max/mean of data)
    data[absZ > zCutOff] <- replace

    if (values == TRUE) {
        return(round(absZ, digits)) #if values == TRUE, return z score for each value
    } else {
        return(round(data, digits)) #otherwise, return values with outliers replaced
    }
}

hos_tidy_omitna$bmi = outliersZ(hos_tidy_omitna$bmi, zCutOff = 3.291, replace = NA, v
alues = FALSE, digits = 2)

hos_tidy_omitna$temperature = outliersZ(hos_tidy_omitna$temperature, zCutOff = 3.291,
replace = NA, values = FALSE, digits = 2)

hos_tidy_omitna$heartrate = outliersZ(hos_tidy_omitna$heartrate, zCutOff = 3.291, rep
lace = NA, values = FALSE, digits = 2)

hos_tidy_omitna$respirationrate = outliersZ(hos_tidy_omitna$respirationrate, zCutOff
= 3.291, replace = NA, values = FALSE, digits = 2)

hos_tidy_omitna$bpdiastolic = outliersZ(hos_tidy_omitna$bpdiastolic, zCutOff = 3.291,
replace = NA, values = FALSE, digits = 2)

hos_tidy_omitna$bpsystolic = outliersZ(hos_tidy_omitna$bpsystolic, zCutOff = 3.291, r
eplace = NA, values = FALSE, digits = 2)

## 99.9% cut off
```

```
fill_na = function(x) {
  if (is.numeric(x)){
   mean = mean(x, na.rm = TRUE)
   x = replace(x, is.na(x), mean)
  } else {x = x}

  return(x)
}

hos_tidy_omitna = map_df(hos_tidy_omitna, fill_na)

hos_tidy_omitna = hos_tidy_omitna %>%
  filter(!o2sat > 100)
```

Summary Statistics

```
summary_losdays2 = hos_tidy_omitna %>%
  dplyr::select(losdays2) %>%
summarize(variable = names(.),
            n = n(), mean = mean(losdays2),
            sd = sd(losdays2),
            minimum = min(losdays2),
            maximum = max(losdays2),
            median = median(losdays2))


summary_ageyear = hos_tidy_omitna %>%
  dplyr::select(ageyear) %>%
  summarize(variable = names(.), n = n(),
            mean = mean(ageyear),
            sd = sd(ageyear),
            minimum = min(ageyear),
            maximum = max(ageyear),
            median = median(ageyear))


summary_evisit = hos_tidy_omitna %>%
 dplyr:: select(evisit) %>%
  summarize(variable = names(.),
            n = n(),
            mean = mean(evisit),
            sd = sd(evisit),
            minimum = min(evisit),
            maximum = max(evisit),
            median = median(evisit))

summary_bmi = hos_tidy_omitna %>%
  dplyr::select(bmi) %>%
  summarize(variable = names(.),
            n = n()-sum(is.na(hos_tidy$bmi)),
            mean = mean(na.omit(bmi)),
            sd = sd(na.omit(bmi)),
            minimum = min(na.omit(bmi)),
            maximum = max(na.omit(bmi)),
            median = median(na.omit(bmi)))


summary_bpsystolic = hos_tidy_omitna %>%
  dplyr::select(bpsystolic) %>%
  summarize(variable = names(.),
            n = n()-sum(is.na(hos_tidy$bpsystolic)),
            mean = mean(na.omit(bpsystolic)),
            sd = sd(na.omit(bpsystolic)),
            minimum = min(na.omit(bpsystolic)),
            maximum = max(na.omit(bpsystolic)),
            median = median(na.omit(bpsystolic)))


summary_o2sat = hos_tidy_omitna %>%
  dplyr::select(o2sat) %>%
  summarize(variable = names(.),
            n = n()-sum(is.na(hos_tidy$o2sat)),
            mean = mean(na.omit(o2sat)),
            sd = sd(na.omit(o2sat)),
            minimum = min(na.omit(o2sat)),
            maximum = max(na.omit(o2sat)),
            median = median(na.omit(o2sat)))


summary_temperature = hos_tidy_omitna %>%
  dplyr::select(temperature) %>%
  summarize(variable = names(.),
            n = n()-sum(is.na(hos_tidy$temperature)),
            mean = mean(na.omit(temperature)),
            sd = sd(na.omit(temperature)),
            minimum = min(na.omit(temperature)),
            maximum = max(na.omit(temperature)),
            median = median(na.omit(temperature)))


