Shawn Urban
NEAS VEE Regression Analysis
Student Project
Spring 2014

Hypothesis

A few months ago I downloaded an app for my IPhone that claimed it could measure my sleep quality through the night. Without proper sleep testing data, I believed this claim to be false. Upon using the app for 24 days I determined that the sleep quality it claimed it could measure appeared to be based almost completely on how long I had slept. This was not tested; I merely perceived that this was what was happening. Given this project, I thought it the perfect opportunity to test my hypothesis: *the sleep quality indicated within my sleep app is purely based on time slept through the night.* In order to test this hypothesis I was able to use data provided by the app itself; namely, day of the week (Day), time going to bed (InBed), amount of sleep received (Time), time getting out of bed, (OutBed) and steps taken throughout the day (Activity). Sleep.Quality is the response variable.

Analyze the Data

Below is the input data provided by the app I use to quantify sleep. All of the columns are factors except for "Activity" which is an integer.

	Day	InBed	OutBed	Time	Activity	Sleep.Quality
1	Wednesday	22:55	5:56	7:01	1804	81%
2	Thursday	22:57	5:59	7:02	3549	59%
3	Friday	21:48	5:34	7:46	2262	73%
4	Saturday	0:32	6:58	6:26	1123	74%
5	Sunday	23:45	6:00	6:14	2779	67%
6	Monday	23:16	5:58	6:42	2668	77%
7	Tuesday	23:03	5:57	6:53	2478	67%
8	Wednesday	23:35	6:02	6:26	3597	73%
9	Thursday	23:31	6:01	6:29	2170	73%
10	Friday	22:50	7:10	8:19	2357	85%
11	Saturday	22:39	7:18	8:39	1995	99%
12	Sunday	23:58	6:01	6:02	1796	66%
13	Monday	23:01	5:58	6:57	4179	71%
14	Tuesday	22:51	6:00	7:09	2459	65%
15	Wednesday	23:09	5:59	6:50	3687	77%
16	Saturday	23:37	6:52	7:15	1463	66%
17	Sunday	22:34	6:01	7:26	4654	73%
18	Monday	0:10	5:58	5:48	2503	59%
19	Tuesday	23:59	6:00	6:01	2019	61%
20	Wednesday	23:48	5:59	6:10	3168	63%
21	Thursday	23:01	5:59	6:58	2349	60%
22	Friday	22:56	7:18	8:22	6267	93%
23	Saturday	22:57	7:13	8:15	3500	79%
24	Sunday	23:03	6:00	6:57	6403	45%

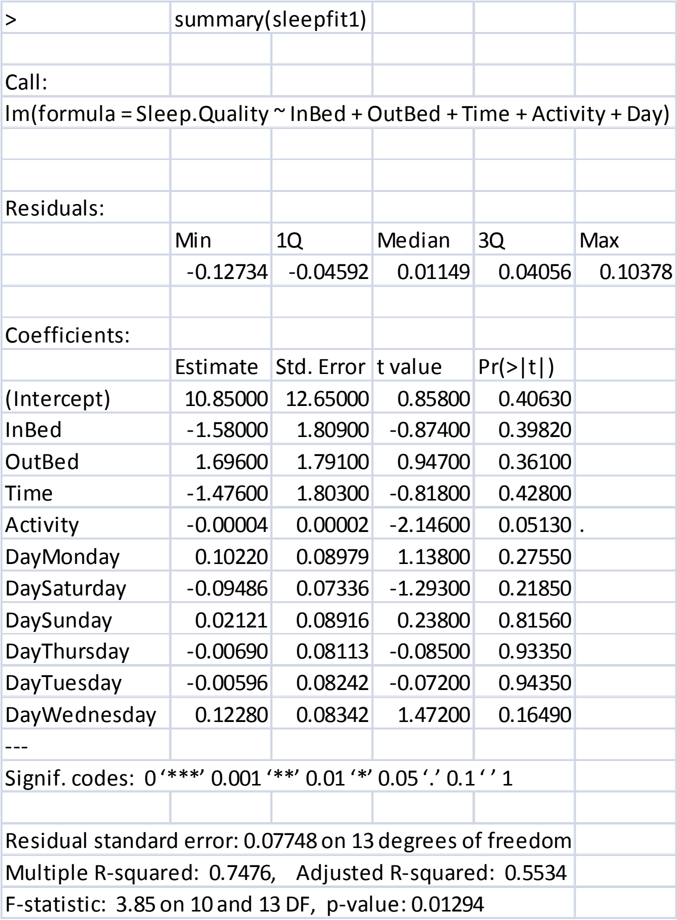
My first step was to get this data into a form more easily usable in a regression setting. First, I took my response variable, Sleep.Quality and noticed that this value should really be in decimal form. Also, the InBed, OutBed and Time columns can be converted into hours and thereby made continuous. For InBed, I will use a measurement of time passed since 21:00 being there are no InBed times before that. In the same vein, I will measure OutBed as time after 4:00 as there are also no out of bed times before that. Time will simply be converted into hours. Day will need to remain as is (qualitative) and will be considered a dummy variable, although this may get lumped into weekdays versus weekends instead of each day of the week. Below is the data updated to a more usable type with only the first 10 rows visible.

	Day	InBed	OutBed	Time	Activity	Sleep.Quality
1	Wednesday	1.92	1.93	7.02	1804	0.81
2	Thursday	1.95	1.98	7.03	3549	0.59
3	Friday	0.8	1.57	7.77	2262	0.73
4	Saturday	3.53	2.97	6.43	1123	0.74
5	Sunday	2.75	2	6.23	2779	0.67
6	Monday	2.27	1.97	6.7	2668	0.77
7	Tuesday	2.05	1.95	6.88	2478	0.67
8	Wednesday	2.58	2.03	6.43	3597	0.73
9	Thursday	2.52	2.02	6.48	2170	0.73
10	Friday	1.83	3.17	8.32	2357	0.85

Before any regression testing with the data I wanted to build my initial hypothesis. Sleep quality as a direct measurement of time slept. Below are the results of this regression. According to the Adj R-squared, Time only accounted for .4071 of the resulting Sleep.Quality. This did not bode well for my hypothesis, but this is also ok, as I also wanted to see if adding the other explanatory variables impacts the regression results or if Time is the only real factor.

```
> summary(sleepfit0)

Call:
lm(formula = Sleep.Quality ~ Time)

Residuals:
                       Min        1Q         Median     3Q         Max
                       -0.25558   -0.04414   0.001953   0.074841   0.120812

Coefficients:
                       Estimate   Std. Error  t value   Pr(>|t|)
(Intercept)            0.03671    0.16551     0.222     0.82651
Time                   0.09624    0.02349     4.098     0.000475 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08928 on 22 degrees of freedom
Multiple R-squared:  0.4329,    Adjusted R-squared:  0.4071
F-statistic: 16.79 on 1 and 22 DF,  p-value: 0.000475
```

