Multiple Regression

Zander Bonnet

2024-05-01

```
library(data.table)
set.seed(12345)
adosleep <- data.table(
  SOLacti = rnorm(150, 4.4, 1.3) ^ 2,
  DBAS =  rnorm(150, 72, 26),
  DAS =  rnorm(150, 125, 32),
  Female = rbinom(150, 1, .53),
  Stress = rnorm(150, 32, 11)
)

adosleep[, SSQ :=
           rnorm(
             150,
             (.36 * 3 / 12.5) * SOLacti +
               (.16 * 3 / 26) * DBAS +
               (.18 * 3 / .5) * Female +
               (.20 * 3 / 11) * Stress,
             2.6
           )]

adosleep[, MOOD :=
           rnorm(
             150,
             (-.07 / 12.5) * SOLacti +
               (.29 / 3) * SSQ +
               (.14 / 26) * DBAS +
               (.21 / 32) * DAS +
               (.12 / 32) * SSQ * (DAS - 50) +
               (.44 / .5) * Female +
               (.28 / 11) * Stress,
             2
           )]

adosleep[, Female := factor(Female, levels=0:1, labels = c("Males", "Females"))]

head(adosleep)
```

```
##      SOLacti     DBAS      DAS  Female   Stress        SSQ     MOOD
##        <num>    <num>    <num>  <fctr>    <num>      <num>    <num>
## 1: 26.637856 29.89746 141.7130   Males 34.46721  0.0351776 3.135512
## 2: 28.326939 86.25835 125.3134 Females 40.65050 10.8613493 5.763634
## 3: 18.129761 77.07734 110.9032   Males 27.34301  5.6395828 2.695476
## 4: 14.519557 51.03105 163.3837 Females 27.95713  5.2300021 4.148444
## 5: 26.911751 69.17577 121.2410 Females 12.56278  5.4454510 3.648391
## 6:  4.147973 65.47539 126.2227 Females 29.98108  2.5245995 5.560322
```

```
sum(is.na(adosleep))
```

```
## [1] 0
```

The data has been simulated, and there are no missing values for us to worry about.

```
library(JWileymisc)

d <- testDistribution(adosleep$MOOD)
paste('MOOD', d$distr  ,'-> LLH', d$Distribution$LL,
      'Outlier:', d$extremevalues)
```

```
## [1] "MOOD normal -> LLH -348.964509848302 Outlier: no"
```

```
d <- testDistribution(adosleep$SSQ)
paste('SSQ', d$distr  ,'-> LLH', d$Distribution$LL,
      'Outlier:', d$extremevalues)
```

```
## [1] "SSQ normal -> LLH -377.142976104801 Outlier: no"
```

