Flu Vaccine Project

Tom DeNatale

Sunday, May 24, 2015

Title: WIth the growing penetration of Flu Vaccination in the United States of America

What is the effect on the Length of Stay in Hospitaization, it may be confounding!

References:

1. Flu Vaccination Coverage, United States, 2013-2014 Influenza Season, last updated September 18,2014, content source: http://www.cdc.gov/flu/fluvaxview/coverage-1314estimates.htm
2. Flu Vaccination Coverage by High-Risk Conditions, Adults, BRFSS http://www.cdc.gov/flu/fluvaxview/brfss-high-risk.htm
3. Reducing major cause-specific hospitalization rates and shortening hospital stays after influenza vaccination. Wang CS1, Wang ST, Lai CT, Lin LJ, Lee CT, Chou P. http://www.ncbi.nlm.nih.gov/pubmed/15578359

Background:

The impact of influenza vaccination has intrigued me for many years. For 2 years, I have helped coordinate a flu vaccination clinic for the poor at a San Jose, California nonprofit. The vaccination is free but it is offerred to adults, as children have another avenue to be vaccinated. We have vaccinated about 100 adults each year that otherwise would not have the opportunity. This year we are planning on a target of 200 people to be vaccinated.

The recent resistance to childhood measle vaccinations rekindled my interests. There are many confounding relationships with vaccinations. One of my current classmates is against having herself and her family participate in flu vaccination.

Well speaking of confounding that is the purpose of my subject for the project: For people that are hospitalized those that have received the flu vaccine are more likely to have a greater length of stay than those that don't.

It is also interesting to see the effect of flu vaccination to high risk patients. The data from the article under the Taiwan health care system was enlightening. In the US article the quoted from the Healthy People 2020 targets are set at 70% which is currently under target. It is also noted that some of the data was not taken consistently, and statistical methods were used to minimize the loss of data. One thing for sure there is a wealth of data out there about flu vaccinations.

Question to the investigator: Why does the patient in the hospital, with the records indicating that the patient was immunized with the influenza vaccination (flu vaccine or flu shot) before they went to the hospital, increase the length of stay in the hospital?

General approach:

- The length of hospital stay is really caused by how ill or sick the patients happen to be.
- The likelihood of being very sick is strongly influenced by prior conditions.
- The likelihood of obtaining a flu vaccination is about twice as great for patients with prior conditions.
- The likelihood of prior conditions increases with age.
- I felt that since all the variables are binary (TRUE or FALSE), I would build a logistical regression model to solve the problem. After creating the data set the professor suggested that I plot the data. After trying to squeeze some results into a logistical regression methodology, I went back and looked at the plots embellishing them with color.
- It is quite clear there are really two clusters of data, ones with patients with the flu vaccination and ones without.
- In hindsight it was also clear how I built the model data that the linear equation was based on almost all the

variables (because they were defined to be independent)
- Separating the data into two clusters and modeling both clusters using a linear model gave ballpark results, with very large residuals or uncertainties. Although this model was not very accurate it might provide insight for a clinic to plan on how much resources in terms of staffing and room availability are needed. It also had a surprising result!
- Variations in the amount and kinds of diseases provides variability, leading to a large residual errors for the simplest models based on age or the presence of a patient that had received a flu shot before entering the hospital.
- So next I tried a linear model based on all the independent variables. This gave very good results
- I descibe the data and rationale interpersed with code and plots below to describe the detail, including the data initiation.

Conclusion:

- “He ain’t heavy he’s my brother”, well that is not it!
- The confounding conclusion is that even though given limited data such as a chart that shows Length of Stay(LOS) increasing as a function of a patient that has had a flu vaccination, versus one that didn’t: the LOS is really a function on how sick a patient really is.
- Given that a sick patient is usually one that is older and under care it is more likely the sicker patient will have a flu shot versus someone that comes into a hospital without prior chronic diseases.
- It goes to show that one needs to know the data, in order to avoid hidden pitfalls, or end up confused or confounded.

Creation of the data set and associated code:

- For this project originally I was going to use the manipulate function to select coefficients for variables. However, this was not necessary, and although I did get the manipulate function to vary results not quite to a normal distribution using 3 parameters, the results did not bear keeping in this project paper. All that is left is the library call.
- In general running models at N = 1000 or N = 1e3 was sufficient to get reasonable results.
- Create data where LOS is greater if not vaccinated based on paper(3) from Taiwan for high risk patients.
- Created data from US paper(2) on rate of vaccination for high risk patient sets.
- Note although the data is reasonable, the data was built as if all results are linear, which is unlikely to be true in real life scenarios
- The following table are the approximate probabilities of patients that enter the hospital with one or more of these high risk conditions:
- COPD | 0.05 | 0.06 | adj to original population
- ASTH | 0.02 | 0.04 | including more age groups
- HRTD | 0.08 | 0.09 |
- DIAB | 0.01 | 0.04 | estimate rise of this disease
- CANC | 0.08 | 0.09 |
- Note: Model data includes rates of disease in hospitalization and rates of obtaining flu vaccination prior to entering the hospital ### The following code chunk creates the model data
- Note that FLUSHOT if true indicates one had the flushot before entering the hospital
- Note that the dependent variable LOWR indicating a low risk patient without any of the high risk maladies has been generated, but could be easily removed.
- The Logistic function was used to fit the input data to hopefully a more realistic base dealing with primarily the older population
- Finally a standard length of stay for each malady is provided.