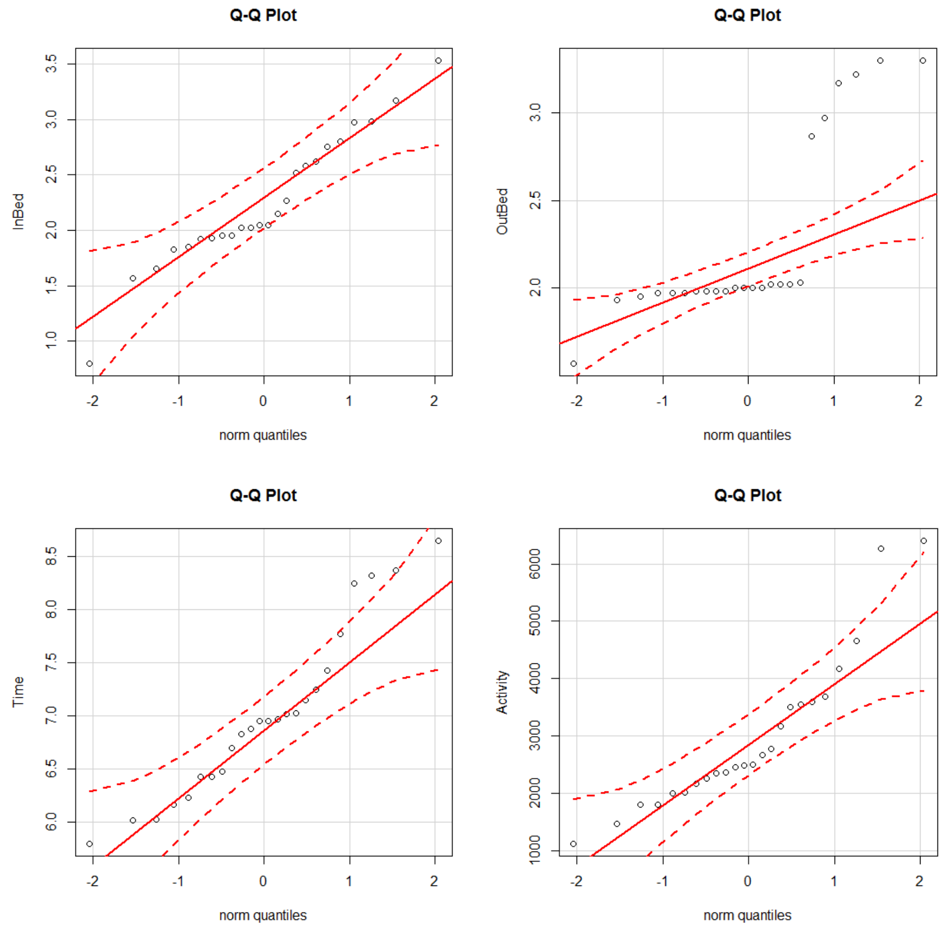
Before moving on and for comparison reasons, I ran a regression using the data as provided from the app (adjusted as described above). The data below shows the linear regression summary for each predictor. The only strong predictor at this point appeared to be Activity; Adj R-squared = .5534. All the other variables did not appear to make any significant contribution.



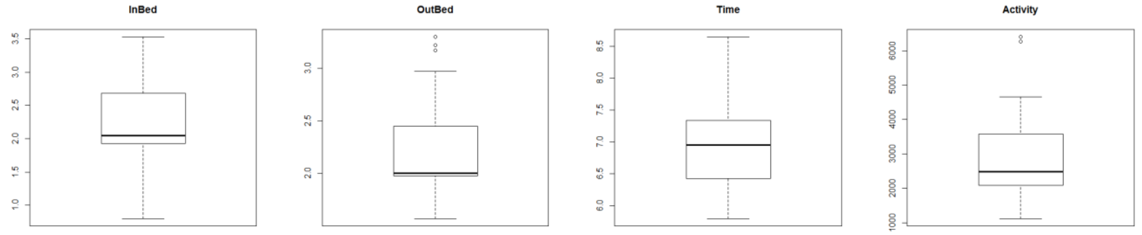
>	summary(sleepfit1)				
Call:					
lm(formula = Sleep.Quality ~ InBed + OutBed + Time + Activity + Day)					
Residuals:					
	Min	1Q	Median	3Q	Max
	-0.12734	-0.04592	0.01149	0.04056	0.10378
Coefficients:					
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	10.85000	12.65000	0.85800	0.40630	
InBed	-1.58000	1.80900	-0.87400	0.39820	
OutBed	1.69600	1.79100	0.94700	0.36100	
Time	-1.47600	1.80300	-0.81800	0.42800	
Activity	-0.00004	0.00002	-2.14600	0.05130	.
DayMonday	0.10220	0.08979	1.13800	0.27550	
DaySaturday	-0.09486	0.07336	-1.29300	0.21850	
DaySunday	0.02121	0.08916	0.23800	0.81560	
DayThursday	-0.00690	0.08113	-0.08500	0.93350	
DayTuesday	-0.00596	0.08242	-0.07200	0.94350	
DayWednesday	0.12280	0.08342	1.47200	0.16490	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					
Residual standard error: 0.07748 on 13 degrees of freedom					
Multiple R-squared: 0.7476, Adjusted R-squared: 0.5534					
F-statistic: 3.85 on 10 and 13 DF, p-value: 0.01294					

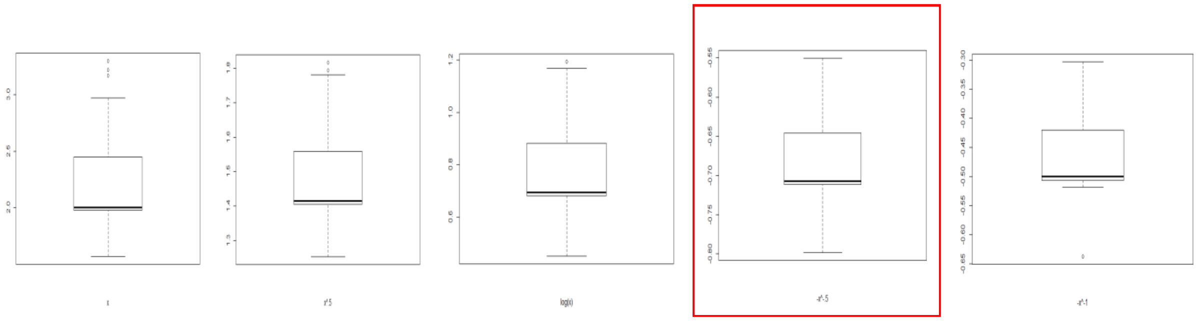
I then checked the variables for normal distribution qualities. The charts below depict the Q-Q plots for each explanatory variable. It is readily apparent that OutBed was definitely not normally distributed, this value needed to be further evaluated. InBed appeared to be normal while Time and Activity both appeared to have a positive skew. It was my intention to transform Time and Activity to a better normal distribution.