summary_heartrate = hos_tidy_omitna %>%
  dplyr::select(heartrate) %>%
  summarize(variable = names(.),
            n = n()-sum(is.na(hos_tidy$heartrate)),
            mean = mean(na.omit(heartrate)),
            sd = sd(na.omit(heartrate)),
            minimum = min(na.omit(heartrate)),
            maximum = max(na.omit(heartrate)),
            median = median(na.omit(heartrate)))


summary_respirationrate = hos_tidy_omitna %>%
  dplyr::select(respirationrate) %>%
summarize(variable = names(.),
            n = n()-sum(is.na(hos_tidy$respirationrate)),
            mean = mean(na.omit(respirationrate)),
            sd = sd(na.omit(respirationrate)),
            minimum = min(na.omit(respirationrate)),
            maximum = max(na.omit(respirationrate)),
            median = median(na.omit(respirationrate)))


summary_bpdiastolic = hos_tidy_omitna %>%
  dplyr::select(bpdiastolic) %>%
  summarize(variable = names(.),
            n = n()-sum(is.na(hos_tidy$bpdiastolic)),
            mean = mean(na.omit(bpdiastolic)),
            sd = sd(na.omit(bpdiastolic)),
            minimum = min(na.omit(bpdiastolic)),
            maximum = max(na.omit(bpdiastolic)),
            median = median(na.omit(bpdiastolic)))

summary = rbind(summary_losdays2, summary_ageyear, summary_evisit, summary_bmi, summa
ry_bpsystolic, summary_o2sat, summary_temperature, summary_heartrate, summary_respira
tionrate, summary_bpdiastolic)
table_summary <-pander(summary)
```

Stepwise selection

```
hos_tidy_omitna = hos_tidy_omitna %>%
  mutate(log_losdays2 = log(losdays2)) %>%
  na.omit()
mult.fit <- lm(log_losdays2 ~ is30dayreadmit + ageyear + evisit+ cindex + maritalstat
us + insurancetype + race + respirationrate + o2sat + heartrate + bmi + temperature +
bpsystolic + bpdiastolic + mews + icu_flag, data = hos_tidy_omitna)
summary_multfit <-summary(mult.fit)

