Biostatistics Methods 1 Final Project

Manqi Cai, Yixi Xu, Peng-Lin Lin, Kaitlin Maciejewski

12/9/2017

The Data Analytics group from Good Health Corporation are interested in improving the overall hospital management and minimizing the cost/resources associated with patients' care. One of the most important outcomes that has a direct effect on these aspects is patient's length of stay (LoS) in the hospital. Thus, they would like to know which variables are associated with LoS, and ultimately build a predictive model to be used for future visits. The group has contacted you to study this problem and make a recommendation.

```
ghproject <- read_excel("GHProject_Dataset.xlsx") %>%
  clean_names()

#summary of continuous variable

attach(ghproject)

sd(loshours, na.rm = T)
```

```
## [1] 142.3524
```

```
sd(losdays2, na.rm = T)
```

```
## [1] 5.931351
```

```
sd(ageyear, na.rm = T)
```

```
## [1] 18.66037
```

```
sd(bmi, na.rm = T)
```

```
## [1] 7.961884
```

```
sd(bpdiastolic, na.rm = T)
```

```
## [1] 9.828815
```

```
sd(o2sat, na.rm = T)
```

```
## [1] 4.864867
```

```
sd(temperature, na.rm = T)
```

```
## [1] 0.9084958
```

```
sd(heartrate, na.rm = T)
```

```
## [1] 12.97154
```

```
sd(respirationrate, na.rm = T)
```

```
## [1] 2.6465
```

```
sd(bpsystolic, na.rm = T)
```

```
## [1] 16.77863
```

```
#summary of categorical data
ghproject %>%
```

```
  group_by(mews) %>%
  summarize(n())
```

```
## # A tibble: 15 x 2
##     mews `n()`
##    <dbl> <int>
##  1     0    39
##  2     1   618
##  3     2   730
##  4     3   801
##  5     4   711
##  6     5   328
##  7     6   151
##  8     7    87
##  9     8    26
## 10     9    14
## 11    10     5
## 12    11     4
## 13    12     1
## 14    14     1
## 15    NA   166
```

```
ghproject %>%
  group_by(cindex) %>%
  summarize(n())
```

```
## # A tibble: 5 x 2
##   cindex `n()`
##    <dbl> <int>
## 1      0  1207
## 2      1   636
## 3      2   722
## 4      3   435
## 5      5   682
```

```
ghproject %>%
  group_by(evisit) %>%
  summarize(n())
```

```
## # A tibble: 5 x 2
##   evisit `n()`
##    <dbl> <int>
## 1      0  1134
## 2      1   752
## 3      2   521
## 4      3   329
## 5      4   946
```

```
ghproject %>%
  group_by(icu_flag) %>%
  summarize(n())
```

```
## # A tibble: 2 x 2
##   icu_flag `n()`
##      <dbl> <int>
## 1        0  3618
```

```
## 2        1    64
```

```
ghproject %>%
  group_by(gender) %>%
  summarize(n())
```

```
## # A tibble: 2 x 2
##   gender `n()`
##    <chr> <int>
## 1 Female  1981
## 2   Male  1701
```

```
ghproject %>%
  group_by(race) %>%
  summarize(n())
```

```
## # A tibble: 6 x 2
##                   race `n()`
##                  <chr> <int>
## 1   African Amer/Black   788
## 2                Asian   253
## 3  Native Amer/Alaskan    22
## 4 Natv Hawaii/Pacf Isl     4
## 5    Other/Multiracial   526
## 6                White  2089
```

```
ghproject %>%
  group_by(religion) %>%
  summarize(n())
```

```
## # A tibble: 11 x 2
##              religion `n()`
##                 <chr> <int>
##  1          Angelican     1
##  2           Catholic  1678
##  3          Christian   909
##  4             Hebrew     1
##  5              Hindu   127
##  6              Islam   112
##  7             Jewish   527
##  8             Mormon     2
##  9     No Affiliation   177
## 10 Non Denominational     1
## 11              Other   147
```

```
ghproject %>%
  group_by(maritalstatus) %>%
  summarize(n())
```

```
## # A tibble: 7 x 2
##   maritalstatus `n()`
##           <chr> <int>
## 1   Civil Union     1
## 2      Divorced   241
## 3       Married  1637
## 4     Separated    52
## 5        Single   967
```

```
## 6       Widowed   704
## 7          <NA>    80
```

```
ghproject %>%
  group_by(facilityname) %>%
  summarize(n())
```

```
## # A tibble: 8 x 2
##                  facilityname `n()`
##                         <chr> <int>
## 1         Lenox Hill Hospital     5
## 2            LIJ Forest Hills   568
## 3           LIJ Valley Stream   331
## 4 Long Island Jewish Hospital   828
## 5                        NSUH  1011
## 6          Plainview Hospital   328
## 7          Southside Hospital   528
## 8            Syosset Hospital    83
```

```
ghproject %>%
  group_by(insurancetype) %>%
  summarize(n())
```

```
## # A tibble: 4 x 2
##   insurancetype `n()`
##           <chr> <int>
## 1      Medicaid   170
## 2      Medicare  1457
## 3       Private  2021
## 4          <NA>    34
```

```
#summary of PatientID and VisitID
ghproject %>%
  distinct(patientid) %>%
  count()
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1  3612
```

```
ghproject %>%
  distinct(visitid) %>%
  count()
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1  3682
```

