Transformations

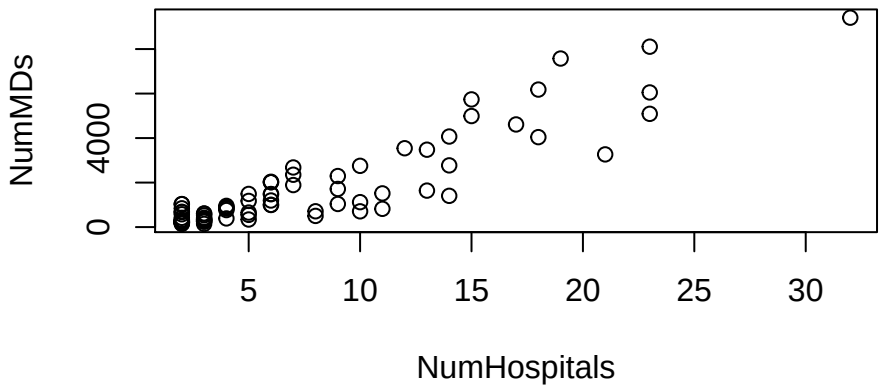
SDS 291

When regression assumptions aren't met, there are four common culprits: 1. 2. 3. 4.

In many cases, you can transform or make changes to your data to address these issues.

#Does the number of hospitals reflect the number of available doctors?

```
##                           City NumMDs RateMDs NumHospitals NumBeds RateBeds
## 1       Holland-Grand Haven, MI    349     140            3     316      127
## 2             Louisville, KY-IN   4042     340           18    3909      328
## 3              Battle Creek, MI    256     184            3     517      372
## 4                   Madison, WI   2679     510            7    1467      279
## 5             Fort Smith, AR-OK    502     179            8     975      348
## 6 Sarasota-Bradenton-Venice, FL   2352     371            7    1899      299
##   NumMedicare PctChangeMedicare MedicareRate SSBNum SSBRate SSBChange
## 1       29533               8.3        11835  34135   13679       8.1
## 2      173845               3.0        14606 202485   17013       3.0
## 3       22972               2.4        16539  27245   19615       3.3
## 4       60530               5.2        11528  68705   13085       4.9
## 5       45185               4.6        16146  55370   19785       5.8
## 6      161625               2.5        25474 175580   27674       2.7
##   NumRetired SSINum SSIRate SqrtMDs
## 1      23165   2070     820 18.6815
## 2     118920  29017    2416 63.5767
## 3      16645   4095    2945 16.0000
## 4      47085   6492    1221 51.7591
## 5      29415   9313    3301 22.4054
## 6     129855   7559    1160 48.4974
```

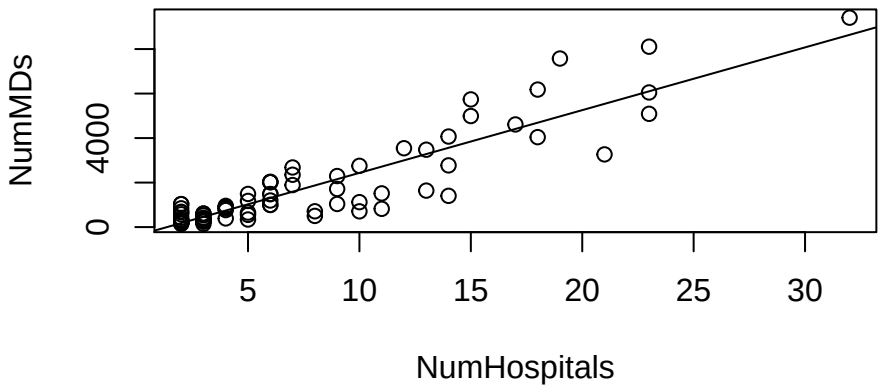


##Choose

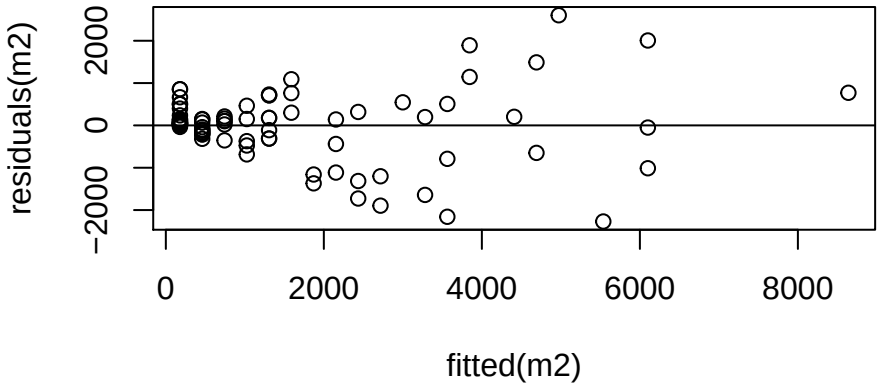
Model we choose to fit:

##Fit

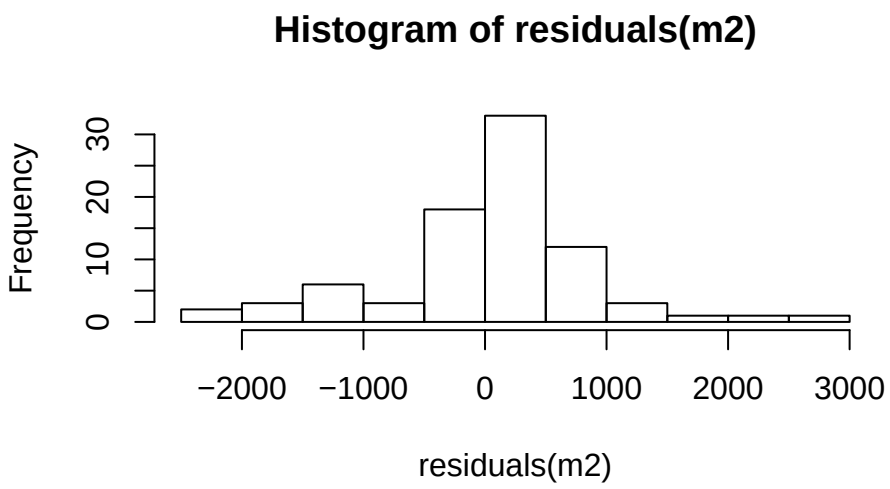
```
##
## Call:
## lm(formula = NumMDs ~ NumHospitals, data = MetroHealth83)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2270.09  -263.44    58.08   309.02  2601.93
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -385.10     138.26  -2.785  0.00666 **
## NumHospitals   282.01      14.42  19.563  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 833.2 on 81 degrees of freedom
## Multiple R-squared:  0.8253, Adjusted R-squared:  0.8232
## F-statistic: 382.7 on 1 and 81 DF,  p-value: < 2.2e-16
```