 z <- step(mult.fit, direction = 'both')
```

```
## Start:  AIC=-1262.19
## log_losdays2 ~ is30dayreadmit + ageyear + evisit + cindex + maritalstatus +
##     insurancetype + race + respirationrate + o2sat + heartrate +
##     bmi + temperature + bpsystolic + bpdiastolic + mews + icu_flag
##
##                    Df Sum of Sq    RSS     AIC
## - maritalstatus    5     4.819 2338.4 -1265.1
## - race             5     5.141 2338.7 -1264.7
## - mews             1     0.126 2333.7 -1264.0
## - icu_flag         1     0.461 2334.0 -1263.5
## <none>                         2333.6 -1262.2
## - bpdiastolic      1     2.537 2336.1 -1260.5
## - bmi              1     3.092 2336.7 -1259.7
## - insurancetype    2     9.468 2343.0 -1252.3
## - is30dayreadmit   1    13.090 2346.7 -1245.0
## - cindex           1    13.622 2347.2 -1244.3
## - bpsystolic       1    13.717 2347.3 -1244.1
## - temperature      1    13.996 2347.6 -1243.7
## - o2sat            1    14.396 2348.0 -1243.1
## - heartrate        1    23.290 2356.8 -1230.2
## - evisit           1    28.012 2361.6 -1223.3
## - respirationrate  1    30.003 2363.6 -1220.4
## - ageyear          1    46.854 2380.4 -1196.1
## 
## Step:  AIC=-1265.12
## log_losdays2 ~ is30dayreadmit + ageyear + evisit + cindex + insurancetype + 
##     race + respirationrate + o2sat + heartrate + bmi + temperature + 
##     bpsystolic + bpdiastolic + mews + icu_flag
## 
##                   Df Sum of Sq    RSS     AIC
## - mews             1     0.212 2338.6 -1266.8
## - icu_flag         1     0.403 2338.8 -1266.5
## - race             5     6.614 2345.0 -1265.5
## <none>                         2338.4 -1265.1
## - bpdiastolic      1     2.751 2341.1 -1263.1
## + maritalstatus    5     4.819 2333.6 -1262.2
## - bmi              1     3.707 2342.1 -1261.7
## - insurancetype    2    11.040 2349.4 -1253.0
## - is30dayreadmit   1    13.033 2351.4 -1248.1
## - cindex           1    13.401 2351.8 -1247.5
## - temperature      1    13.451 2351.8 -1247.5
## - bpsystolic       1    13.736 2352.1 -1247.1
## - o2sat            1    14.798 2353.2 -1245.5
## - heartrate        1    23.596 2362.0 -1232.7
## - evisit           1    29.514 2367.9 -1224.2
## - respirationrate  1    29.926 2368.3 -1223.6
## - ageyear          1    46.883 2385.3 -1199.1
## 
## Step:  AIC=-1266.81
## log_losdays2 ~ is30dayreadmit + ageyear + evisit + cindex + insurancetype + 
##     race + respirationrate + o2sat + heartrate + bmi + temperature + 
##     bpsystolic + bpdiastolic + icu_flag
## 
##                   Df Sum of Sq    RSS     AIC
## - icu_flag         1     0.393 2339.0 -1268.2
## - race             5     6.683 2345.3 -1267.0
## <none>                         2338.6 -1266.8
## + mews              1     0.212 2338.4 -1265.1
## - bpdiastolic       1     2.959 2341.6 -1264.5
## + maritalstatus     5     4.905 2333.7 -1264.0
## - bmi               1     3.529 2342.1 -1263.7
## - insurancetype     2    11.194 2349.8 -1254.5
## - is30dayreadmit    1    12.975 2351.6 -1249.9
## - cindex            1    13.404 2352.0 -1249.2
## - temperature       1    13.629 2352.2 -1248.9
## - bpsystolic        1    13.686 2352.3 -1248.8
## - o2sat             1    14.942 2353.5 -1247.0
## - heartrate         1    25.937 2364.5 -1231.0
## - evisit            1    29.582 2368.2 -1225.8
## - respirationrate   1    31.166 2369.8 -1223.5
## - ageyear           1    60.823 2399.4 -1180.9
##
## Step:  AIC=-1268.24
## log_losdays2 ~ is30dayreadmit + ageyear + evisit + cindex + insurancetype +
##     race + respirationrate + o2sat + heartrate + bmi + temperature +
##     bpsystolic + bpdiastolic
##
##                   Df Sum of Sq    RSS     AIC
## - race             5     6.668 2345.7 -1268.5
## <none>                         2339.0 -1268.2
## + icu_flag         1     0.393 2338.6 -1266.8
## + mews             1     0.201 2338.8 -1266.5
## - bpdiastolic      1     3.080 2342.1 -1265.7
## + maritalstatus    5     4.844 2334.1 -1265.3
## - bmi              1     3.506 2342.5 -1265.1
## - insurancetype    2    11.030 2350.0 -1256.1
## - is30dayreadmit   1    12.962 2351.9 -1251.3
## - cindex           1    13.259 2352.2 -1250.9
## - temperature      1    13.606 2352.6 -1250.4
## - bpsystolic       1    13.619 2352.6 -1250.3
## - o2sat            1    14.803 2353.8 -1248.6
## - heartrate        1    26.045 2365.0 -1232.3
## - evisit           1    29.276 2368.3 -1227.6
## - respirationrate  1    30.957 2369.9 -1225.2
## - ageyear          1    60.481 2399.5 -1182.8
##
## Step:  AIC=-1268.49
## log_losdays2 ~ is30dayreadmit + ageyear + evisit + cindex + insurancetype +
##     respirationrate + o2sat + heartrate + bmi + temperature +
##     bpsystolic + bpdiastolic
##
##                   Df Sum of Sq    RSS     AIC
## <none>                         2345.7 -1268.5
## + race             5     6.668 2339.0 -1268.2
## + maritalstatus    5     6.367 2339.3 -1267.8
## + icu_flag         1     0.379 2345.3 -1267.0
## + mews             1     0.269 2345.4 -1266.9
## - bpdiastolic      1     2.620 2348.3 -1266.7
## - bmi              1     2.842 2348.5 -1266.3
## - insurancetype    2    10.255 2355.9 -1257.5
## - o2sat            1    11.665 2357.3 -1253.5
## - bpsystolic       1    12.089 2357.7 -1252.9
## - is30dayreadmit   1    12.429 2358.1 -1252.4
## - temperature      1    14.164 2359.8 -1249.9
## - cindex           1    14.726 2360.4 -1249.0
## - heartrate        1    25.298 2370.9 -1233.8
## - respirationrate  1    31.441 2377.1 -1224.9
## - evisit           1    32.828 2378.5 -1222.9
## - ageyear          1    58.048 2403.7 -1186.8
```

