solution

Begin by reading in the data.

library(readr)

## Warning: package 'readr' was built under R version 4.1.2

data <- read\_csv("data.csv")

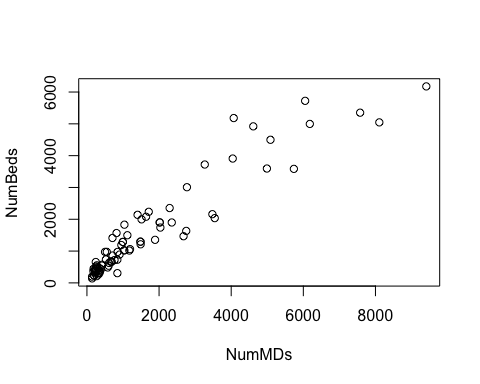
## Rows: 83 Columns: 16  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (1): City  
## dbl (15): NumMDs, RateMDs, NumHospitals, NumBeds, RateBeds, NumMedicare, Pct...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

attach(data)  
data

## # A tibble: 83 × 16  
## City NumMDs RateMDs NumHo…¹ NumBeds RateB…² NumMe…³ PctCh…⁴ Medic…⁵ SSBNum  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Hollan… 349 140 3 316 127 29533 8.3 11835 34135  
## 2 Louisv… 4042 340 18 3909 328 173845 3 14606 202485  
## 3 Battle… 256 184 3 517 372 22972 2.4 16539 27245  
## 4 Madiso… 2679 510 7 1467 279 60530 5.2 11528 68705  
## 5 Fort S… 502 179 8 975 348 45185 4.6 16146 55370  
## 6 Saraso… 2352 371 7 1899 299 161625 2.5 25474 175580  
## 7 Anders… 200 153 2 231 176 22828 1.1 17408 26740  
## 8 Honolu… 3478 389 13 2160 242 126752 5.2 14188 136730  
## 9 Ashevi… 1489 389 5 1213 317 76397 5.1 19970 87520  
## 10 Winsto… 2018 462 6 1901 435 66298 6 15165 76180  
## # … with 73 more rows, 6 more variables: SSBRate <dbl>, SSBChange <dbl>,  
## # NumRetired <dbl>, SSINum <dbl>, SSIRate <dbl>, SqrtMDs <dbl>, and  
## # abbreviated variable names ¹​NumHospitals, ²​RateBeds, ³​NumMedicare,  
## # ⁴​PctChangeMedicare, ⁵​MedicareRate

Next, let x=NumMDs, y=NumBeds. Visualize the relationship between the 2 variables using a scatterplot.

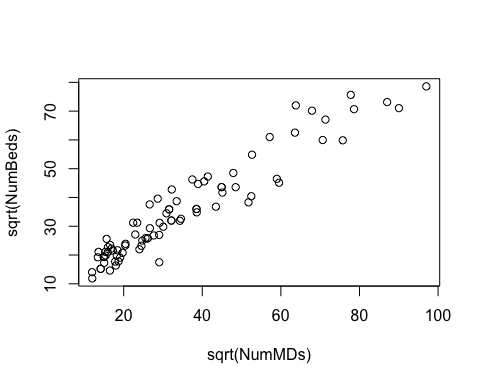
plot(NumBeds~NumMDs)

 Notice that there appears to be some clustering near the origin, as well as some “fanning” at larger x-values.

Since the scatterplot does not satisfy the conditions for linearity, we must perform transformations on one (or both) of the variables to get a better linear model.

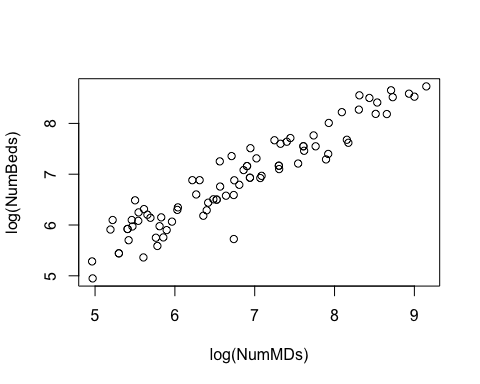
We start by attempting a square-root transformation on both variables.

plot(sqrt(NumBeds)~sqrt(NumMDs))

 Notice that the data appear much better than before, though there are still signs of clustering near the origin.