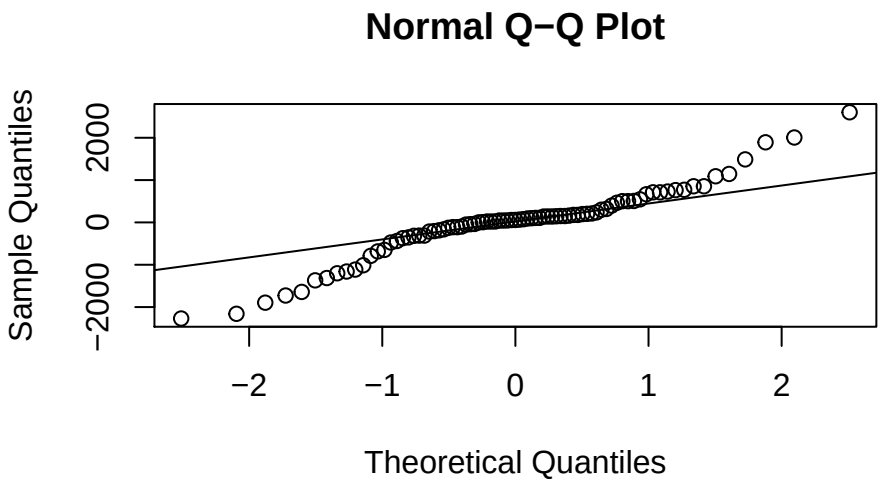


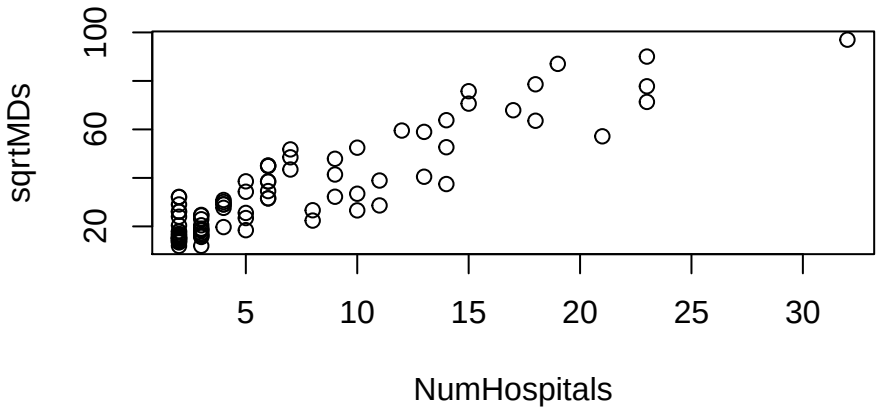
Interpretation of findings:



##Assess


Histogram of residuals(m2)
Frequency
0
10
20
30
−2000
−1000
0
1000
2000
3000
residuals(m2)


Normal Q−Q Plot
Sample Quantiles
−2000
0
2000
−2
−1
0
1
2
Theoretical Quantiles

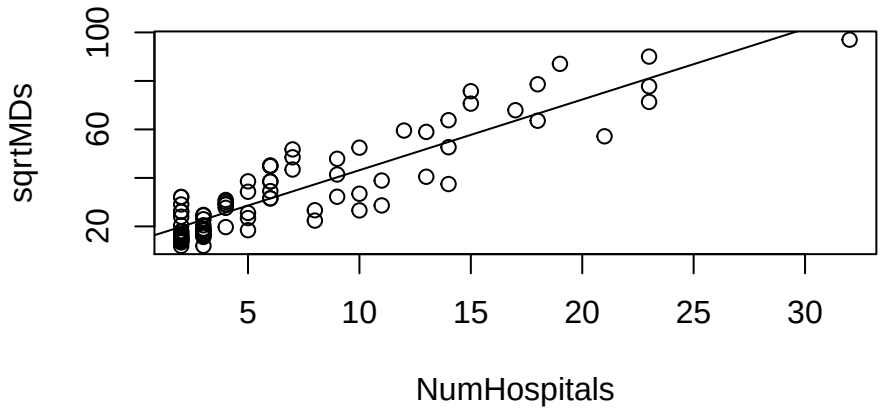


##Choose (Again)

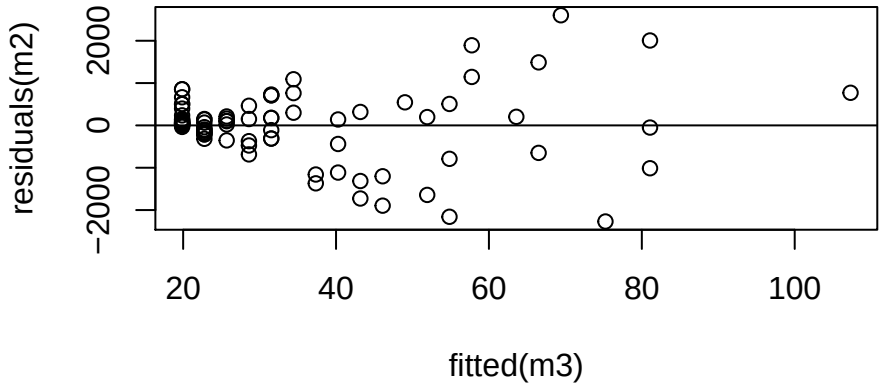
What Model are we choosing?

##Fit (Again)

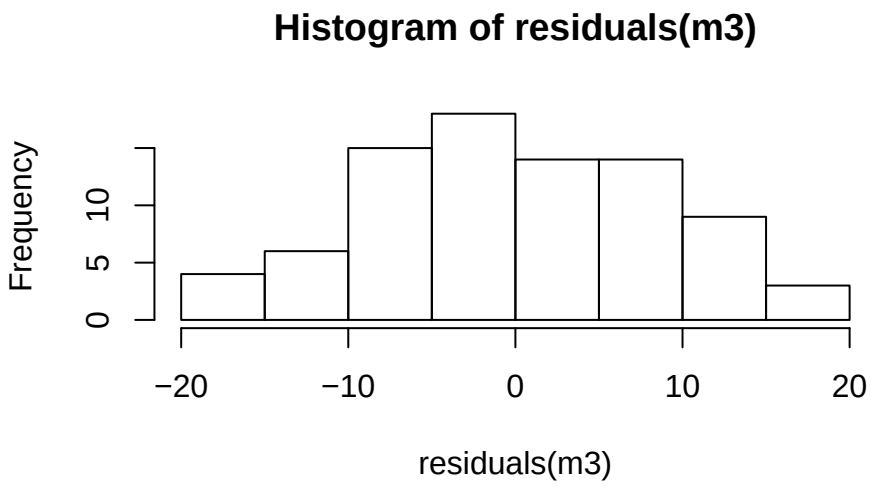
```
##
## Call:
## lm(formula = sqrtMDs ~ NumHospitals, data = MetroHealth83)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.086  -5.845  -2.030   7.001  17.994
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   14.0329     1.4686   9.555 6.36e-15 ***
## NumHospitals   2.9148     0.1531  19.036  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.85 on 81 degrees of freedom
## Multiple R-squared:  0.8173, Adjusted R-squared:  0.8151
## F-statistic: 362.4 on 1 and 81 DF,  p-value: < 2.2e-16
```