```
sum_z <- summary(z)

#mews score based on bp, respiration, heartrate & temp and is a less significant vari
able than bp, resp $ temp
#the two bp readings are correlated
#of all the dummy variables, only insurancetype was significant
```

Criterion-based procedures

```
mult.fit <- lm(log_losdays2 ~ is30dayreadmit + evisit+ cindex + ageyear + maritalstat
us + insurancetype + race + respirationrate + o2sat + heartrate + bmi + temperature +
bpsystolic + bpdiastolic + mews + icu_flag, data = hos_tidy_omitna)

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)
}

best_fit <- best(mult.fit, nbest = 1)
```

Final model

```
library(car)
```

```
## Warning: package 'car' was built under R version 3.4.3
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
## The following object is masked from 'package:boot':
##
##     logit
```

```
mult.fit2 <- lm(log_losdays2 ~ is30dayreadmit + evisit+ cindex + ageyear +  respirati
onrate + heartrate + temperature + bpsystolic, data = hos_tidy_omitna)

best_Fit2 <- best(mult.fit2, nbest = 1)

vif_multfit2 <- vif(mult.fit2)
```

Checking outliers

```
# # Simple linear regression
# reg_hos<-lm(hos_tidy_omitna$log_losdays2~hos_tidy_omitna$is30dayreadmit)


stu_res<-rstandard(mult.fit2)
outliers_y<-stu_res[abs(stu_res)>2.5]

#removing outliers
hos_tidy_omitna_outl <- hos_tidy_omitna[c(-6,-232,-277,-368,-411,-514,-535,-557,-562,
-604,-629,-704,-772,-824,-838,-852,-879,-982,-996,-1114,-1337,-1395,-1438,-1446, -147
1,-1491,-1517,-1552,-1605,-1639,-1682,-1697,-1882,-2002,-2024,-2071,-2153,-2395,-2460
,-2525,-2554,-2769,-2786,-2828,-2852, -2926, -3086, -3103, -3104, -3105, -3116, -3131
, -3170,-3174,-3232,-3298,-3299,-3318,-3329,-3332, -3405),]

checking_influence <- influence.measures(mult.fit2)
```

Vif

```
library(HH)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'lattice'
```

```
## The following object is masked from 'package:boot':
##
##     melanoma
```

```
## Loading required package: grid
```

```
## Loading required package: latticeExtra
```

```
## Loading required package: RColorBrewer
```

```
##
## Attaching package: 'latticeExtra'
```

```
## The following object is masked from 'package:ggplot2':
##
##     layer
```

```
## Loading required package: multcomp
```

```
## Loading required package: mvtnorm
```

```
## Loading required package: survival
```

```
##
## Attaching package: 'survival'
```

```
## The following object is masked from 'package:boot':
##
##     aml
```

```
## Loading required package: TH.data
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
##
## Attaching package: 'TH.data'
```

```
## The following object is masked from 'package:MASS':
##
##     geyser
```

```
## Loading required package: gridExtra
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
##
## Attaching package: 'HH'
```

```
## The following objects are masked from 'package:car':
##
##     logit, vif
```

```
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
## The following object is masked from 'package:boot':
##
##     logit
```