The charts below are the box-plots depicting the distribution of each variable. Again, we see that OutBed had a large positive skew. In this case, InBed and Activity also showed positive skews. Time looked pretty good.

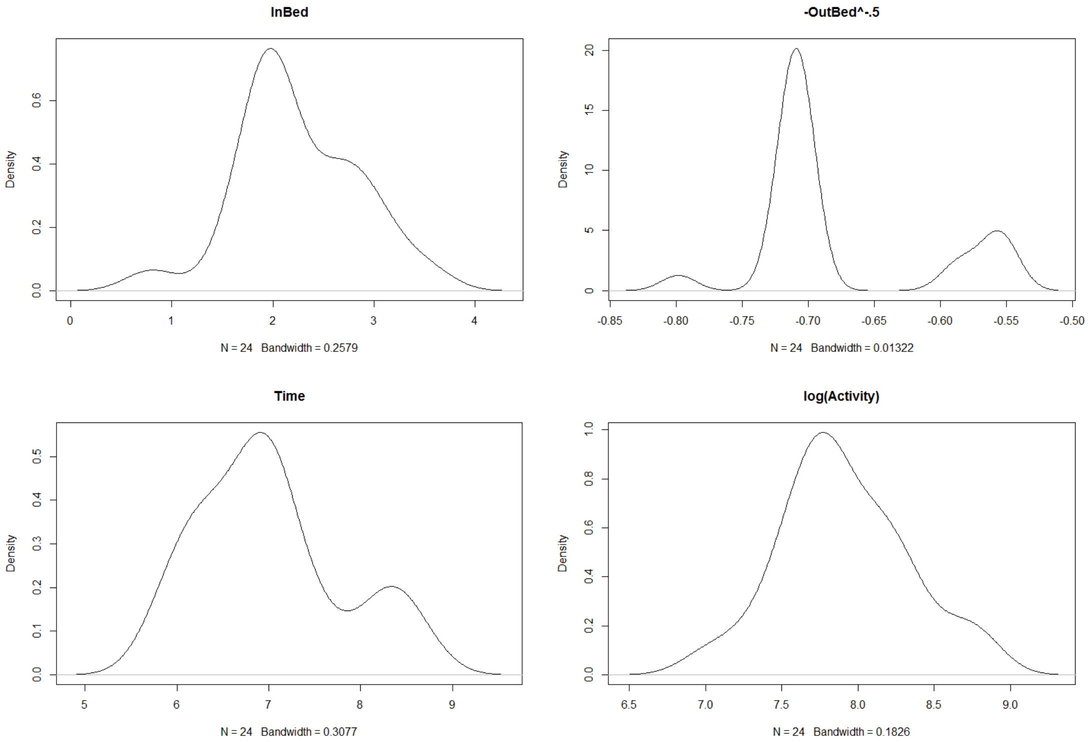


Using OutBed as an example I used several transformation “down the ladder” in order to offset the positive skew. While still skewed it looks like –x^-.5 was the best transformation here. This transformation is the one in the border below.

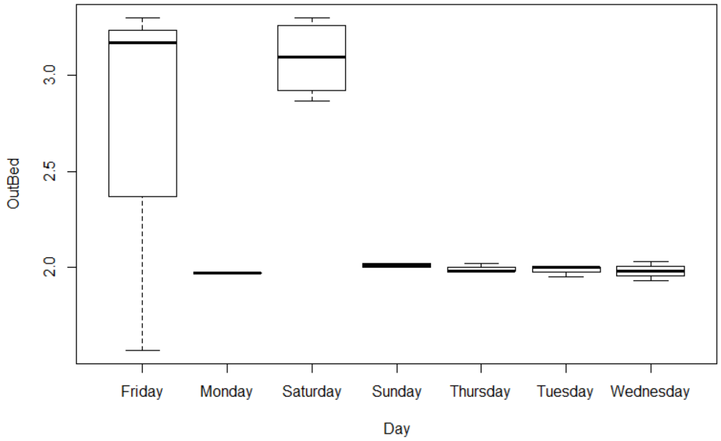


I used the same process with Activity and found that the log(Activity) was the best transformation.

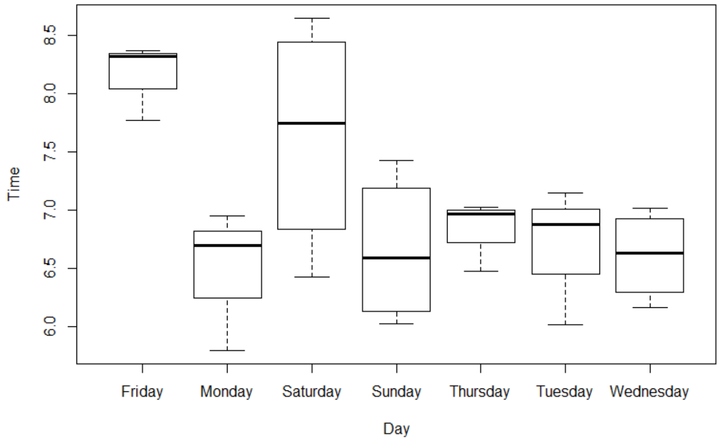
Here are the density distributions of the all the variables.



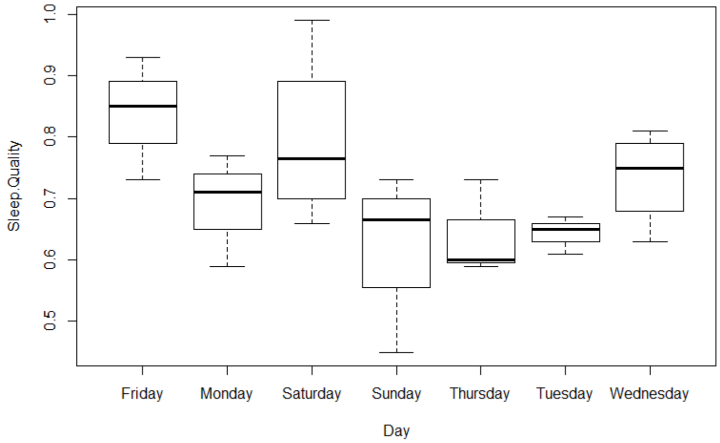
It appeared the OutBed transformation had two modes and perhaps Time did as well. This may have been due to weekends where I tended to get out of bed later. If I do a box-plot of OutBed against Day it yields the following results.