There are 3682 records from 3612 patients

Table 1: Summary of Numerical Patient Data

	Min	Max	Mean	Median	Sd	Missing
Length of Stay, hours	1	2111	131.8	92.0	142.35	
Length of Stay, days	0.04	87.96	5.49	3.83	5.93	
Age	18.00	105	65.74	68.00	18.66	
BMI	3.10	122.65	28.33	27.10	7.96	697
BP Systolic mmHg	88.78	193.96	130.52	129.17	9.83	5
O2 Saturation %	80	236.53	97.85	97.58	4.86	3
Temperature °C	11.85	52.27	36.73	36.73	0.91	3
Heart Rate bpm	37.58	242.58	80.09	79.20	12.97	5
Respiration Rate bpm	12	67.72	18.20	17.76	2.65	3
BP Diastolic mmHg	29.56	154.4	72.53	71.85	16.78	1

SAS output:

Forward

significant variables:

ageyear, evisit, bpsystolic, cindex, heartrate, is30dayreadmit, respirationrate

AdjR2: 0.1169

AIC: 1743.8

Backward

significant variables:

month, mews, icu_flag, temperature, bmi, o2sat, religion, gender, bpdiastolic, maritalstatus, insurancetype

AdjR2: 0.1169

AIC: 1743.8

Stepwise

significant variables:

ageyear, evisit, bpsystolic, cindex, heartrate, is30dayreadmit, respirationrate

AdjR2: 0.1169

AIC: 1743.8

All models agree

```
#identifies which observations to remove
stu_res <- rstandard(reg)
outlier <- stu_res[abs(stu_res) > 2.5]

is.outlier = function(value, vector){
  value %in% outlier
```