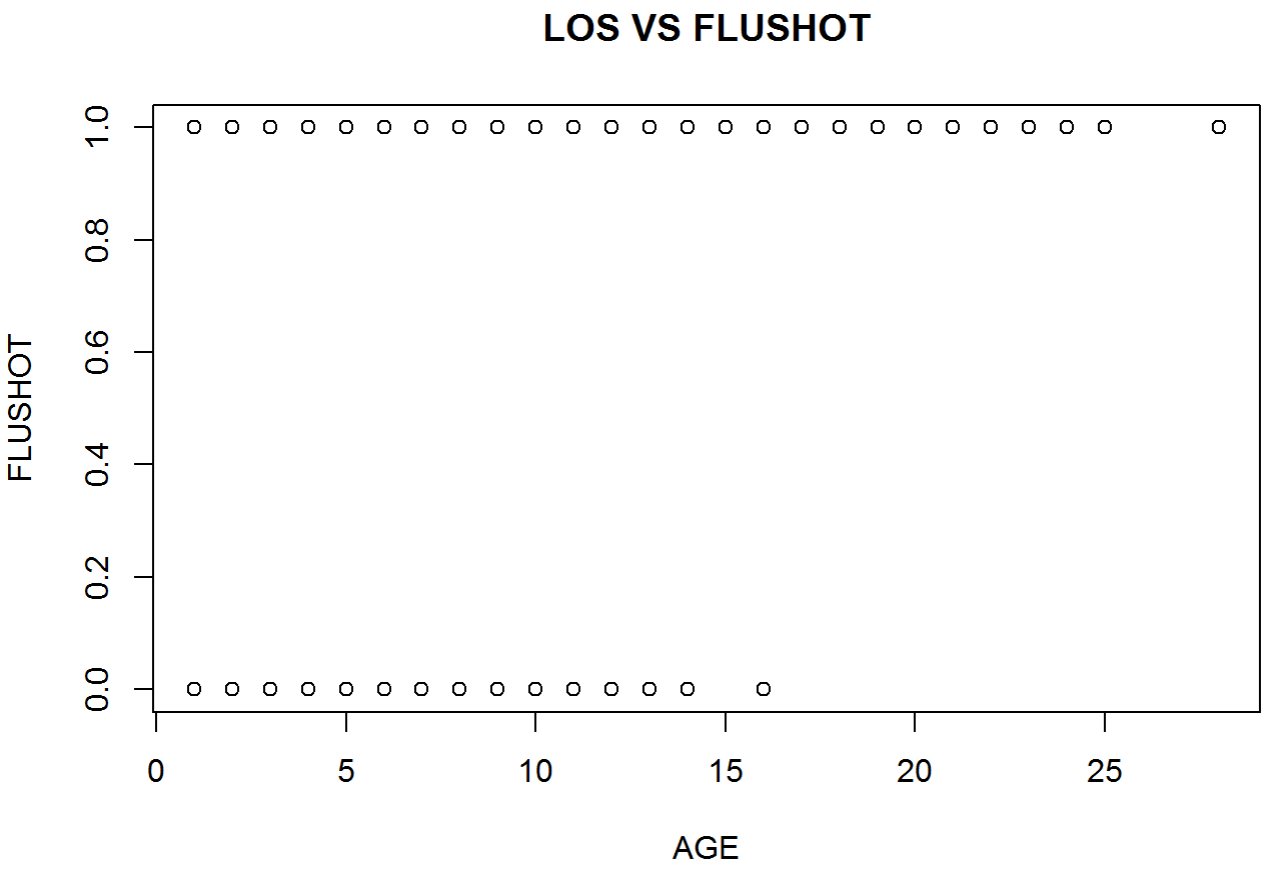
```
library(manipulate)
```

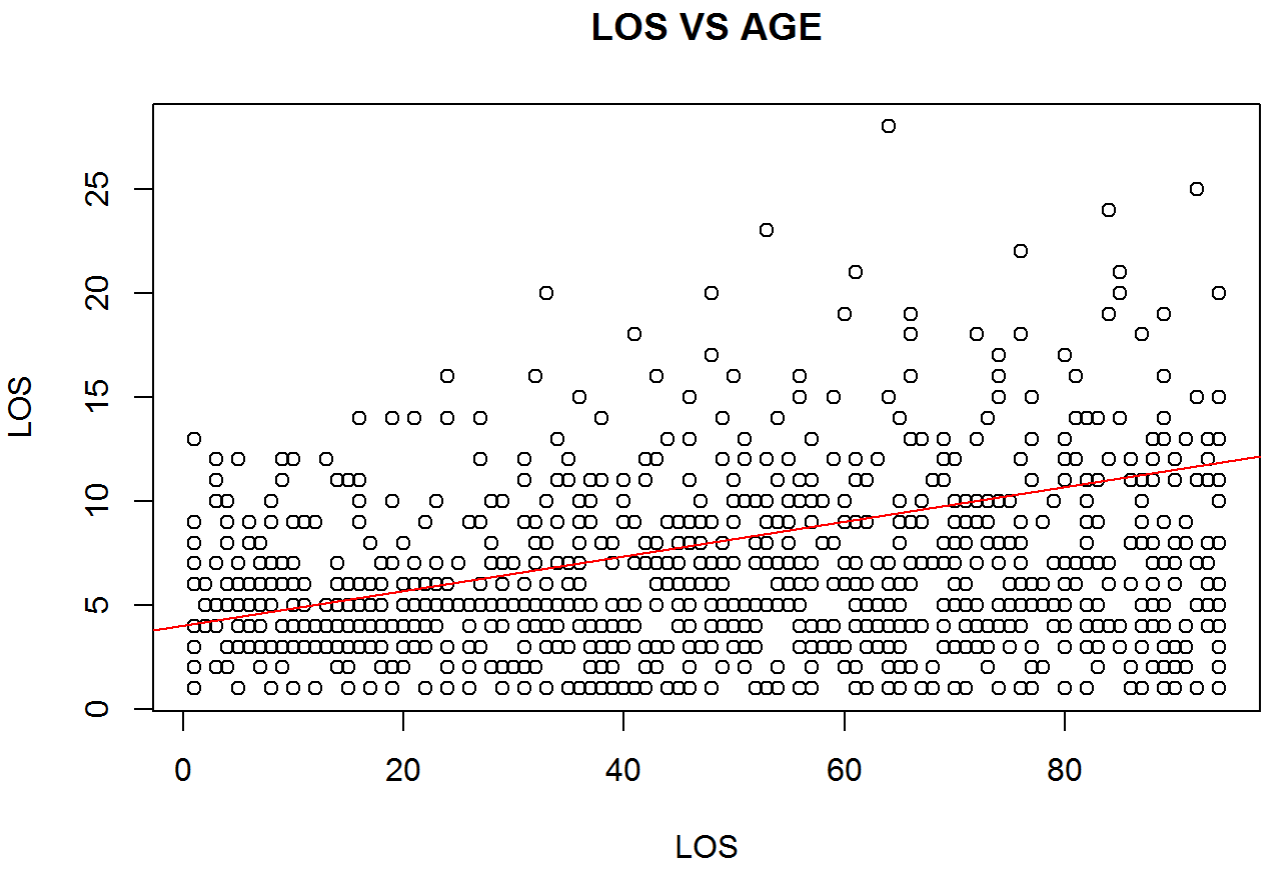
```
N <- 1e3
logistic <- function(t) 1 / (1 + exp(-t))


simLOS <- function(N) {
    AGE  <- sample(1:94, N, replace=T)
    COPD <- runif(N)< .12*logistic((AGE-50)/10)
    ASTH <- runif(N)< .095
    HRTD <- runif(N)< .19*logistic((AGE-50)/10)
    DIAB <- runif(N)< .08*logistic((AGE-50)/10)
    CANC <- runif(N)< .19*logistic((AGE-40)/10)
    LOWR <- !(COPD+ASTH+HRTD+DIAB+CANC)
    FLUSHOT <- runif(N)< .34 + .5*(COPD+ASTH+HRTD+DIAB+CANC)
    pflu  <- ifelse(FLUSHOT,.05,.20)
    FLU  <- runif(N)<pflu
    LOS <- 4 + COPD*8 + ASTH*6 +HRTD*6 + DIAB*4 + CANC*5 +FLU*2 +round(rnorm(N,sd=2,mean=0),di
gits=0)
    LOS[LOS<1]<- 1
    data.frame(AGE,COPD,ASTH,HRTD,DIAB,CANC,LOWR,FLUSHOT,FLU,LOS)
}
datafr<- (simLOS(N))


head(datafr)
```

```
##   AGE  COPD  ASTH  HRTD  DIAB  CANC  LOWR FLUSHOT   FLU LOS
## 1   3 FALSE FALSE FALSE FALSE FALSE  TRUE    TRUE FALSE  11
## 2  50 FALSE FALSE FALSE FALSE  TRUE FALSE    TRUE FALSE   9
## 3  50 FALSE FALSE FALSE FALSE FALSE  TRUE   FALSE FALSE   4
## 4  68 FALSE FALSE FALSE FALSE  TRUE FALSE    TRUE FALSE  11
## 5   5 FALSE FALSE FALSE FALSE FALSE  TRUE   FALSE FALSE   1
## 6  16 FALSE FALSE FALSE FALSE FALSE  TRUE   FALSE FALSE   4
```