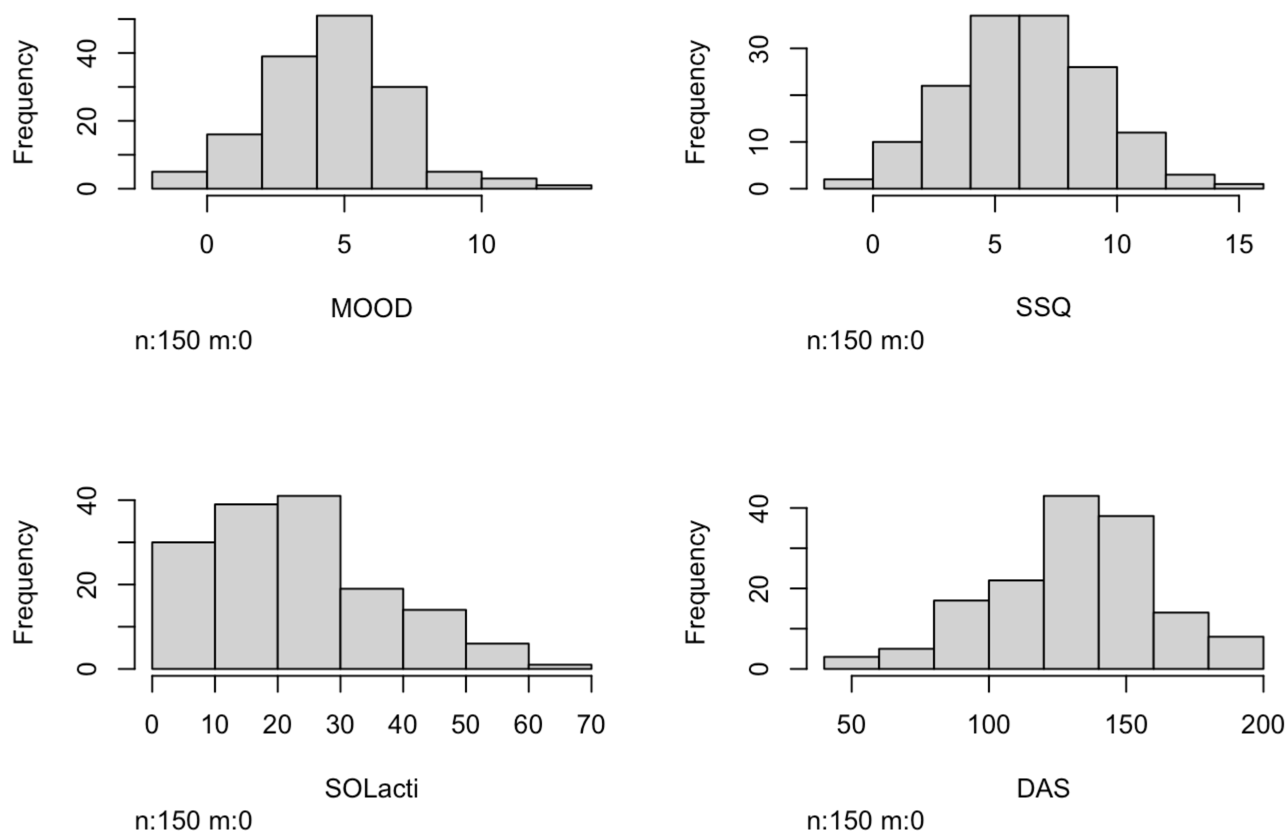
```
d <- testDistribution(adosleep$SOLacti)
paste('SOLacti', d$distr  ,'-> LLH', d$Distribution$LL,
      'Outlier:', d$extremevalues)
```

```
## [1] "SOLacti normal -> LLH -603.880229585631 Outlier: no"
```

```
d <- testDistribution(adosleep$DAS)
paste('DAS', d$distr  ,'-> LLH', d$Distribution$LL,
      'Outlier:', d$extremevalues)
```

```
## [1] "DAS normal -> LLH -724.738607314043 Outlier: no"
```

```
hist(adosleep[,c('MOOD','SSQ','SOLacti','DAS')])
```