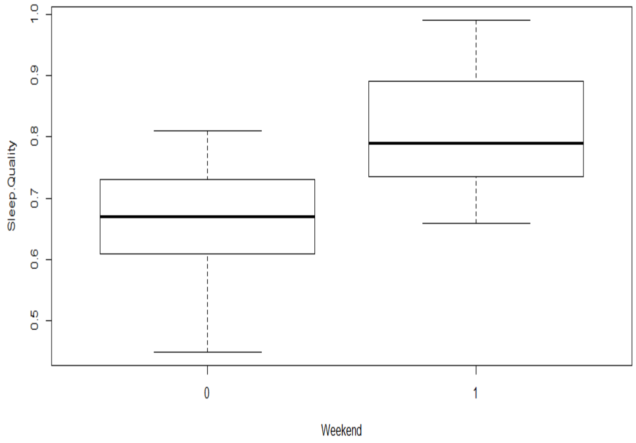
Without a doubt there is a correlation between OutBed and Day of the week. Doing the same for Time yielded the following. Again there appeared to be a strong correlation between Time in bed and day of the week. This further supported my initial thought that I would create a dummy variable indication weekend of weekday in place of day of the week.



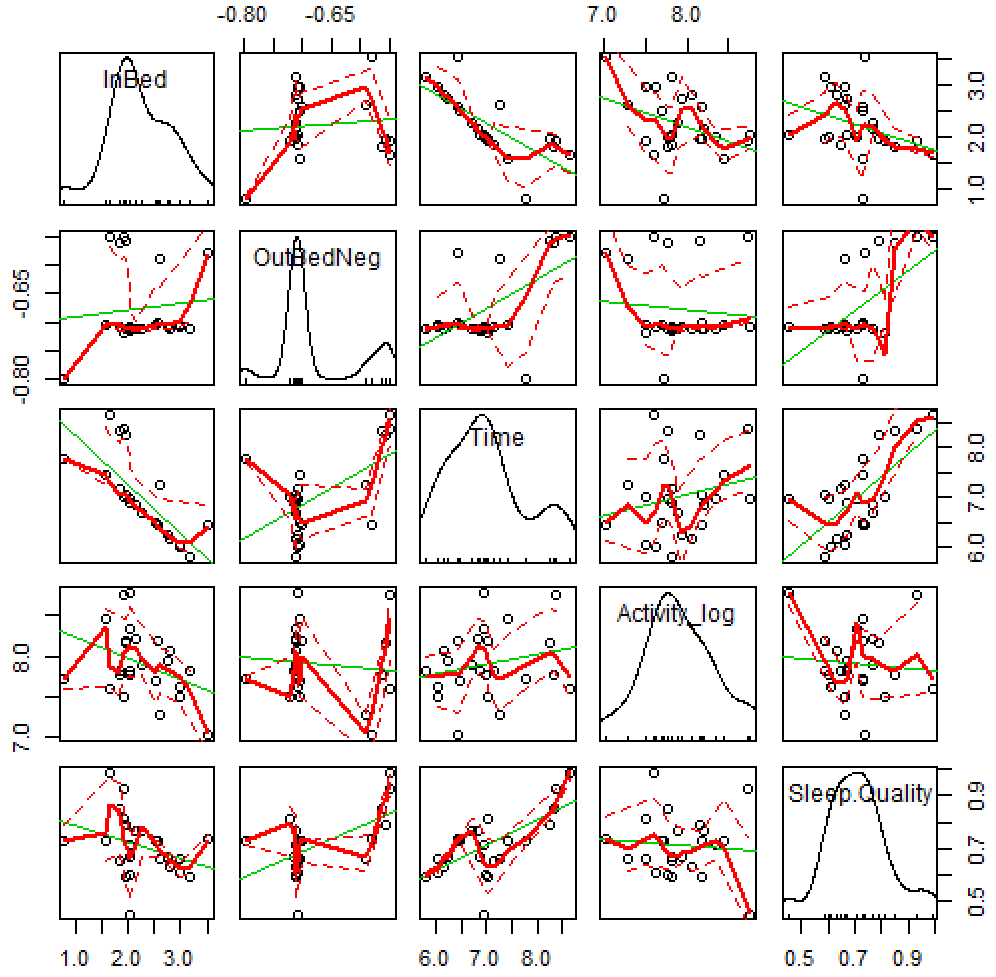
With this in mind, I focused my attention on the variable Day against Sleep.Quality.



I saw again the same pattern. This led me to believe that a dummy variable was necessary. Here is the box-plot after the dummy variable had been created. Comparing this to the graphs above it was easy to see that there was now two strong indicators for Sleep.Quality.