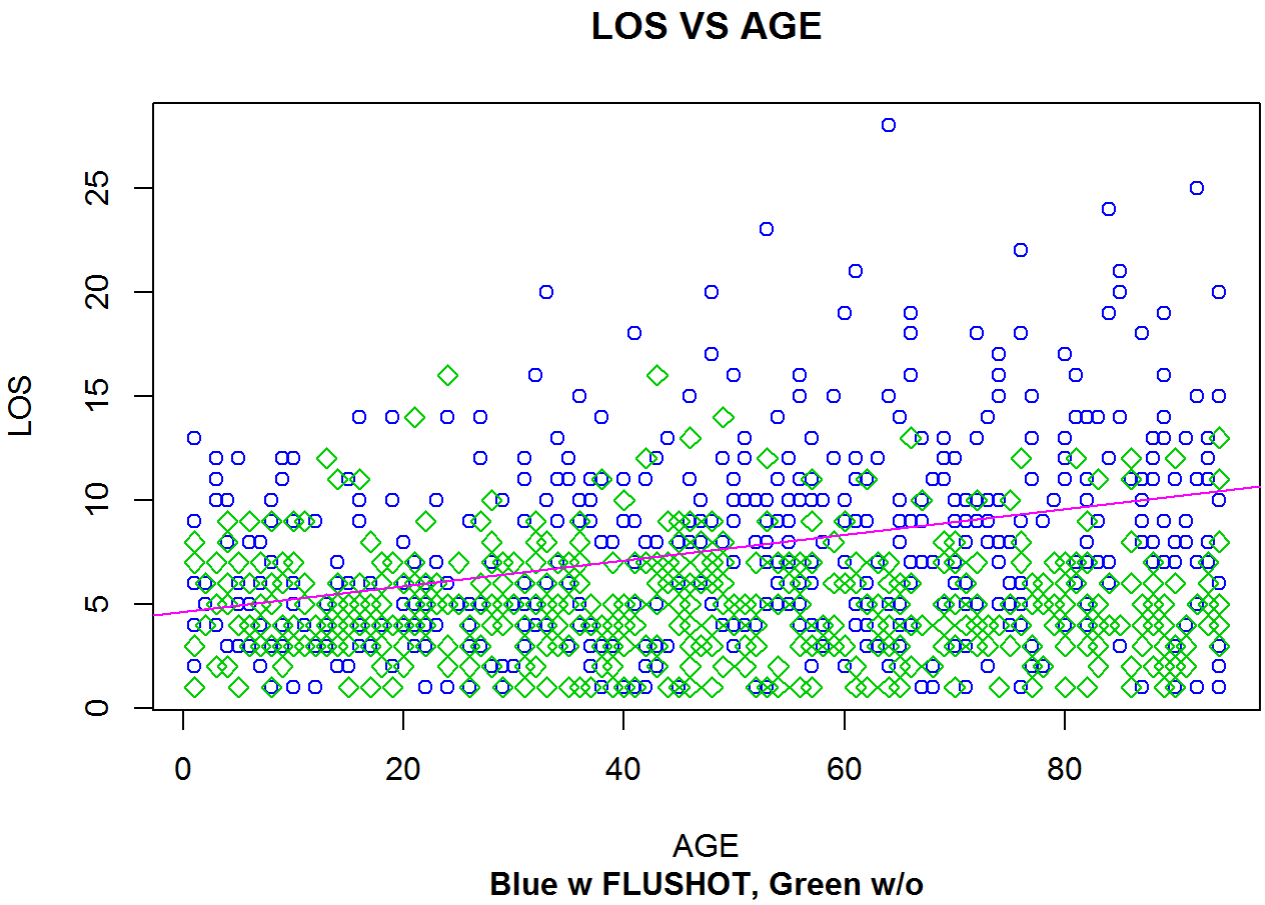
The data above is the heasder or sample that is contained in the dataframe