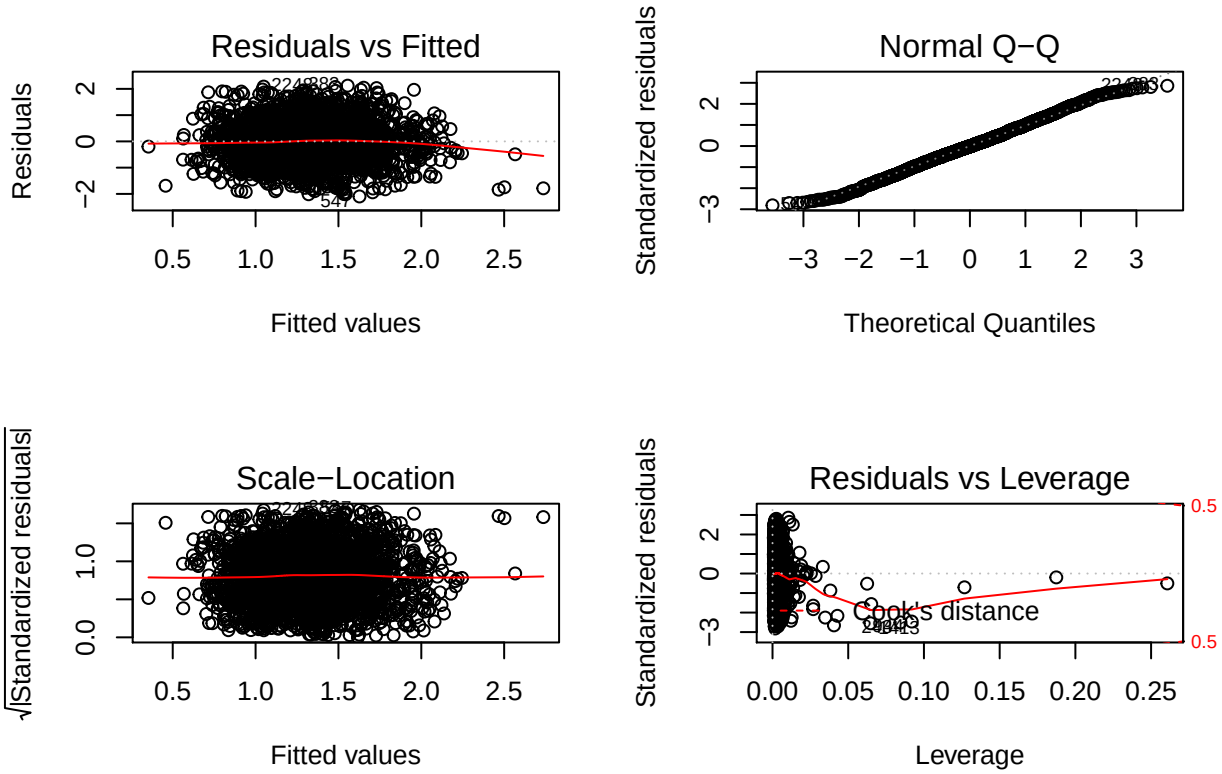
```
}

data_no_outliers = ghproject_tidy %>%
  mutate(outlier = is.outlier(stu_res, outlier))%>%
  filter(outlier == FALSE)

#refit model with no outliers
reg_no_outliers <- lm(losdays2_log ~ is30dayreadmit + ageyear +  cindex  + insurancetype + bpsystolic +

summary(reg_no_outliers)
```

```
##
## Call:
## lm(formula = losdays2_log ~ is30dayreadmit + ageyear + cindex +
##     insurancetype + bpsystolic + o2sat + heartrate + respirationrate,
##     data = data_no_outliers)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.09710 -0.48366 -0.00958  0.48504  2.12177
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.7095302  0.3274788   2.167 0.030351 *
## is30dayreadmit   0.2674520  0.0419803   6.371 2.21e-10 ***
## ageyear          0.0096754  0.0008669  11.161  < 2e-16 ***
## cindex           0.1271358  0.0198810   6.395 1.89e-10 ***
## insurancetype    0.0961222  0.0256676   3.745 0.000184 ***
## bpsystolic      -0.0059429  0.0009011  -6.595 5.10e-11 ***
## o2sat           -0.0032183  0.0027216  -1.182 0.237121
## heartrate        0.0053301  0.0011077   4.812 1.58e-06 ***
## respirationrate  0.0212500  0.0053352   3.983 6.99e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7458 on 2657 degrees of freedom
## Multiple R-squared:  0.1341, Adjusted R-squared:  0.1315
## F-statistic: 51.42 on 8 and 2657 DF,  p-value: < 2.2e-16
```

```
confint(reg_no_outliers)
```

```
##                       2.5 %       97.5 %
## (Intercept)      0.067391071  1.351669387
## is30dayreadmit   0.185134664  0.349769276
## ageyear          0.007975497  0.011375263
## cindex           0.088152003  0.166119534
## insurancetype    0.045791699  0.146452626
## bpsystolic      -0.007709847 -0.004176027
## o2sat           -0.008555028  0.002118464
## heartrate        0.003158059  0.007502067
## respirationrate  0.010788480  0.031711433
```