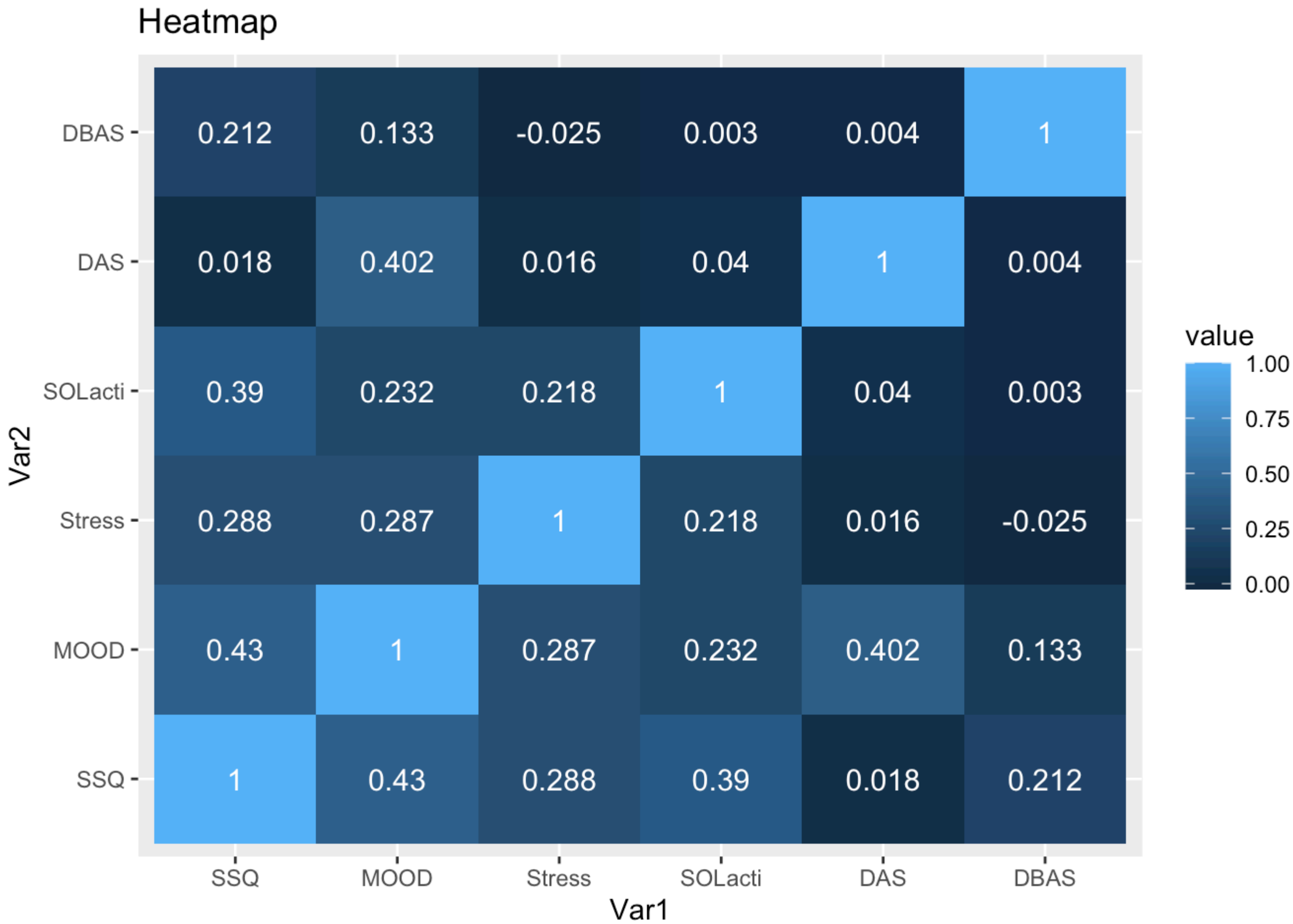
All of the LLH values are sufficiently large to say that the distributions fit a normal distribution. There are also no extreme values in the data. We can also see in the histograms that all of the variables are roughly normal.

```
library(ggplot2)
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
```

```
## The following objects are masked from 'package:data.table':
##
##     dcast, melt
```

```
corr_mat <- round(cor(adosleep[,c('SSQ', 'MOOD', 'Stress', 'SOLacti', 'DAS', 'DBAS')]),
3)
melt_corr_mat <- melt(corr_mat)
plt <- ggplot(data = melt_corr_mat, aes(x = Var1, y = Var2, fill = value))
plt <- plt + geom_tile()
plt <- plt + geom_text(aes(Var2, Var1, label = value), color = "white", size = 4)
plt <- plt + labs(title = 'Heatmap')
plt
```




```
egltable(adosleep)
```

```
##               M (SD)/N (%)
##       <char>        <char>
## 1:   SOLacti  23.33 (13.60)
## 2:      DBAS  72.10 (23.88)
## 3:       DAS 130.57 (30.45)
## 4:    Female
## 5:     Males     67 (44.7%)
## 6:   Females     83 (55.3%)
## 7:    Stress  32.84 (10.92)
## 8:       SSQ   6.18 (3.00)
## 9:      MOOD   4.53 (2.49)
```

```
stan_lacti <- as.vector(scale(adosleep$SOLacti))
stan_dbas <- as.vector(scale(adosleep$DBAS))
stan_das <- as.vector(scale(adosleep$DAS))
stan_stress <- as.vector(scale(adosleep$Stress))

standardized <- data.frame(
  SOLacti = stan_lacti,
  DBAS = stan_dbas,
  DAS = stan_das,
  Female = adosleep$Female,
  STRESS = stan_stress,
  SSQ = adosleep$SSQ,
  MOOD =adosleep$MOOD
  )

head(standardized)
```

```
##      SOLacti       DBAS        DAS  Female     STRESS        SSQ     MOOD
## 1  0.2429460 -1.7671947  0.3658192   Males  0.1491628  0.0351776 3.135512
## 2  0.3671180  0.5931063 -0.1728046 Females  0.7152324 10.8613493 5.763634
## 3 -0.3825213  0.2086207 -0.6460895   Males -0.5030461  5.6395828 2.695476
## 4 -0.6479233 -0.8821548  1.0775621 Females -0.4468246  5.2300021 4.148444
## 5  0.2630813 -0.1222839 -0.3065568 Females -1.8561515  5.4454510 3.648391
## 6 -1.4103839 -0.2772498 -0.1429394 Females -0.2615348  2.5245995 5.560322
```

```
mod1 <- lm(MOOD~Female + STRESS, data = standardized)
summary(mod1)
```

```
##
## Call:
## lm(formula = MOOD ~ Female + STRESS, data = standardized)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.455 -1.395  0.204  1.449  7.534
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.9195     0.2853  13.740  < 2e-16 ***
## FemaleFemales   1.0970     0.3836   2.860 0.004859 **
## STRESS          0.7337     0.1914   3.834 0.000186 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.334 on 147 degrees of freedom
## Multiple R-squared:  0.1306, Adjusted R-squared:  0.1187
## F-statistic: 11.04 on 2 and 147 DF,  p-value: 3.422e-05
```