Let’s plot the data with several views


LOS VS FLUSHOT
FLUSHOT
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
AGE


LOS VS AGE
LOS
0
5
10
15
20
25
0
20
40
60
80
LOS

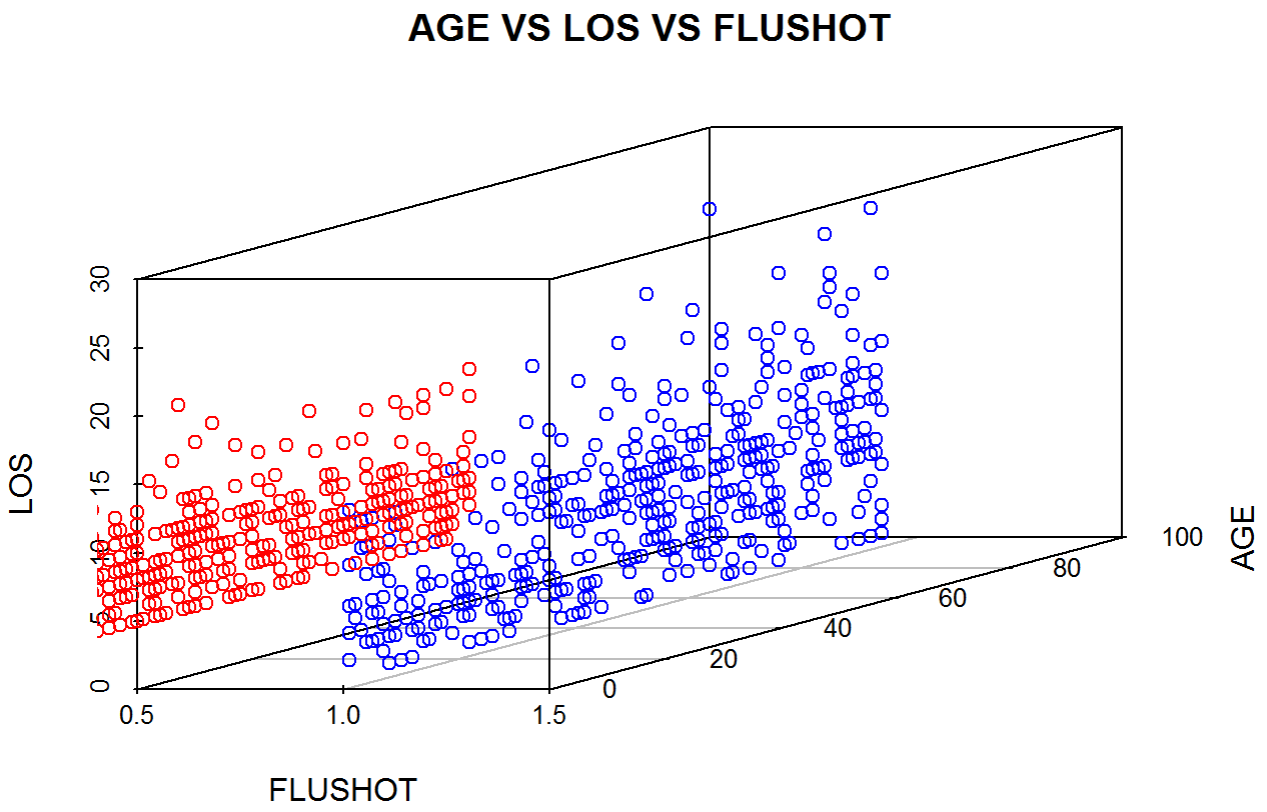
```
#Plot AGE VS LOS with FLUSHOT vs NOFLUSHOT
plot(datafr$AGE[datafr$FLUSHOT==1],datafr$LOS[datafr$FLUSHOT==1],col=4, xlab= "AGE", ylab= "LOS",
     main="LOS VS AGE",sub="Blue w FLUSHOT, Green w/o",font.sub=2)

points(datafr$AGE[datafr$FLUSHOT==0],datafr$LOS[datafr$FLUSHOT==0],col=3,pch = 5)
abline(a=4.64,b=.0615,col=6)
```




```
# 3D Scatterplot
library(scatterplot3d)
```

```
## Warning: package 'scatterplot3d' was built under R version 3.1.3
```

```
#Note reversed labels for xyz to xzy because it was clearerAGE sb the z axis and LOS the y axis
s3d <-scatterplot3d(datafr$FLUSHOT[datafr$FLUSHOT==1],datafr$AGE[datafr$FLUSHOT==1],datafr$LOS[datafr$FLUSHOT==1],
                    color = "blue",
              xlab= "FLUSHOT",ylab= "AGE",zlab= "LOS",
              main="AGE VS LOS VS FLUSHOT")
s3d$points3d(datafr$FLUSHOT[datafr$FLUSHOT==0],datafr$AGE[datafr$FLUSHOT==0],datafr$LOS[datafr$FLUSHOT==0],
              col = "red")
```



Wow what was that?

Each plot delivers the story that there appears to be a relationship between having a flu shot and Length of Stay (LOS)

- Plot 1 shows clearly that for some instances LOS is greater for events where a flu shot had been given
- Plot 2 shows that there appears to an effect of AGE on instances for LOS is larger. It also shows a line that could be fit to the data, but with a large variance to contain the points
- plot 3 once again shows a relationship between FLUSHOT and LOS
- Plot 4 shows this in “3D”

So let’s fit some models and review the results

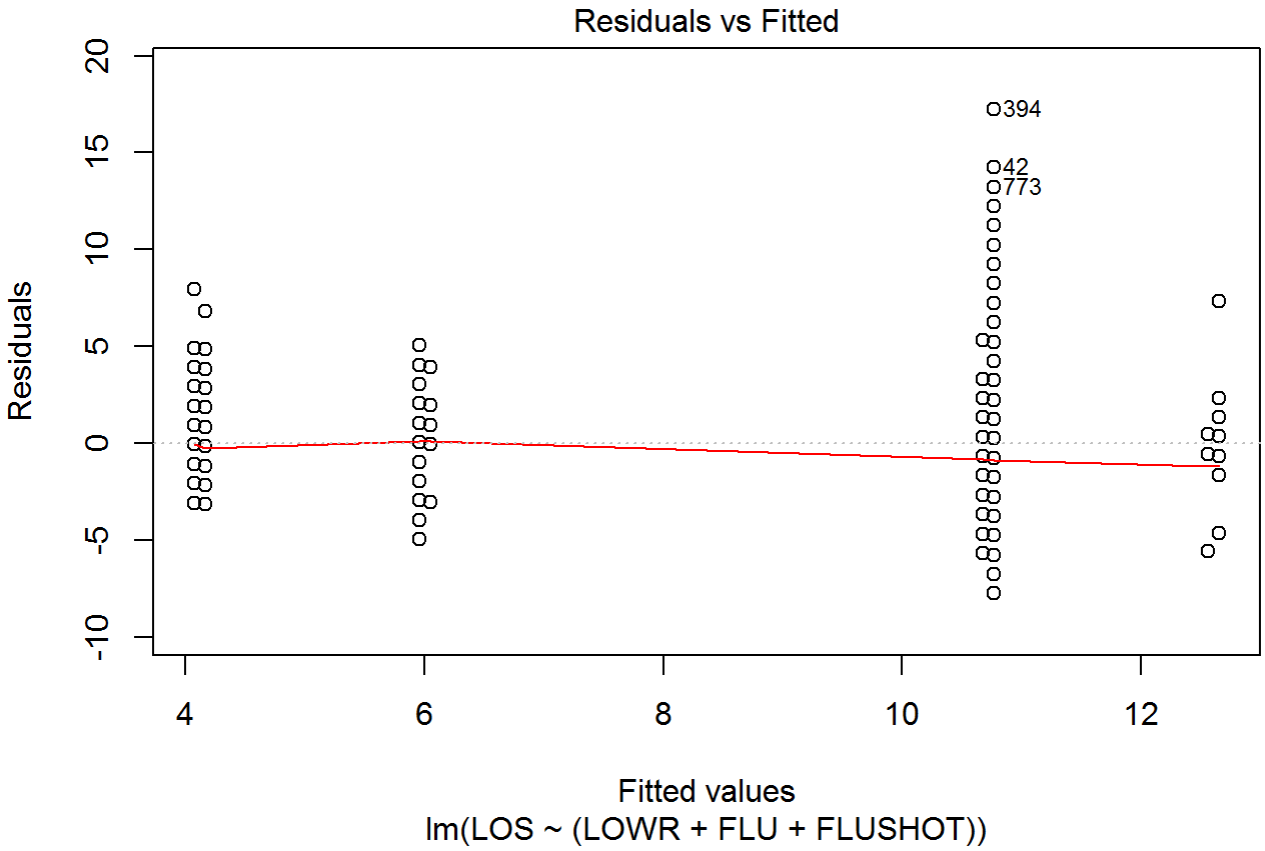
- Someones first attempt may be very basic, let’s include FLUSHOT as a parameter of interest
- Results are not very good, unless you want to know that the average stay of all patients is about 11 days with a large residual or variance