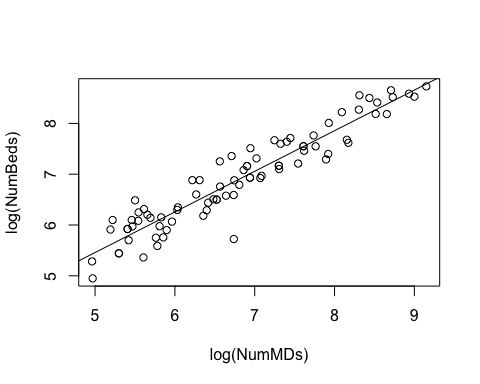
Let’s compare our results using a log transform on both variables.

plot(log(NumBeds)~log(NumMDs))

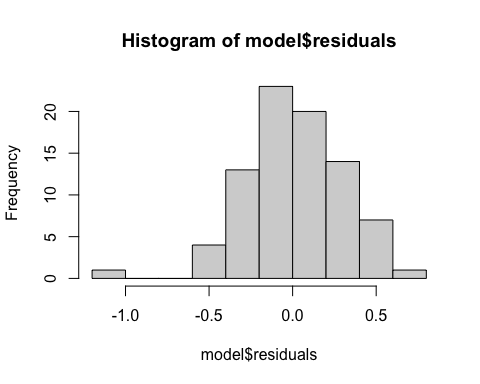
 Indeed, not only do the data appear more linear, but they also seem to be more well-dispersed. Thus, we gather that using log transforms on both variables should suffice!

To be sure, let’s fit a linear model & verify that the conditions for linearity are met. Specifically, 1. Zero-mean 2. Homoscedasticity 3. Independence 4. Normality

plot(log(NumBeds)~log(NumMDs))  
model = lm(log(NumBeds)~log(NumMDs))  
abline(model)

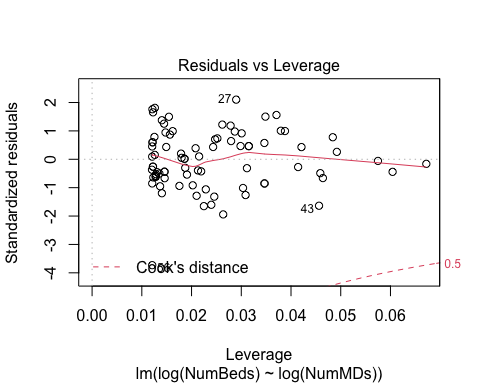
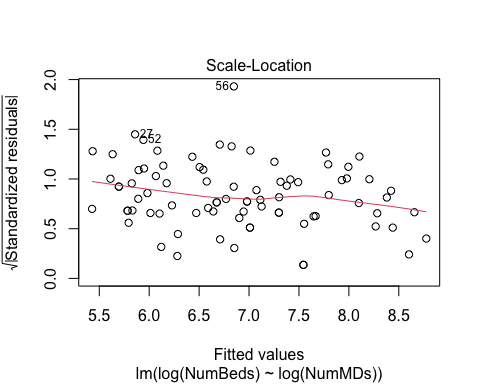
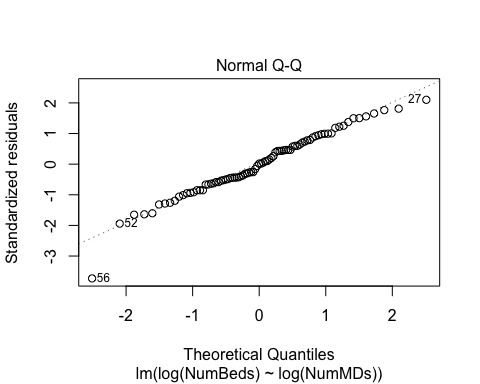
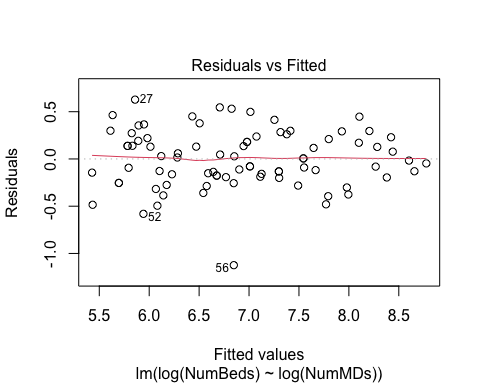
 We can plot a histogram of the model residuals to verify normality.

hist(model$residuals)

 Indeed, the residuals appear fairly normal, with the exception of an outlier or two. Hence, we conclude that the condition for normality are met.

Then, we check the conditions of zero-mean & constant variance by plotting the model residuals against the model’s fitted values.

plot(model)

 Notice, the residuals are well dispersed about the zero-line, with no signs of clustering or grouping, nor any trends like fanning or tailing. Hence, we determine that the 2 conditions are indeed satisfied.

Finally, we verify independence using the model’s p-value.

summary(model)

##   
## Call:  
## lm(formula = log(NumBeds) ~ log(NumMDs))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.1239 -0.1822 0.0056 0.2153 0.6277   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.45967 0.20894 6.986 7.17e-10 \*\*\*  
## log(NumMDs) 0.79960 0.03033 26.367 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3031 on 81 degrees of freedom  
## Multiple R-squared: 0.8956, Adjusted R-squared: 0.8944   
## F-statistic: 695.2 on 1 and 81 DF, p-value: < 2.2e-16

At a p-value: < 2.2e-16, we can be confident the transformed variables are independent, thus concluding that all 4 conditions are satisfied!