Both Female and stress are significant in the model. The model has a very small r-squared, so we need more facotrs to explain the variance in mood.

```r
mod2 <- lm(MOOD~., data = standardized)
summary(mod2)
```

```
##
## Call:
## lm(formula = MOOD ~ ., data = standardized)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8272 -0.9979  0.0788  1.2248  4.5272
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.10647    0.43526   4.840 3.33e-06 ***
## SOLacti        0.09713    0.17164   0.566 0.572376
## DBAS           0.16493    0.16116   1.023 0.307841
## DAS            1.02807    0.15708   6.545 9.93e-10 ***
## FemaleFemales  1.24759    0.31502   3.960 0.000118 ***
## STRESS         0.46076    0.16517   2.790 0.005997 **
## SSQ            0.28008    0.05986   4.679 6.63e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.908 on 143 degrees of freedom
## Multiple R-squared:  0.4348, Adjusted R-squared:  0.4111
## F-statistic: 18.34 on 6 and 143 DF,  p-value: 9.986e-16
```

```r
mod2 <-  lm(MOOD~. -  SOLacti - DBAS, data = standardized)
summary(mod2)
```

```
##
## Call:
## lm(formula = MOOD ~ . - SOLacti - DBAS, data = standardized)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8534 -1.0726  0.0517  1.2704  4.6355
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.9544     0.4054   4.820 3.58e-06 ***
## DAS             1.0314     0.1566   6.587 7.74e-10 ***
## FemaleFemales   1.2467     0.3141   3.969 0.000113 ***
## STRESS          0.4563     0.1630   2.799 0.005817 **
## SSQ             0.3048     0.0543   5.613 9.79e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.903 on 145 degrees of freedom
## Multiple R-squared:  0.4297, Adjusted R-squared:  0.414
## F-statistic: 27.32 on 4 and 145 DF,  p-value: < 2.2e-16
```