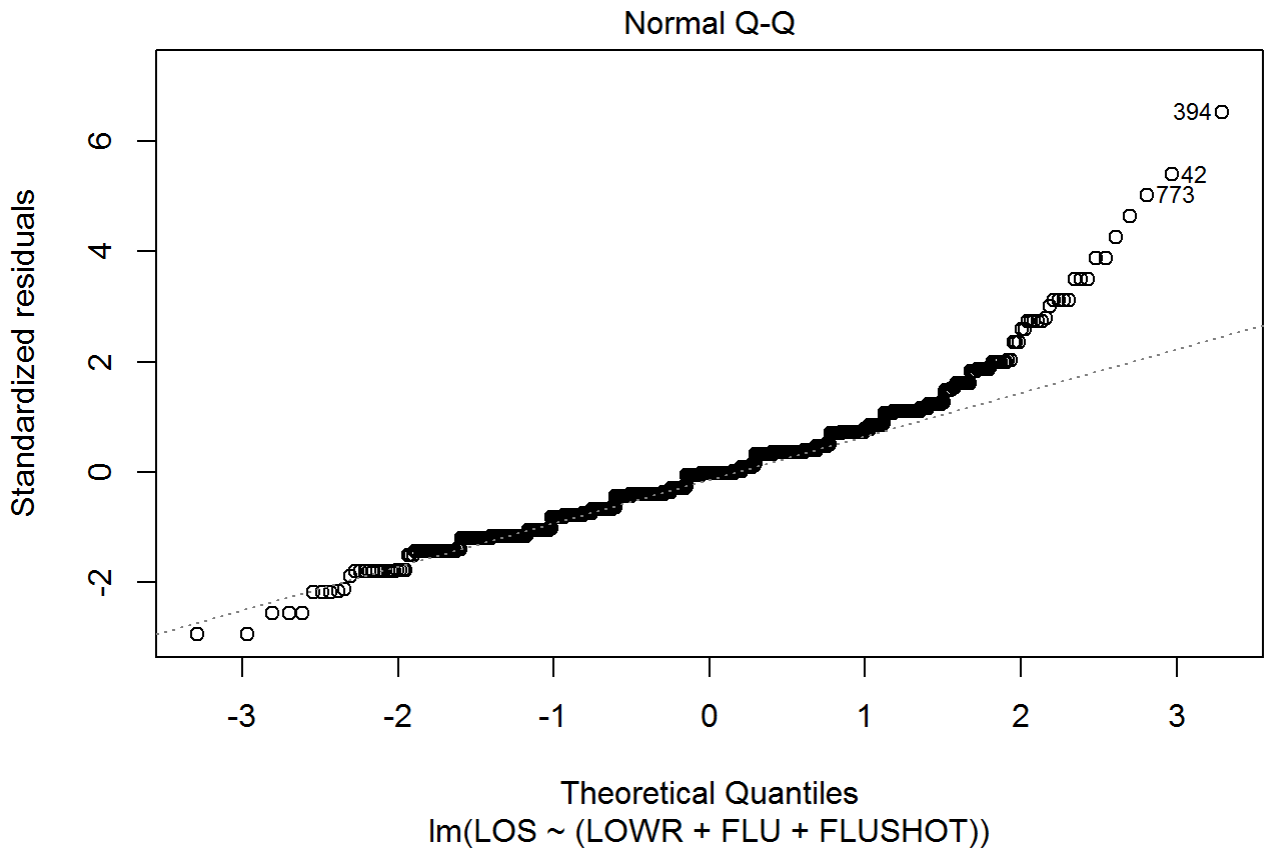
```
LOS.lm = lm(LOS ~ (LOWR+FLU+FLUSHOT),data=datafr)
summary(LOS.lm)
```

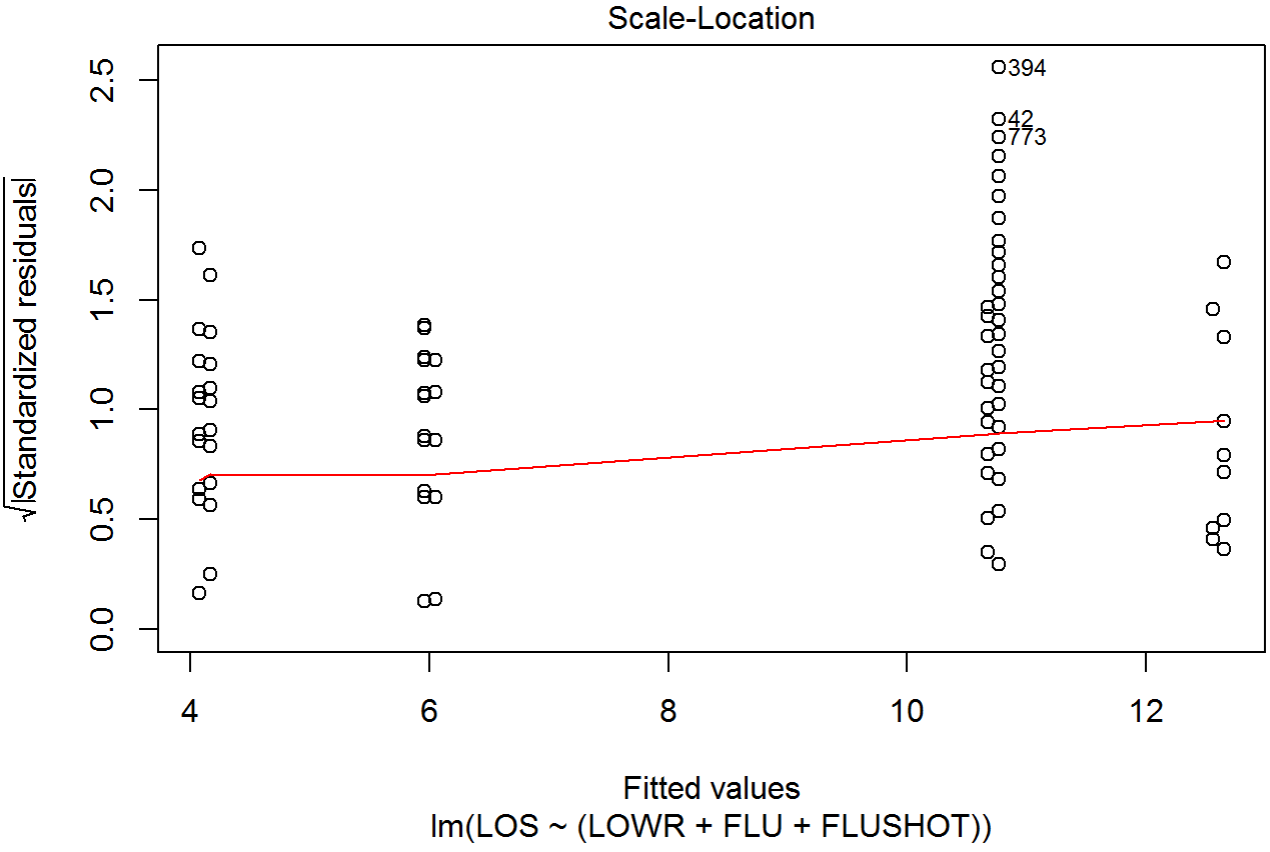
```
##
## Call:
## lm(formula = LOS ~ (LOWR + FLU + FLUSHOT), data = datafr)
##
## Residuals:
```