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



The MSE we get from bootstrap model is 0.7430972, and MSE in our model is 0.7458, the difference between these two MSE is about 0.8%, which is a lot less than the threshold 10%. Therefore, after model validation, we think our model is reasonalbe for predict LOS.

```
set.seed(1)
##cross-validation method
# load the library
library(caret)
# define training control
train_control <- trainControl(method="cv", number=10)
# fix the parameters of the algorithm
grid <- expand.grid(.fL=c(0), .usekernel=c(FALSE))
# train the model
model <- train(losdays2_log ~ + is30dayreadmit + ageyear + cindex + evisit + gender + maritalstatus + ir
# summarize results
print(model)
```

```
## Linear Regression
##
## 2666 samples
##   12 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 2400, 2399, 2399, 2400, 2399, 2398, ...
## Resampling results:
##
##   RMSE       Rsquared   MAE
##   0.7398153  0.1465013  0.5865414
##
```

```
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
#SLR for each variable
SLR_is30dayreadmit <- lm(ghproject_tidy$losdays2 ~ ghproject_tidy$is30dayreadmit)
#confint(SLR_is30dayreadmit)
#coef(summary(SLR_is30dayreadmit))
row_is30dayreadmit <- cbind(coef(summary(SLR_is30dayreadmit))[, c(1:2, 4)], confint(SLR_is30dayreadmit)


SLR_cindex <- lm(ghproject_tidy$losdays2 ~ ghproject_tidy$cindex)
row_cindex <- cbind(coef(summary(SLR_cindex))[, c(1:2, 4)], confint(SLR_cindex))[-1, ]


SLR_evisit <- lm(ghproject_tidy$losdays2 ~ ghproject_tidy$evisit)
row_evisit <- cbind(coef(summary(SLR_evisit))[, c(1:2, 4)], confint(SLR_evisit))[-1, ]


SLR_ageyear <- lm(ghproject_tidy$losdays2 ~ ghproject_tidy$ageyear)
row_ageyear <- cbind(coef(summary(SLR_ageyear))[, c(1:2, 4)], confint(SLR_ageyear))[-1, ]


SLR_gender <- lm(ghproject_tidy$losdays2 ~ ghproject_tidy$gender)
row_gender <- cbind(coef(summary(SLR_gender))[, c(1:2, 4)], confint(SLR_gender))[-1, ]


SLR_maritalstatus <- lm(ghproject_tidy$losdays2 ~ ghproject_tidy$maritalstatus)
row_maritalstatus <- cbind(coef(summary(SLR_maritalstatus))[, c(1:2, 4)], confint(SLR_maritalstatus))[-

SLR_insurancetype <- lm(ghproject_tidy$losdays2 ~ ghproject_tidy$insurancetype)
row_insurancetype <- cbind(coef(summary(SLR_insurancetype))[, c(1:2, 4)], confint(SLR_insurancetype))[-

SLR_bpsystolic <- lm(ghproject_tidy$losdays2 ~ ghproject_tidy$bpsystolic)
row_bpsystolic <- cbind(coef(summary(SLR_bpsystolic))[, c(1:2, 4)], confint(SLR_bpsystolic))[-1, ]


SLR_o2sat <- lm(ghproject_tidy$losdays2 ~ ghproject_tidy$o2sat)
row_o2sat <- cbind(coef(summary(SLR_o2sat))[, c(1:2, 4)], confint(SLR_o2sat))[-1, ]


SLR_heartrate <- lm(ghproject_tidy$losdays2 ~ ghproject_tidy$heartrate)
row_heartrate <- cbind(coef(summary(SLR_heartrate))[, c(1:2, 4)], confint(SLR_heartrate))[-1, ]


SLR_respirationrate <- lm(ghproject_tidy$losdays2 ~ ghproject_tidy$respirationrate)
row_respirationrate <- cbind(coef(summary(SLR_respirationrate))[, c(1:2, 4)], confint(SLR_respirationra


SLR_bpdiastolic <- lm(ghproject_tidy$losdays2 ~ ghproject_tidy$bpdiastolic)
bpdiastolic <- cbind(coef(summary(SLR_bpdiastolic))[, c(1:2, 4)], confint(SLR_bpdiastolic))[-1, ]

SLR_res <- rbind(row_is30dayreadmit, row_cindex, row_evisit, row_ageyear, row_gender, row_maritalstatus


#MLR
```

```r
MLR_ghprojrct <- lm(ghproject_tidy$losdays2 ~ ghproject_tidy$is30dayreadmit + ghproject_tidy$cindex + gh
#summary(MLR_ghprojrct)
MLR_coef <- coef(summary(MLR_ghprojrct))[,c(1:2,4)]
MLR_conf<- confint(MLR_ghprojrct)
MLR_res <- cbind(MLR_coef, MLR_conf)[-1,]

cbind(round(SLR_res,digits = 3), round(MLR_res, digits = 3))
```

```
##                     Estimate Std. Error Pr(>|t|)  2.5 % 97.5 % Estimate
## row_is30dayreadmit     1.895      0.321    0.000  1.266  2.524    0.991
## row_cindex             1.321      0.145    0.000  1.036  1.606    0.853
## row_evisit             0.478      0.071    0.000  0.339  0.616    0.282
## row_ageyear            0.038      0.006    0.000  0.026  0.049    0.029
## row_gender             0.507      0.225    0.024  0.066  0.947    0.808
## row_maritalstatus     -0.401      0.225    0.075 -0.842  0.040   -0.324
## row_insurancetype      1.056      0.189    0.000  0.685  1.426    0.508
## row_bpsystolic        -0.026      0.007    0.000 -0.039 -0.012   -0.022
## row_o2sat             -0.038      0.021    0.070 -0.080  0.003   -0.026
## row_heartrate          0.043      0.008    0.000  0.026  0.059    0.040
## row_respirationrate    0.221      0.040    0.000  0.142  0.301    0.154
## bpdiastolic           -0.059      0.011    0.000 -0.081 -0.037   -0.033
##                     Std. Error Pr(>|t|)  2.5 % 97.5 %
## row_is30dayreadmit       0.326    0.002  0.352  1.629
## row_cindex               0.150    0.000  0.560  1.147
## row_evisit               0.072    0.000  0.140  0.424
## row_ageyear              0.007    0.000  0.016  0.043
## row_gender               0.224    0.000  0.369  1.246
## row_maritalstatus        0.222    0.145 -0.760  0.112
## row_insurancetype        0.192    0.008  0.131  0.885
## row_bpsystolic           0.008    0.005 -0.038 -0.007
## row_o2sat                0.020    0.206 -0.066  0.014
## row_heartrate            0.008    0.000  0.024  0.057
## row_respirationrate      0.040    0.000  0.077  0.232
## bpdiastolic              0.013    0.014 -0.059 -0.007
```