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.

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	88.78	193.96	130.52	129.17	9.83	5
O2 Saturation	80	236.53	97.85	97.58	4.86	3
Temperature	11.85	52.27	36.73	36.73	0.91	3
Heart Rate	37.58	242.58	80.09	79.20	12.97	5
Respiration Rate	12	67.72	18.20	17.76	2.65	3
BP Diastolic	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

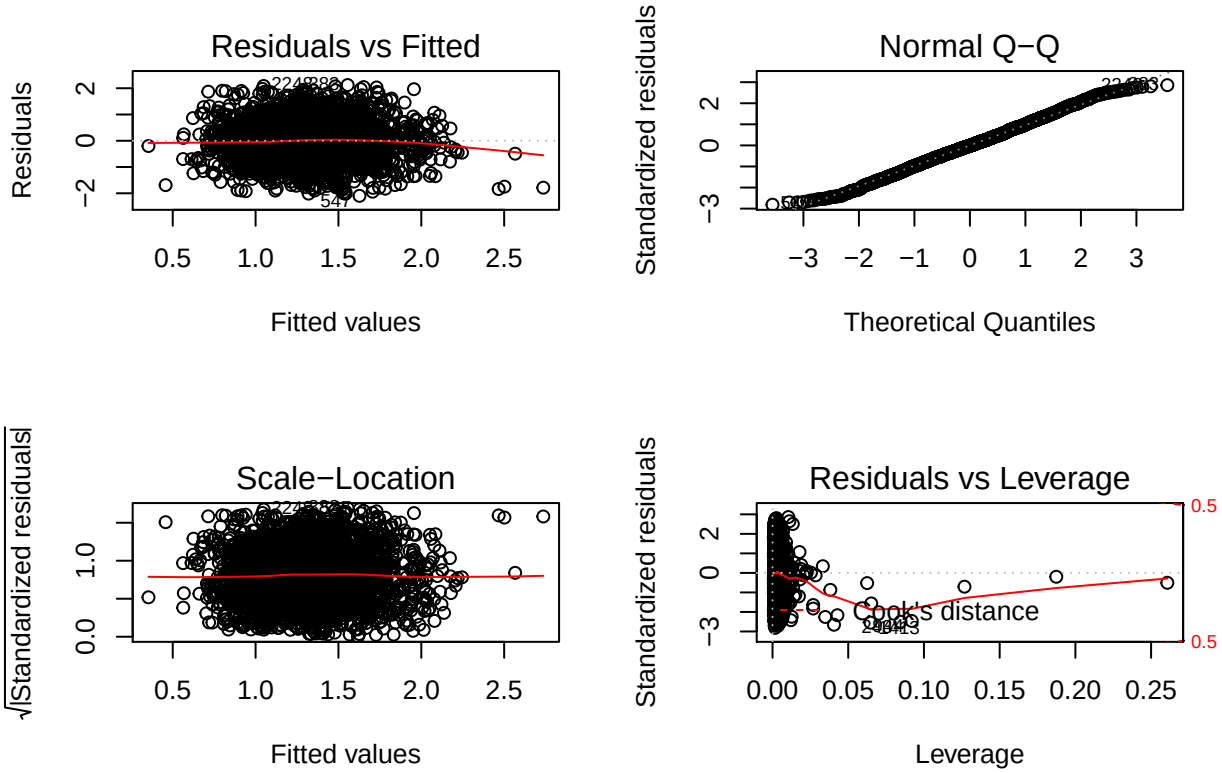
```
#identifies which observations to remove
stu_res <- rstandard(reg)
outlier <- stu_res[abs(stu_res) > 2.5]

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 0.832, and MSE in our model is 0.831, the difference between these two MSE is about 0.1%, which is a lot less than 10%. Therefore, after model validation, we think our model is reasonalbe for predict LOS.