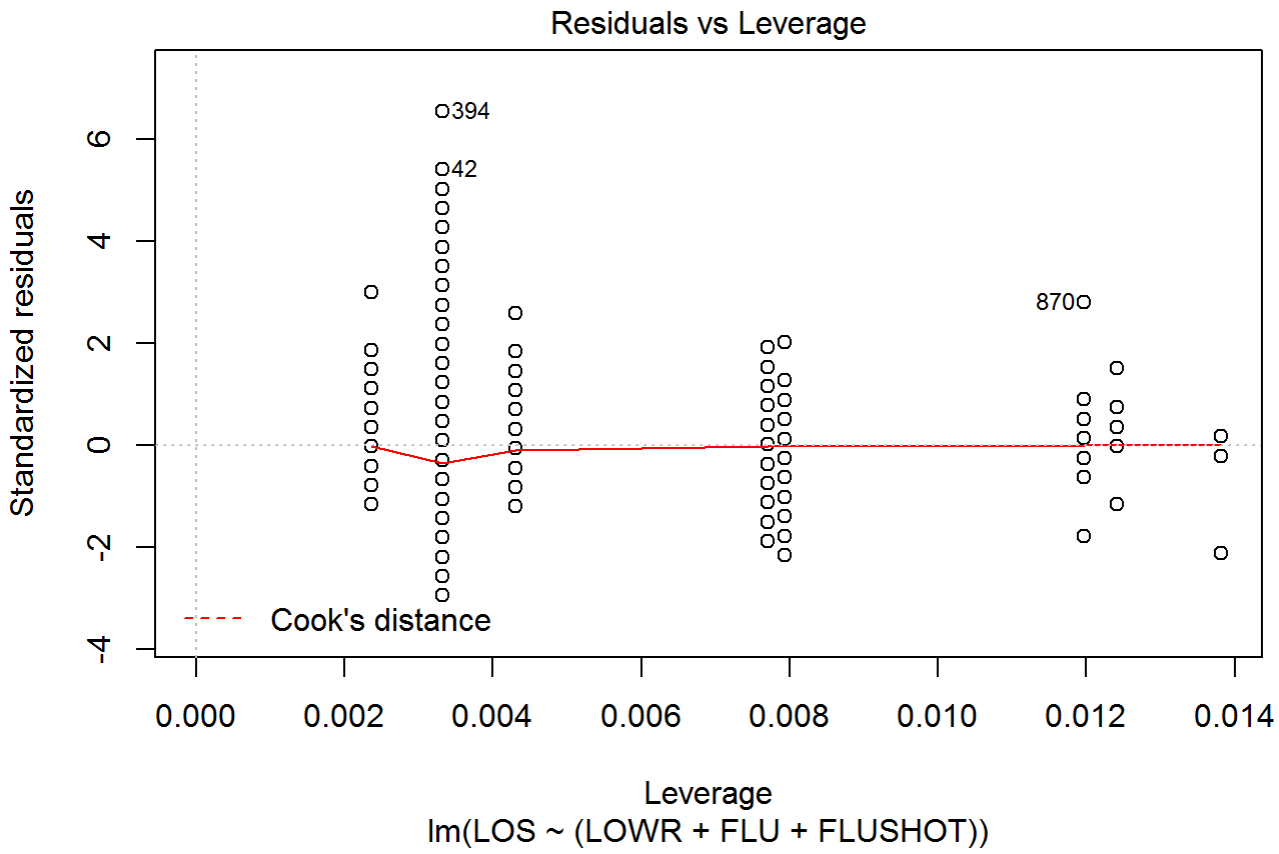
```
##     Min      1Q  Median      3Q     Max
## -7.7654 -1.7654 -0.0714  1.0429 17.2346
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.67228    0.23527  45.361  < 2e-16 ***
## LOWRTRUE    -6.60089    0.21493 -30.712  < 2e-16 ***
## FLUTRUE      1.88571    0.25443   7.412 2.66e-13 ***
## FLUSHOTTRUE  0.09312    0.20251   0.460    0.646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.641 on 996 degrees of freedom
## Multiple R-squared:  0.5711, Adjusted R-squared:  0.5698
## F-statistic: 442.1 on 3 and 996 DF,  p-value: < 2.2e-16
```

```
plot(LOS.lm)
```




Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(LOS ~ (LOWR + FLU + FLUSHOT))
394
42
773


Scale-Location
√|Standardized residuals|
Fitted values
lm(LOS ~ (LOWR + FLU + FLUSHOT))
394
42
773



Oops that model albeit not very good, says LOS is somewhat dependent on getting a FLUSHOT!

But that is not the real answer, let's do it right!

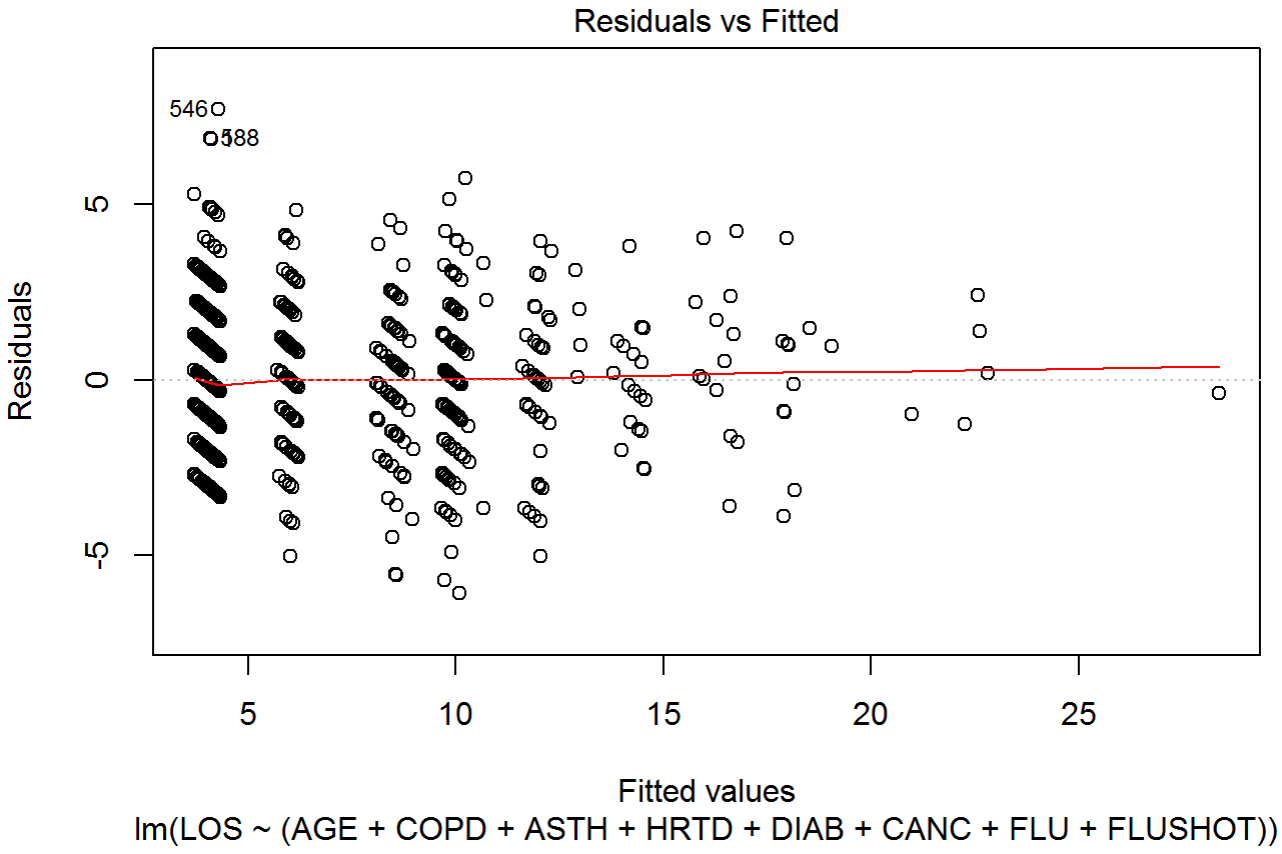
- Let's fit a model that might give good results using all the datapoints:

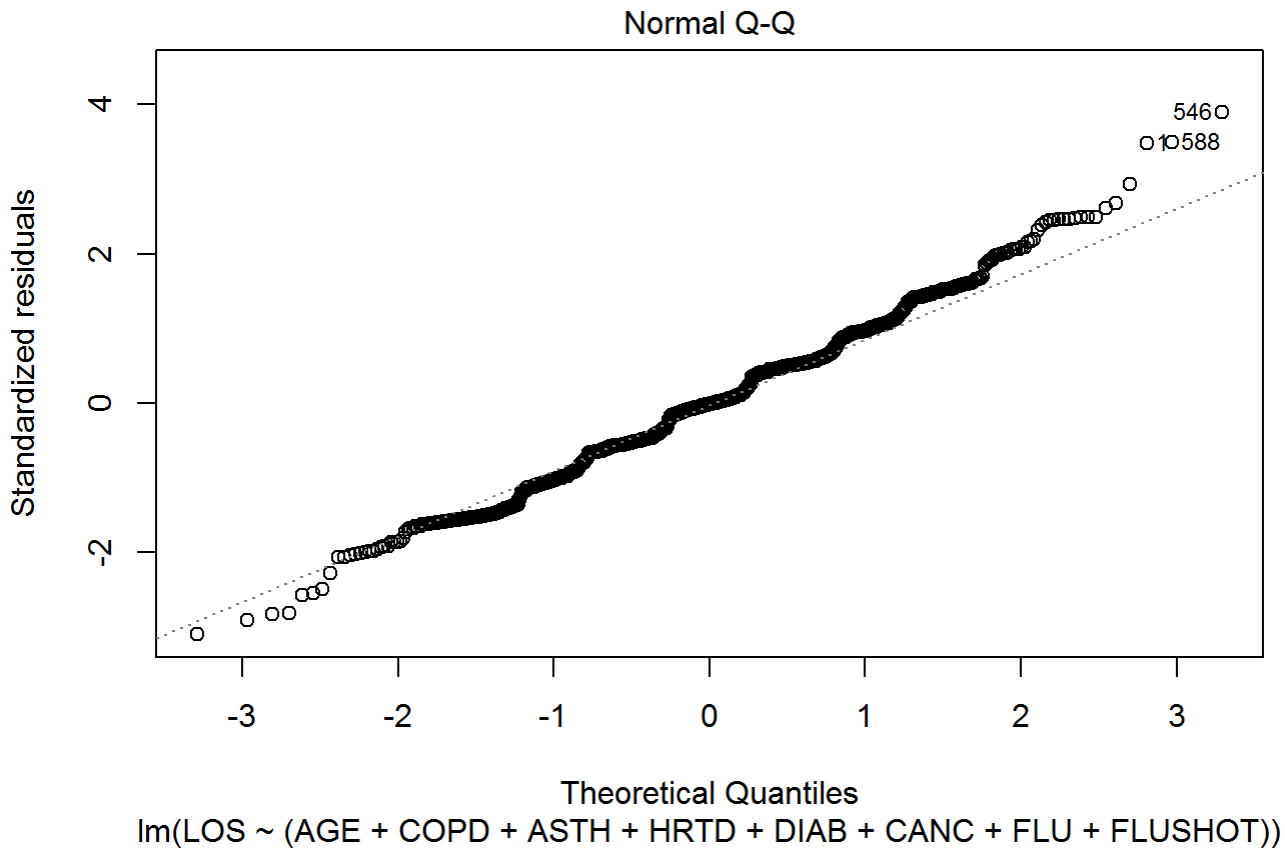
```
LOS.lm = lm(LOS ~ (AGE+COPD+ASTH+HRTD+DIAB+CANC+FLU+FLUSHOT),data=datafr)
summary(LOS.lm)
```

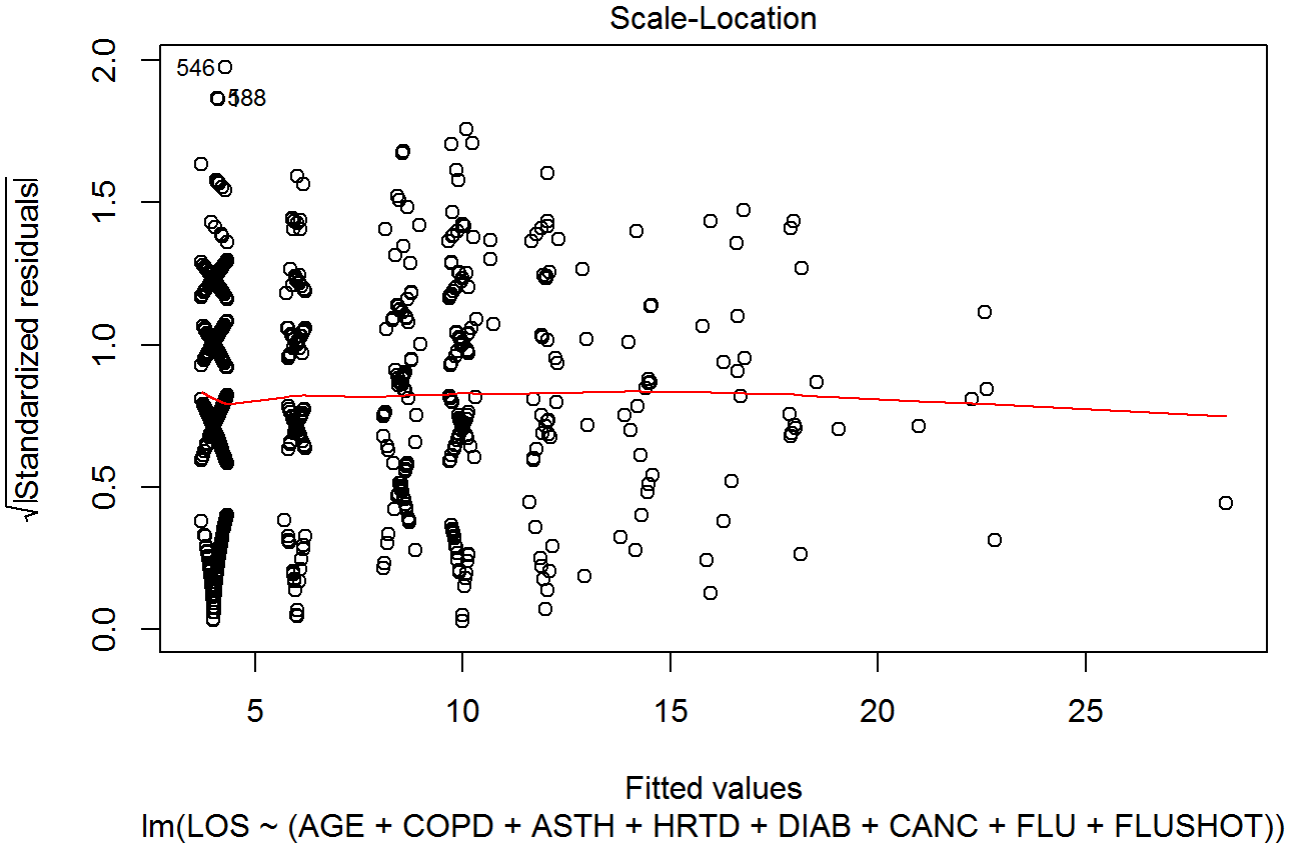
```
##
## Call:
## lm(formula = LOS ~ (AGE + COPD + ASTH + HRTD + DIAB + CANC +
##     FLU + FLUSHOT), data = datafr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1007 -1.2211 -0.0303  1.1149  7.7184
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.342034   0.147142  29.509   <2e-16 ***
## AGE         -0.004652   0.002515  -1.850   0.0646 .
## COPDTRUE     8.173764   0.279562  29.238   <2e-16 ***
```