The R-Squared value increased significantly, but there were two factors that are not significant, DBAS and SOLacti, so I removed them and this again improved the R-Squared and significance of the model.

```
mod3 <-  lm(MOOD~. + (SSQ*DAS) -  SOLacti - DBAS, data = standardized)
summary(mod3)
```

```
##
## Call:
## lm(formula = MOOD ~ . + (SSQ * DAS) - SOLacti - DBAS, data = standardized)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.856 -1.089  0.144  1.278  4.068
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.92870    0.40178   4.800 3.92e-06 ***
## DAS            0.43403    0.34304   1.265 0.207827
## FemaleFemales  1.24120    0.31115   3.989 0.000105 ***
## STRESS         0.49778    0.16283   3.057 0.002666 **
## SSQ            0.30851    0.05381   5.733 5.58e-08 ***
## DAS:SSQ        0.10671    0.05466   1.952 0.052847 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.885 on 144 degrees of freedom
## Multiple R-squared:  0.4444, Adjusted R-squared:  0.4252
## F-statistic: 23.04 on 5 and 144 DF,  p-value: < 2.2e-16
```

```r
mod3 <- lm(MOOD~. + (SSQ*DAS) -  SOLacti - DBAS - DAS, data = standardized)
summary(mod3)
```

```
##
## Call:
## lm(formula = MOOD ~ . + (SSQ * DAS) - SOLacti - DBAS - DAS, data = standardized)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7960 -1.1179  0.0765  1.1311  4.3503
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.91965    0.40255   4.769 4.47e-06 ***
## FemaleFemales  1.22212    0.31143   3.924 0.000134 ***
## STRESS         0.52244    0.16200   3.225 0.001557 **
## SSQ            0.31116    0.05388   5.775 4.52e-08 ***
## DAS:SSQ        0.16840    0.02476   6.800 2.54e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.889 on 145 degrees of freedom
## Multiple R-squared:  0.4383, Adjusted R-squared:  0.4228
## F-statistic: 28.28 on 4 and 145 DF,  p-value: < 2.2e-16
```

Adding the interaction of the DAS and SSQ made DAS an insignificant factor. Removing it made the model more significant and only slightly lowered the R-Squared.

```r
library(texreg)
```

```
## Version:  1.39.3
## Date:     2023-11-09
## Author:   Philip Leifeld (University of Essex)
##
## Consider submitting praise using the praise or praise_interactive functions.
## Please cite the JSS article in your publications -- see citation("texreg").
```

```r
screenreg(list(mod1,mod2,mod3))
```

```
##
## ==================================================
##                 Model 1     Model 2     Model 3
## --------------------------------------------------
## (Intercept)      3.92 ***    1.95 ***    1.92 ***
##                 (0.29)      (0.41)      (0.40)
## FemaleFemales    1.10 **     1.25 ***    1.22 ***
##                 (0.38)      (0.31)      (0.31)
## STRESS           0.73 ***    0.46 **     0.52 **
##                 (0.19)      (0.16)      (0.16)
## DAS                          1.03 ***
##                             (0.16)
## SSQ                          0.30 ***    0.31 ***
##                             (0.05)      (0.05)
## DAS:SSQ                                  0.17 ***
##                                         (0.02)
## --------------------------------------------------
## R^2              0.13        0.43        0.44
## Adj. R^2         0.12        0.41        0.42
## Num. obs.      150         150         150
## ==================================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

Model 1 does not include DAS or SSQ

Model 2 finds that both DAS and SSQ are significant predictors at the .001 level. Meaning that they are very significant predictors of mood.

Model 3 did not find DAS to be significant, but SSQ is still significant. The interaction between DAS and SSQ is also significant.

Model 3 has a slightly larger impact of SSQ as well in model 2 SSQ's constant is .3 and in model 3 it is .31.