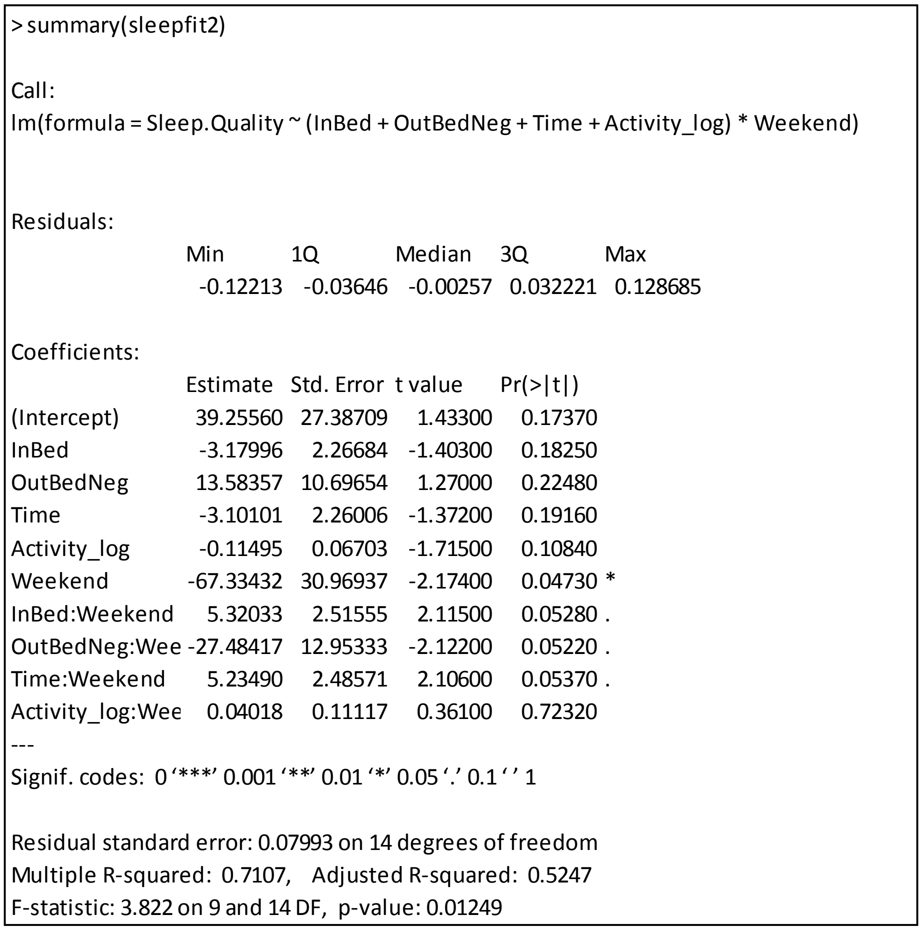


Unfortunately, while all these transformations normalized the variables, it appeared that it did very little for the linearization of them. Looking below, it showed evidence that only Time and Activity_log had some sort of relationship, although it was hard to tell.



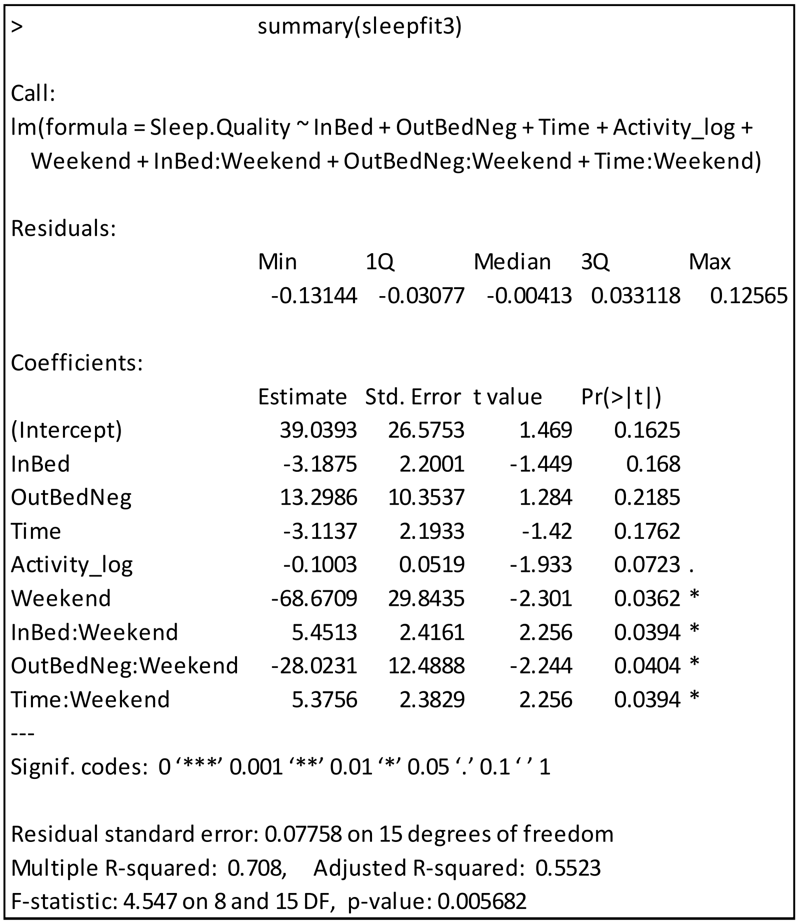
Fitting the Regression

The next step was to fit the variables to a regression. I used 'R's lm function to do this. Using the following function: sleepfit2<-lm(Sleep.Quality~(InBed+OutBedNeg+Time+Activity_log)*Weekend), I was able to measure each individual variable as well as its effects with the dummy variable Weekend. Weekend also came through on its own to make a direct adjustment to the intercept if necessary. Compared to sleepfit1 I saw the variables are all much stronger with only a slight diminishing of Adj R-squared.

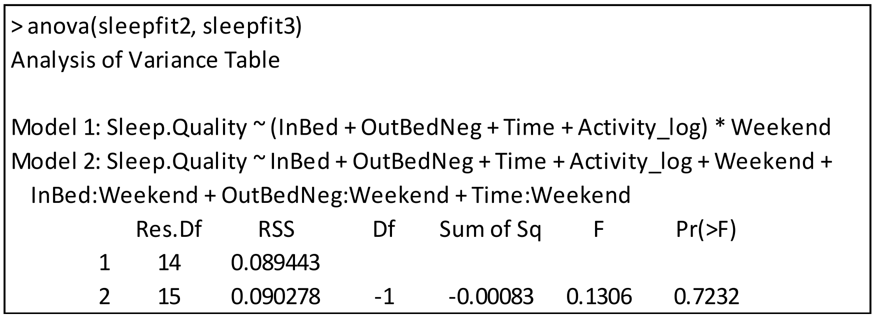


```
> summary(sleepfit2)

Call:
lm(formula = Sleep.Quality ~ (InBed + OutBedNeg + Time + Activity_log) * Weekend)


Residuals:
                  Min        1Q         Median     3Q          Max
                  -0.12213   -0.03646   -0.00257   0.032221   0.128685

Coefficients:
                  Estimate   Std. Error  t value    Pr(>|t|)
(Intercept)       39.25560   27.38709    1.43300    0.17370
InBed             -3.17996    2.26684   -1.40300    0.18250
OutBedNeg         13.58357   10.69654    1.27000    0.22480
Time              -3.10101    2.26006   -1.37200    0.19160
Activity_log      -0.11495    0.06703   -1.71500    0.10840
Weekend          -67.33432   30.96937   -2.17400    0.04730 *
InBed:Weekend      5.32033    2.51555    2.11500    0.05280 .
OutBedNeg:Wee    -27.48417   12.95333   -2.12200    0.05220 .
Time:Weekend       5.23490    2.48571    2.10600    0.05370 .
Activity_log:Wee   0.04018    0.11117    0.36100    0.72320
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07993 on 14 degrees of freedom
Multiple R-squared:  0.7107,    Adjusted R-squared:  0.5247
F-statistic: 3.822 on 9 and 14 DF,  p-value: 0.01249
```

In order to make this an even stronger regression model I began removing all variables that did not provide much significance. The most obvious choice here was Activity_log:Weekend. Looking at the summary below, I saw that this dramatically improved the significance of the explanatory variables. There was now four beneath .05 significance and one under .1 significance. The Adj R-Squared also went up to .5523. Looking at the ANOVA for each model I saw that removing Activity_log:Weekend had no significance on the model (Pr>F=.7232). Overall, this appeared to be a very good decision to remove.