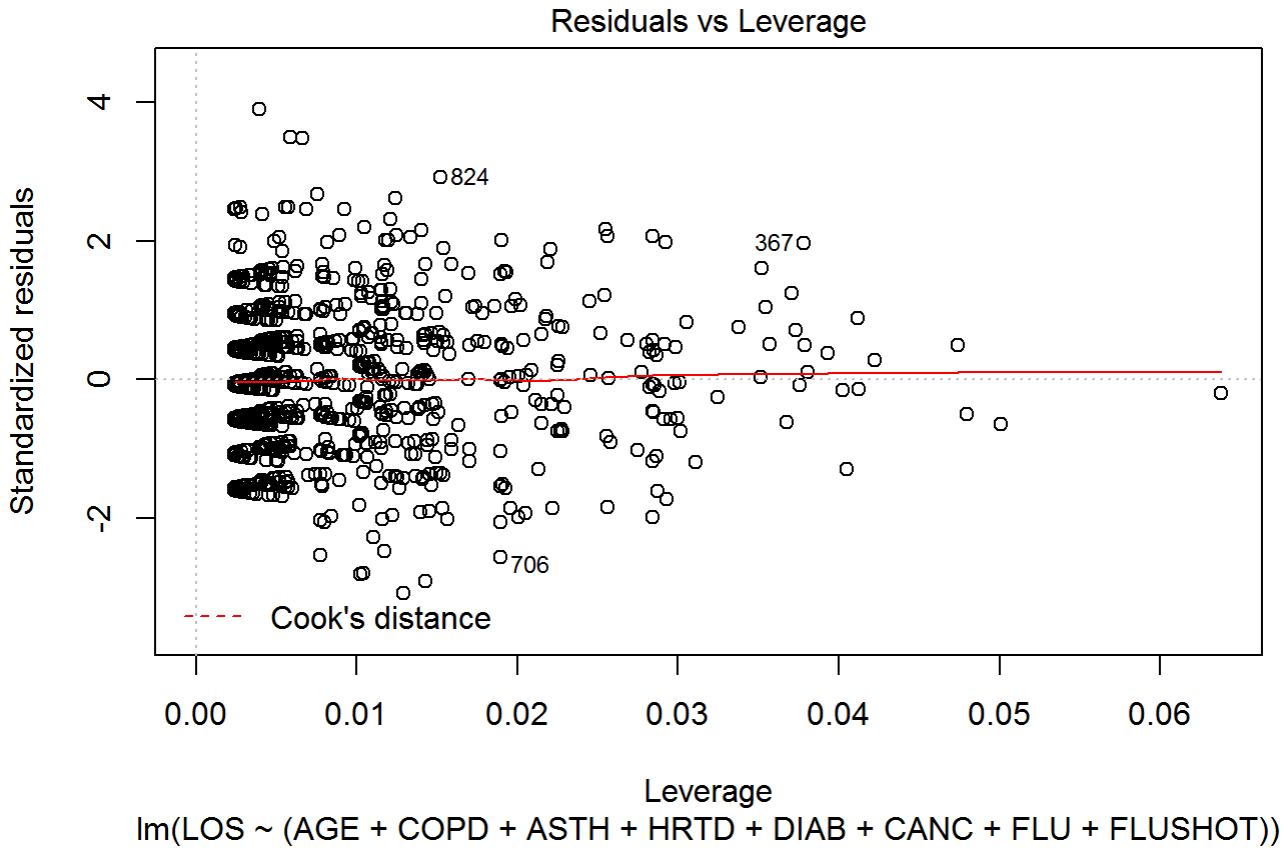
```
## ASTHTRUE     6.014550   0.228433  26.330   <2e-16 ***
## HRTDTRUE     5.964770   0.249981  23.861   <2e-16 ***
## DIABTRUE     4.386155   0.336113  13.050   <2e-16 ***
## CANCTRUE     4.729113   0.219784  21.517   <2e-16 ***
## FLUTRUE      1.881145   0.191239   9.837   <2e-16 ***
## FLUSHOTTRUE -0.204724   0.149424  -1.370   0.1710
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.983 on 991 degrees of freedom
## Multiple R-squared:  0.7593, Adjusted R-squared:  0.7574
## F-statistic: 390.8 on 8 and 991 DF,  p-value: < 2.2e-16
```

```
plot(LOS.lm)
```




Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(LOS ~ (AGE + COPD + ASTH + HRTD + DIAB + CANC + FLU + FLUSHOT))
546
588
-3
-2
-1
0
1
2
3
-2
0
2
4


Scale-Location
√|Standardized residuals|
Fitted values
lm(LOS ~ (AGE + COPD + ASTH + HRTD + DIAB + CANC + FLU + FLUSHOT))
546
588
5
10
15
20
25
0.0
0.5
1.0
1.5
2.0



Wow! the model works

- Identifies each independent condition with a very close approximation of the coefficients used above
- Identifies that FLUSHOT = 1 is NOT a significant contributing factor to LOS
- Identifies AGE as not a significant contributing factor as its effects are reflected in High Risk Conditions