```r
library(regclass)
```

```
## Loading required package: bestglm
```

```
## Loading required package: leaps
```

```
## Loading required package: VGAM
```

```
## Loading required package: stats4
```

```
## Loading required package: splines
```

```
## Loading required package: rpart
```

```
## Loading required package: randomForest
```

```
## randomForest 4.7-1.1
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
## Important regclass change from 1.3:
## All functions that had a . in the name now have an _
## all.correlations -> all_correlations, cor.demo -> cor_demo, etc.
```

```
print('mod1')
```

```
## [1] "mod1"
```

```
print(VIF(mod1))
```

```
##  Female  STRESS
## 1.00146 1.00146
```

```
print(testDistribution(mod1$residuals)$Distribution$LL)
```

```
## 'log Lik.' -338.4719 (df=2)
```

```
print('mod2')
```

```
## [1] "mod2"
```

```
print(VIF(mod2))
```

```
##      DAS   Female   STRESS      SSQ
## 1.008485 1.009891 1.092778 1.091457
```

```
print(testDistribution(mod2$residuals)$Distribution$LL)
```

```
## 'log Lik.' -306.8403 (df=2)
```

```
print('mod3')
```

```
## [1] "mod3"
```

```
print(VIF(mod3))
```

```
##   Female   STRESS      SSQ  DAS:SSQ
## 1.007601 1.095757 1.091192 1.008555
```

```
print(testDistribution(mod3$residuals)$Distribution$LL)
```

```
## 'log Lik.' -305.7102 (df=2)
```

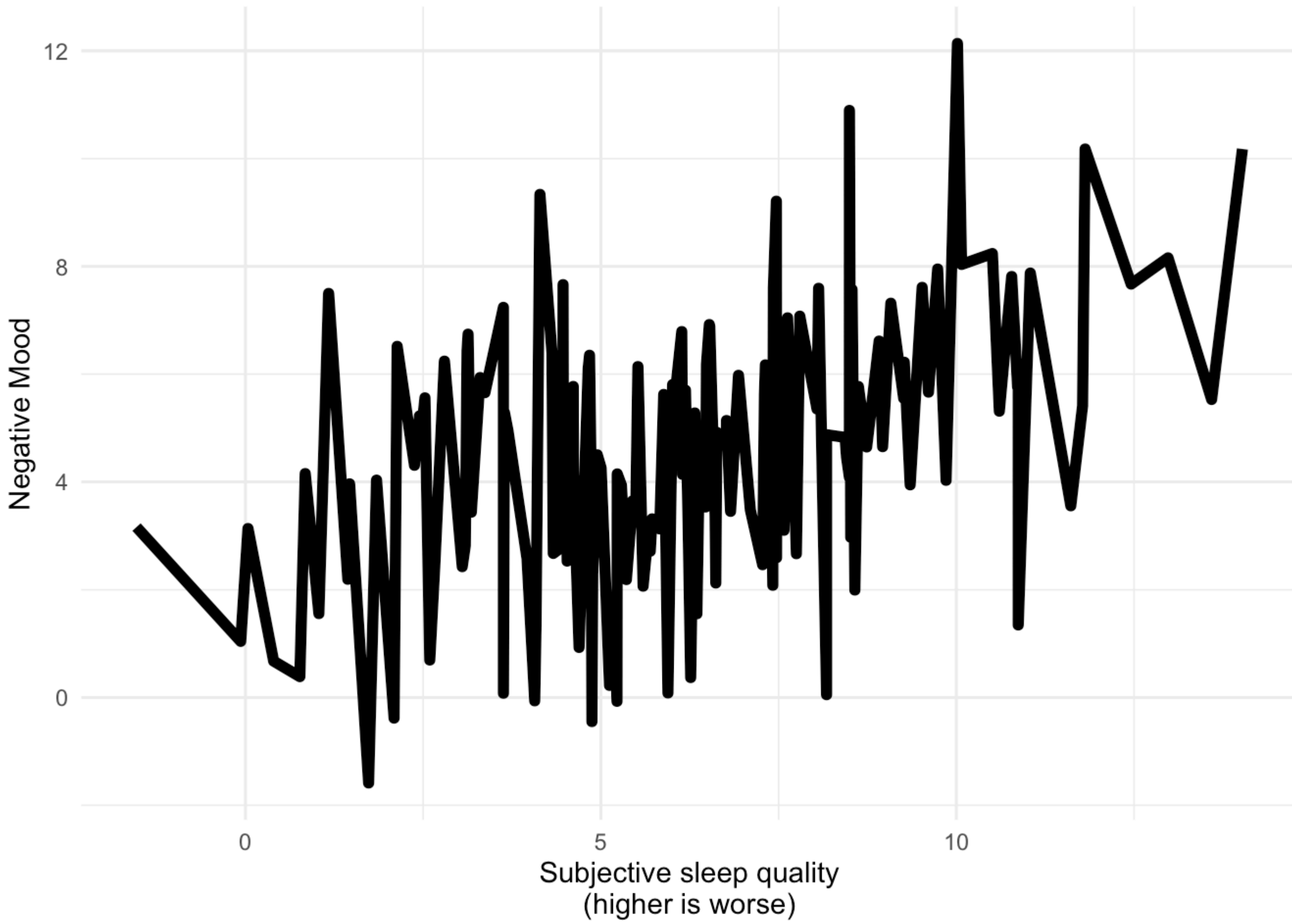
All models have residuals that fit the normality assumption and have very low VIF's. This shows that the residuals are homoscedastic, and the assumptions of the model are met.