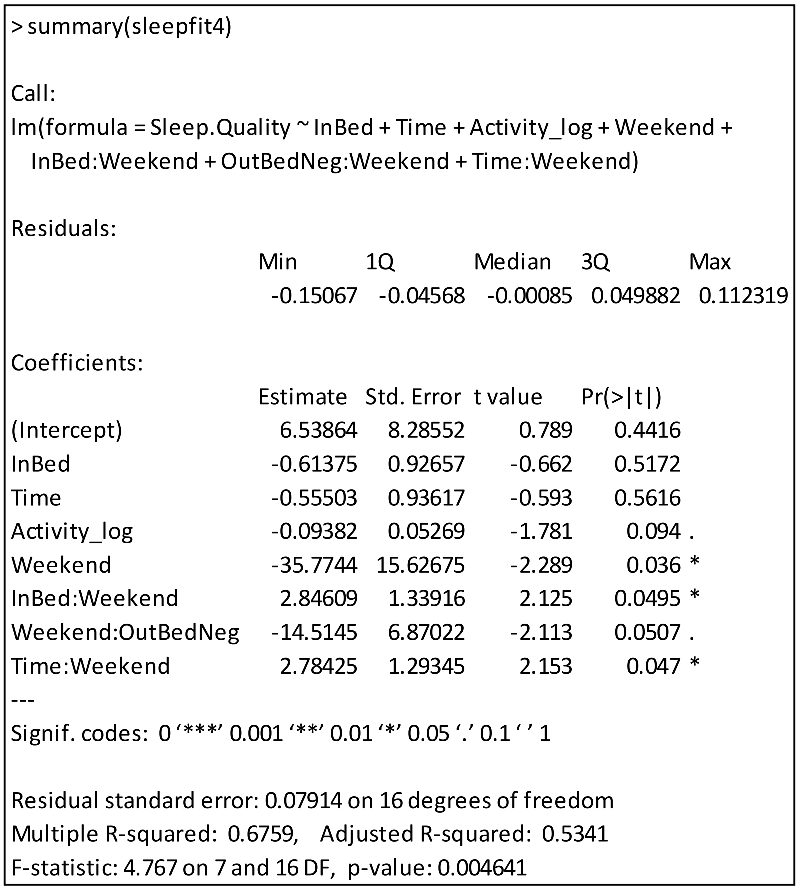


```
>                        summary(sleepfit3)

Call:
lm(formula = Sleep.Quality ~ InBed + OutBedNeg + Time + Activity_log +
  Weekend + InBed:Weekend + OutBedNeg:Weekend + Time:Weekend)

Residuals:
                        Min        1Q         Median     3Q         Max
                         -0.13144   -0.03077   -0.00413   0.033118   0.12565

Coefficients:
                        Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)              39.0393    26.5753     1.469    0.1625
InBed                    -3.1875     2.2001    -1.449     0.168
OutBedNeg                13.2986    10.3537     1.284    0.2185
Time                     -3.1137     2.1933     -1.42    0.1762
Activity_log             -0.1003     0.0519    -1.933    0.0723 .
Weekend                 -68.6709    29.8435    -2.301    0.0362 *
InBed:Weekend             5.4513     2.4161     2.256    0.0394 *
OutBedNeg:Weekend       -28.0231    12.4888    -2.244    0.0404 *
Time:Weekend              5.3756     2.3829     2.256    0.0394 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07758 on 15 degrees of freedom
Multiple R-squared:  0.708,    Adjusted R-squared:  0.5523
F-statistic: 4.547 on 8 and 15 DF,  p-value: 0.005682
```

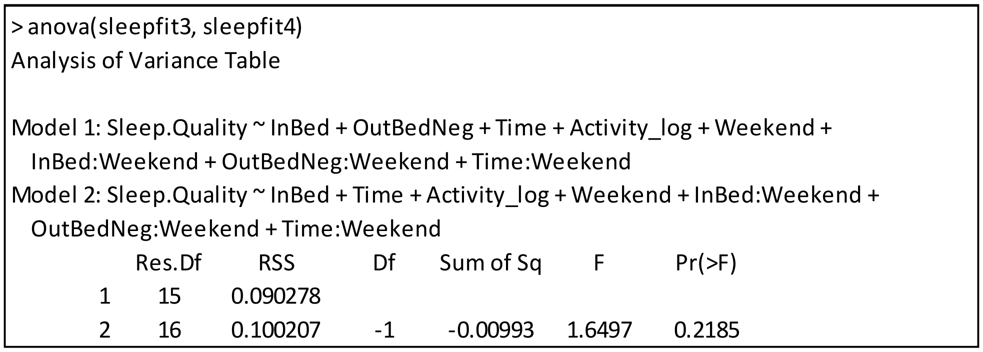


```
> anova(sleepfit2, sleepfit3)
Analysis of Variance Table

Model 1: Sleep.Quality ~ (InBed + OutBedNeg + Time + Activity_log) * Weekend
Model 2: Sleep.Quality ~ InBed + OutBedNeg + Time + Activity_log + Weekend +
  InBed:Weekend + OutBedNeg:Weekend + Time:Weekend
          Res.Df      RSS         Df     Sum of Sq      F        Pr(>F)
      1     14     0.089443
      2     15     0.090278      -1     -0.00083     0.1306     0.7232
```

Moving on, I removed the next least significant predictor. This was the variable OutBedNeg. Removing this one gave the following results. This step actually lowered the Adj R-squared but improved the significance level. The ANOVA showed that there was still no significant differences between models.