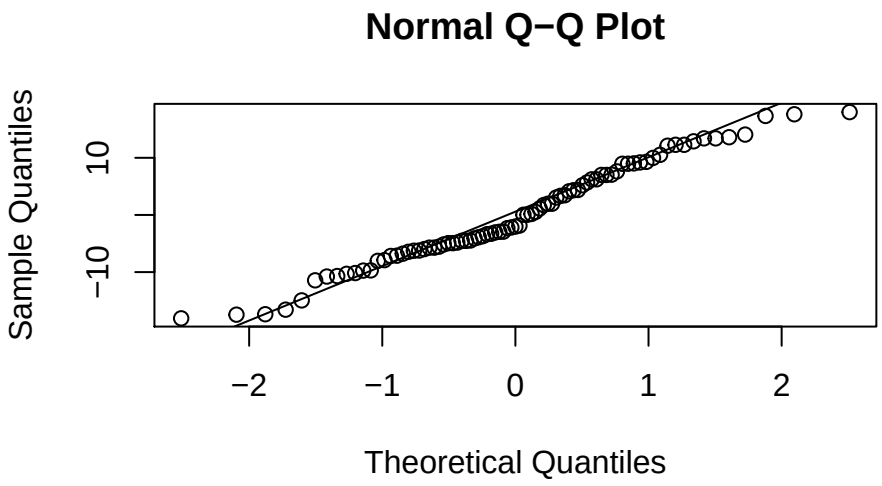


Interpretation of findings:



##Assess (Again)





##Use (Again)

What if we predict for a city like Louisville that has 18 hospitals how many doctors there would be. Be sure you have the units right!

```
#This is a part of the mosaic package that makes prediction somewhat easier
predictedMDs<-makeFun(m3)
predictedMDs(NumHospitals=18)
```

```
##        1
## 66.49963
```

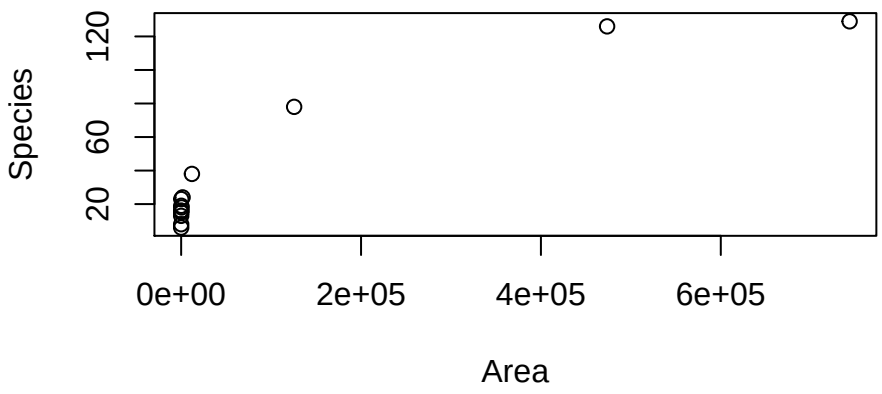
```
predictedMDs(NumHospitals=18)^2
```

```
##        1
## 4422.201
```

#Does the number of species vary by size of the island in Southeast Asia?

```
data("SpeciesArea")

plot(Species~Area, data=SpeciesArea)
```




```
SpeciesArea = SpeciesArea %>%
  mutate(ln_Species = log(Species),
         ln_Area = log(Area))

m4<-lm(ln_Species~ln_Area, data=SpeciesArea)
plot(ln_Species~ln_Area, data=SpeciesArea)
abline(m4)
```