```
mod4 <- lm(MOOD~. + (SSQ*DAS) -  SOLacti - DBAS - DAS, data = adosleep)
summary(mod4)
```

```
##
## Call:
## lm(formula = MOOD ~ . + (SSQ * DAS) - SOLacti - DBAS - DAS, data = adosleep)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7960 -1.1179  0.0765  1.1311  4.3503
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.3490615  0.5599174   0.623 0.533991
## FemaleFemales  1.2221162  0.3114295   3.924 0.000134 ***
## Stress         0.0478286  0.0148307   3.225 0.001557 **
## SSQ           -0.4110235  0.1189889  -3.454 0.000724 ***
## DAS:SSQ        0.0055308  0.0008133   6.800 2.54e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.889 on 145 degrees of freedom
## Multiple R-squared:  0.4383, Adjusted R-squared:  0.4228
## F-statistic: 28.28 on 4 and 145 DF,  p-value: < 2.2e-16
```

The model of raw data is significant but the intercept is not found to be significant. This means that our model cannot be accurate. The model does not know what the baseline for the origin of the data is, so we cannot use this model for small predictions. The R-Squared is similar to the standardized models, so it does explain the variance in a similar way, but fails to know where to start the estimate.

```
(ggplot(standardized, aes(SSQ, MOOD))
  +geom_line(linewidth = 2)
  +scale_x_continuous("Subjective sleep quality\n(higher is worse)")
  +ylab("Negative Mood")
  +theme_minimal()
  +theme(legend.position.inside = c(.85, .15),legend.key.width = unit(2, "cm")))
```



This plot shows that as SSQ increases the negative mood also increases on average.

** linetype = DAS caused an error message as DAS is a continuous variable

** the theme cowplot is also no longer supported

We were able to address the original objective of what factors can we use to predict the mood of an individual. It was found that SOLacti was not significant in predicting mood, but SSQ was very significant in the prediction of mood. It was also found that SSQ is a very significant factor in predicting the the mood of an individual. This was found by conducting linear analysis. In this analysis we discovered what factors were significant in predicting mood, and how significant they are. We found that SSQ and being a Female are the most significant factors in predicting mood, and in a model with the interaction between SSQ and DAS that is the most significant factor in predicting mood.