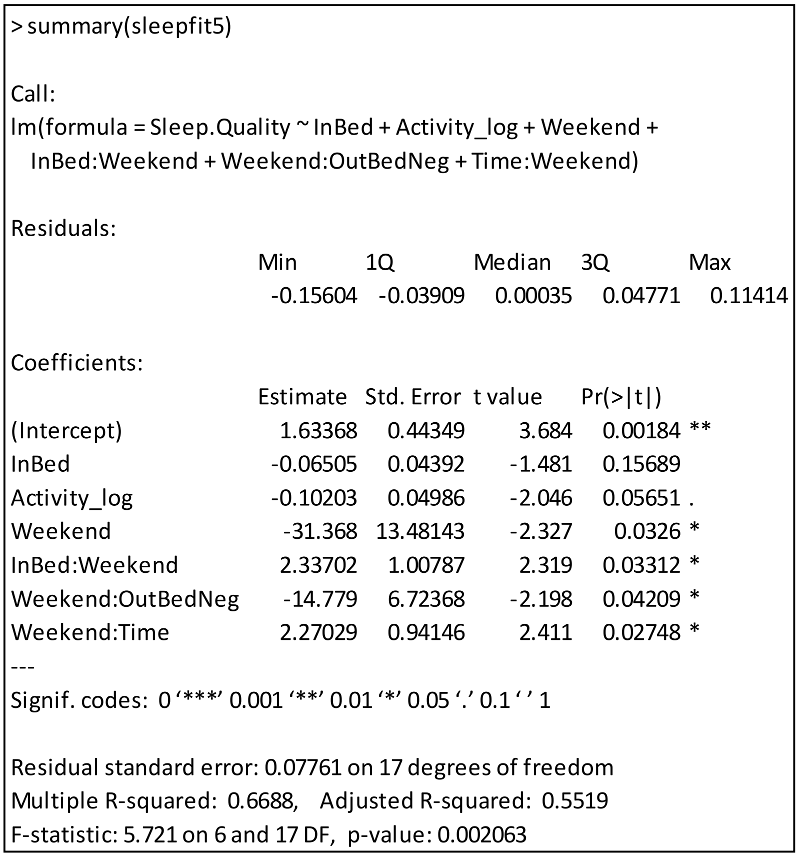


```
> summary(sleepfit4)

Call:
lm(formula = Sleep.Quality ~ InBed + Time + Activity_log + Weekend +
   InBed:Weekend + OutBedNeg:Weekend + Time:Weekend)

Residuals:
                      Min        1Q         Median     3Q         Max
                      -0.15067   -0.04568   -0.00085   0.049882   0.112319

Coefficients:
                      Estimate   Std. Error  t value   Pr(>|t|)
(Intercept)            6.53864    8.28552     0.789     0.4416
InBed                 -0.61375    0.92657    -0.662     0.5172
Time                  -0.55503    0.93617    -0.593     0.5616
Activity_log          -0.09382    0.05269    -1.781     0.094 .
Weekend               -35.7744   15.62675    -2.289     0.036 *
InBed:Weekend          2.84609    1.33916     2.125     0.0495 *
Weekend:OutBedNeg     -14.5145    6.87022    -2.113     0.0507 .
Time:Weekend           2.78425    1.29345     2.153     0.047 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07914 on 16 degrees of freedom
Multiple R-squared:  0.6759,    Adjusted R-squared:  0.5341
F-statistic: 4.767 on 7 and 16 DF,  p-value: 0.004641
```

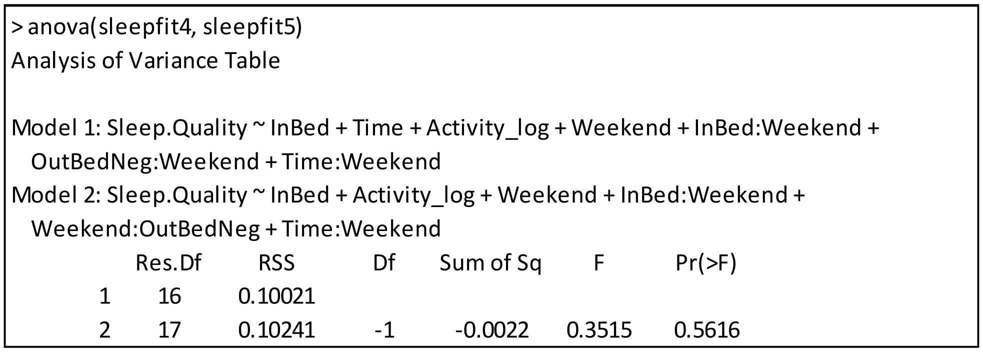


```
> anova(sleepfit3, sleepfit4)
Analysis of Variance Table

Model 1: Sleep.Quality ~ InBed + OutBedNeg + Time + Activity_log + Weekend +
   InBed:Weekend + OutBedNeg:Weekend + Time:Weekend
Model 2: Sleep.Quality ~ InBed + Time + Activity_log + Weekend + InBed:Weekend +
   OutBedNeg:Weekend + Time:Weekend
          Res.Df    RSS        Df    Sum of Sq    F         Pr(>F)
      1    15       0.090278
      2    16       0.100207   -1    -0.00993     1.6497    0.2185
```

Looking at the next variable removal, I realized that now I was getting somewhere: I removed Time from the model and saw the following results. Nearly all the variables now had significance and the Adj R-squared was back up to .5519. This was looking fairly strong.



```
> summary(sleepfit5)

Call:
lm(formula = Sleep.Quality ~ InBed + Activity_log + Weekend +
   InBed:Weekend + Weekend:OutBedNeg + Time:Weekend)

Residuals:
                        Min        1Q         Median     3Q         Max
                        -0.15604   -0.03909   0.00035    0.04771    0.11414

Coefficients:
                        Estimate   Std. Error t value    Pr(>|t|)
(Intercept)              1.63368   0.44349     3.684     0.00184 **
InBed                   -0.06505   0.04392    -1.481     0.15689
Activity_log            -0.10203   0.04986    -2.046     0.05651 .
Weekend                  -31.368   13.48143   -2.327      0.0326 *
InBed:Weekend            2.33702   1.00787     2.319     0.03312 *
Weekend:OutBedNeg        -14.779   6.72368    -2.198     0.04209 *
Weekend:Time             2.27029   0.94146     2.411     0.02748 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07761 on 17 degrees of freedom
Multiple R-squared:  0.6688,    Adjusted R-squared:  0.5519
F-statistic: 5.721 on 6 and 17 DF,  p-value: 0.002063
```



```
> anova(sleepfit4, sleepfit5)
Analysis of Variance Table

Model 1: Sleep.Quality ~ InBed + Time + Activity_log + Weekend + InBed:Weekend +
   OutBedNeg:Weekend + Time:Weekend
Model 2: Sleep.Quality ~ InBed + Activity_log + Weekend + InBed:Weekend +
   Weekend:OutBedNeg + Time:Weekend
           Res.Df     RSS        Df     Sum of Sq     F          Pr(>F)
         1    16      0.10021
         2    17      0.10241    -1     -0.0022      0.3515     0.5616
```

Going one step further, I removed the InBed variable which produced the following results. I concluded there was less variables below the .05 significance level and I also lost some value on the Adj R-squared amount (.0297). This model did not not appear to be any better and indicators show it was actually worse. Because of this, I believed sleepfit5 was the best model to project Sleep.Quality.