```
vif_multfit2 <- vif(mult.fit2)
```

Building the model without the outliers

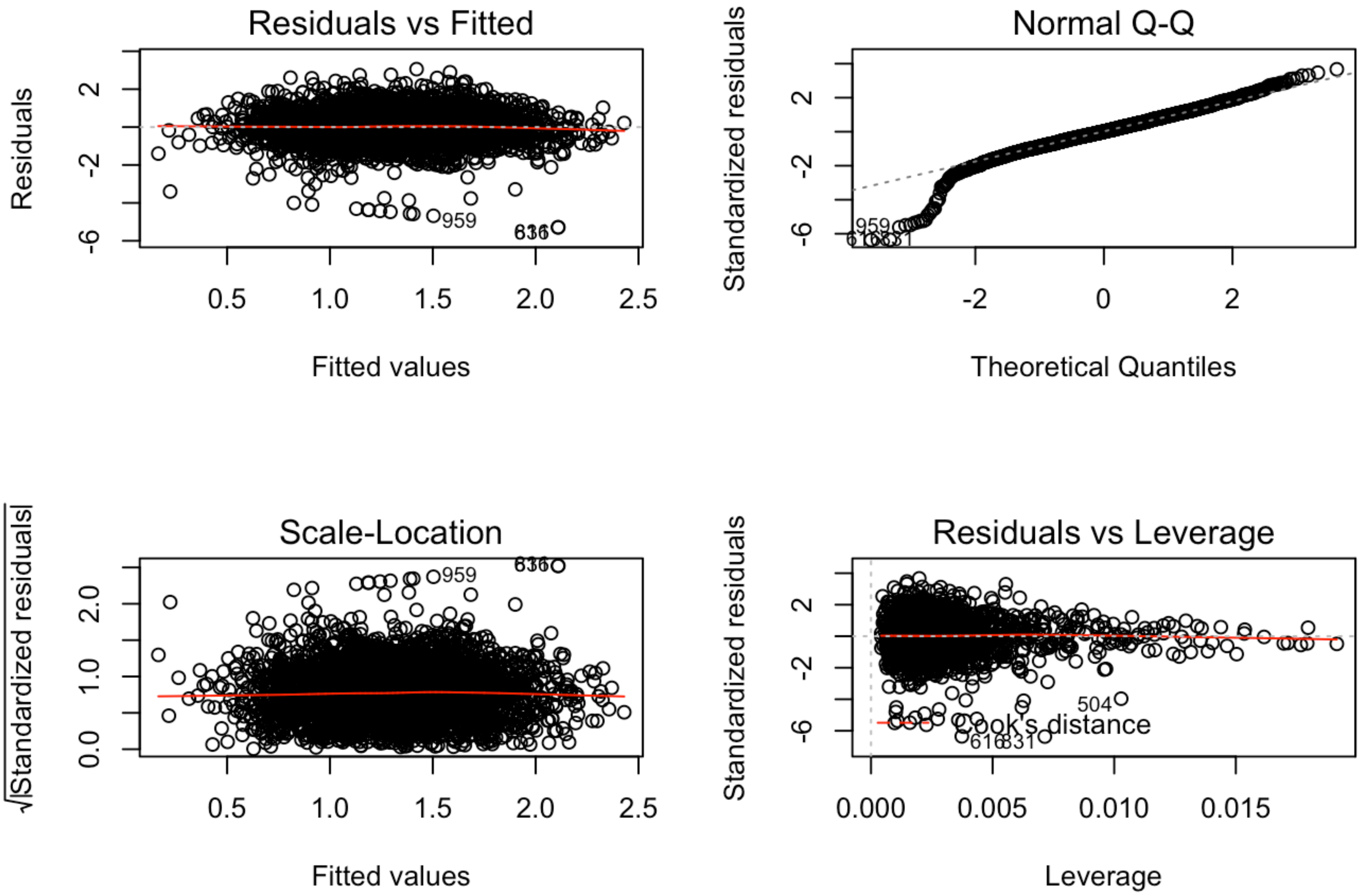
```
mult.fit3 <- lm(log_losdays2 ~ is30dayreadmit + evisit+ cindex + ageyear +  respirati
onrate + heartrate + temperature + bpsystolic, data = hos_tidy_omitna_outl)

best_fit3 <- best(mult.fit3, nbest = 1)
summary_multfit3 <-summary(mult.fit3)
```

Bootstrap

```
set.seed(1)


boot.fn<-function(data, index){
    return(coef(lm(log_losdays2 ~ is30dayreadmit + evisit+ cindex + ageyear +  respir
ationrate + heartrate  + temperature + bpsystolic , data = hos_tidy_omitna_outl, subs
et=index)))
}
boot_result <- boot.fn(hos_tidy_omitna_outl,1:3431)


set.seed(1)


boot_result2 <- boot.fn(hos_tidy_omitna_outl,sample(3431,3431,replace=T))



boot_result3 <- boot(hos_tidy_omitna_outl, boot.fn, 1000)

# How does it compare to the original (non-bootstrap) estimates?
summary_Boot_3<- summary(mult.fit3)
```

Residuals

```
par(mfrow=c(2,2))
plot_residuals <- plot(mult.fit2)
```


Residuals vs Fitted
Residuals
2
-2
-6
959
616
831
0.5
1.0
1.5
2.0
2.5
Fitted values
Normal Q-Q
Standardized residuals
2
-2
-6
959
616
831
-2
0
2
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
0.0
1.0
2.0
959
616
831
0.5
1.0
1.5
2.0
2.5
Fitted values
Residuals vs Leverage
Standardized residuals
2
-2
-6
504
Cook's distance
616
831
0.000
0.005
0.010
0.015
Leverage