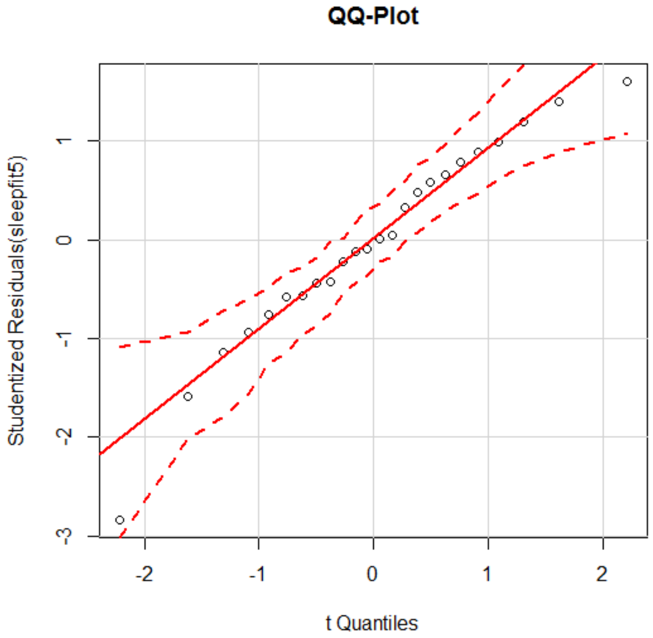
```
> summary(sleepfit6)

Call:
lm(formula = Sleep.Quality ~ Activity_log + Weekend + InBed:Weekend +
   Weekend:OutBedNeg + Weekend:Time)

Residuals:
                       Min        1Q         Median     3Q         Max
                       -0.15791   -0.04543   -0.00111   0.046663   0.119754

Coefficients:
                       Estimate   Std. Error  t value   Pr(>|t|)
(Intercept)             1.28014    0.38596     3.317    0.00384 **
Activity_log           -0.0767     0.04836    -1.586    0.13017
Weekend               -29.476     13.85857    -2.127    0.04751 *
Weekend:InBed           2.14968    1.03252     2.082    0.05189 .
Weekend:OutBedNeg     -13.9628     6.91965    -2.018    0.05877 .
Weekend:Time            2.14354    0.96815     2.214    0.03997 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08015 on 18 degrees of freedom
Multiple R-squared:  0.626,    Adjusted R-squared:  0.5222
F-statistic: 6.027 on 5 and 18 DF,  p-value: 0.001908
```

Regression Diagnostics

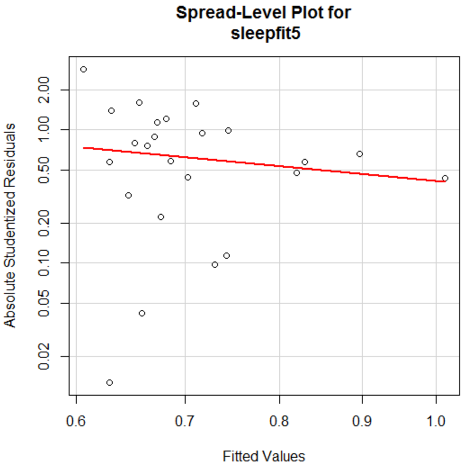
Now that I had chosen the final model, I wanted to make sure the residuals were behaving properly. I specifically mean that the outliers needed to be examined, a normal distribution checked of the residuals as well as the constant variance. Running an outlier test on sleepfit5 gave the following result.

```
> outlierTest(sleepfit5)

No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
   rstudent unadjusted p-value Bonferonni p
24 -2.834573          0.011956      0.28696
```

This shows that for outlying purposes, there were no outliers with a p-value less than .05. For this study, outliers were not an issue. When I looked at Normality, the graph below depicted the qq-plot of the residuals. All values fell within the confidence interval of .95.



For a final check I ran the ncvTest through R which tests the hypothesis of constant error variance against the alternative that the error variance changes with the level of response. The following results were found showing that since the level does not fall below .05 we do not reject the original hypothesis. However, the P-value of .1033 indicates that this is getting close to having non-constant error variance and at the very least should be noted. The graph is a depiction of Studentized Residuals vs. Fitted Values; here too we see that the error variance appeared to have non-constant tendencies.



```
> ncvTest(sleepfit5)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 2.65364    Df = 1     p = 0.1033131
```

Conclusion

My initial hypothesis was that the Sleep.Quality score was directly influenced by Time and Time alone. The first portion was proven wrong immediately when I ran a regression on Sleep.Quality~Time. The results showed an Adj R-Squared of .4071 which hardly supports Time alone being the influence. I continued to build a regression in order to see if Time was the main driver or if other factors were involved in the final Sleep.Quality number. Running an ANOVA test of sleepfit0 and sleepfit5 the additional variables did make a small significant change <.10 to the model. Therefore, proving again my hypothesis incorrect in that Time was the only influencing variable in Sleep.Quality.

```
> anova(sleepfit0, sleepfit5)
Analysis of Variance Table

Model 1: Sleep.Quality ~ Time
Model 2: Sleep.Quality ~ InBed + Activity_log + Weekend + InBed:Weekend +
   Weekend:OutBedNeg + Time:Weekend
          Res.Df     RSS       Df    Sum of Sq      F        Pr(>F)
      1     22     0.17534
      2     17     0.10241     5     0.072933    2.4214    0.07849  .
```

The final regression equation is as follows:

$$Sleep.Quality = 1.63368 - InBed * .06505 - \log(Activity) * .10203 - Weekend * 31.368 + InBed * Weekend * 2.33702 - Outbed^{-\frac{1}{2}} * Weekend * 14.779 + Time * Weekend * 2.27029$$

The first important thing to notice within this equation is that anything with the term Weekend is a dummy variable. This means that any weekday zeroes out these factors. Specifically, on a weekday, the regression is reduced to time InBed minus the log of the Activity. This leads to an interesting observation. Intuitively, I would think that the more activity I did the better sleep I would get but this suggests otherwise. It actually appears that the more activity I did throughout the day the worse sleep I got. It also indicates that the later I went to bed the less sleep quality I received. That one makes more sense.

The weekend portion of the equation is harder to explain, but it essentially shows that in addition to the statements above I must also now include time I got out of bed as well as the total time I slept to get the sleep quality. This makes sense because weekends are not consistent for me and those factors would need to be included in order to give a valid sleep assessment. It looks like the later I got out of bed and also the more time I slept, the better my sleep quality. Well this matches my gut assessment and perhaps I was basing my initial hypothesis more on weekend quality of sleep than what was happening throughout the week.

Based upon these findings, I now believe the app to be true in that it measures Sleep.Quality by more than simply measuring the time I have spent in bed. It is interesting that based on my initial feelings, I believed it to solely be based upon Time, but it would appear that Time only made up ~41% of the quality of sleep. Not only that, but in the final regression I actually removed the Time variable from the equation (at least for weekdays). In addition, by adding the other variables, I was only able to see an improvement to ~55% of explanation of the Sleep.Quality variable. This leads me to believe that there must truly be something else occurring within the apps process to produce the Sleep.Quality number.