MLR Analysis Technical Report

Josh Houlding

2024-02-06

Task 2

Create a synthetic dataset using the code provided in the “DSC-520 Multiple Regression – Dataset.” Note that this snippet of R code is designed to generate data that approximates the data described in the article.

```
# Generate synthetic dataset
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"
))]
# Display the synthetic dataset
adosleep
```

```
##       SOLacti     DBAS       DAS  Female   Stress        SSQ     MOOD
##   1: 26.63786 29.89746 141.71303   Males 34.46721  0.0351776 3.135512
##   2: 28.32694 86.25835 125.31340 Females 40.65050 10.8613493 5.763634
##   3: 18.12976 77.07734 110.90316   Males 27.34301  5.6395828 2.695476
##   4: 14.51956 51.03105 163.38367 Females 27.95713  5.2300021 4.148444
##   5: 26.91175 69.17577 121.24101 Females 12.56278  5.4454510 3.648391
##  ---
## 146: 26.03265 76.77716 141.02786   Males 27.77196  7.6259684 7.044688
## 147:  5.70503 16.09193  95.89703 Females 16.28171  1.0325290 1.557178
## 148: 30.30006 55.80287  54.68933 Females 42.59759  4.8613652 4.189277
## 149: 30.96719 52.09858 139.03345 Females 38.75160 12.9784500 8.158078
## 150: 20.97913 84.07204  92.05873 Females 27.94590  5.8835742 5.623150
```

Task 3

Instructions: Inspect the core variables to assess their distribution (using the testdistr function) and identify outliers: MOOD, SSQ, SOLacti, DAS. Plot the distribution of each variable.

Now that we have our data, we will inspect the distributions and identify outliers for the following variables: `SOLacti`, `DAS`, `SSQ`, and `MOOD`.

Variable `SOLacti`

```
# Assess the distribution of 'SOLacti'
dist <- testDistribution(adosleep$SOLacti)
type <- dist$distr
paste("Distribution type:", type)
```

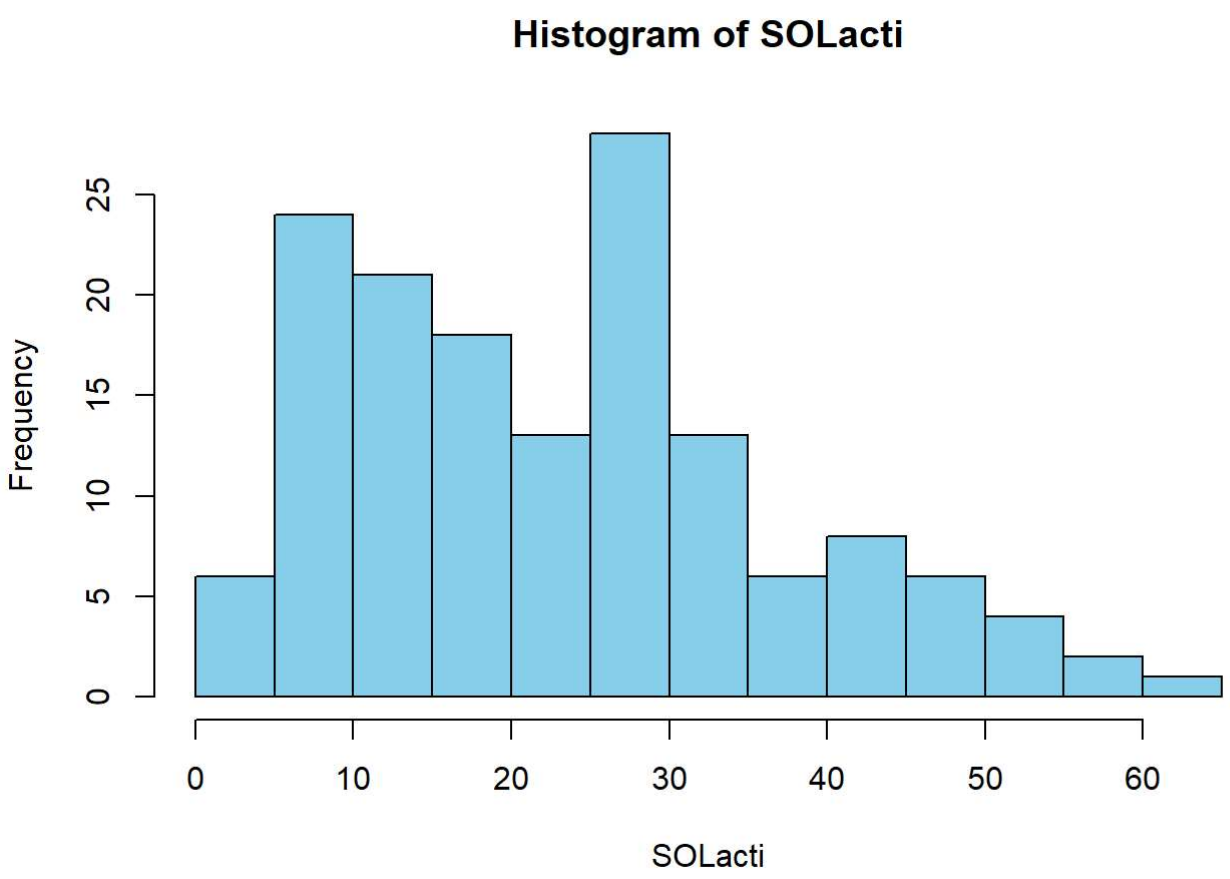
```
## [1] "Distribution type: normal"
```

We see from the `$distr` variable that the distribution of `SOLacti` is roughly normal.

```
# Identify outliers of 'SOLacti'
Q1 <- quantile(adosleep$SOLacti, 0.25)
Q3 <- quantile(adosleep$SOLacti, 0.75)
IQR <- Q3 - Q1
outliers <- which(adosleep$SOLacti < (Q1 - 1.5 * IQR) | adosleep$SOLacti > (Q3 + 1.5 * IQR))
outliers <- adosleep[outliers, ]
head(outliers)
```

```
##     SOLacti     DBAS      DAS  Female   Stress      SSQ     MOOD
## 1: 58.06812 78.89844 107.8661   Males 42.42037 13.59092 5.530014
## 2: 61.66215 88.81154 137.3947 Females 52.73468 12.45876 7.670935
```

```
# Plot distribution of 'SOLacti'
hist(adosleep$SOLacti, main = "Histogram of SOLacti", xlab = "SOLacti", col = "skyblue", border
= "black")
```



Variable `DAS`

```
# Assess the distribution of 'DAS'
dist <- testDistribution(adosleep$DAS)
type <- dist$distr
paste("Distribution type:", type)
```

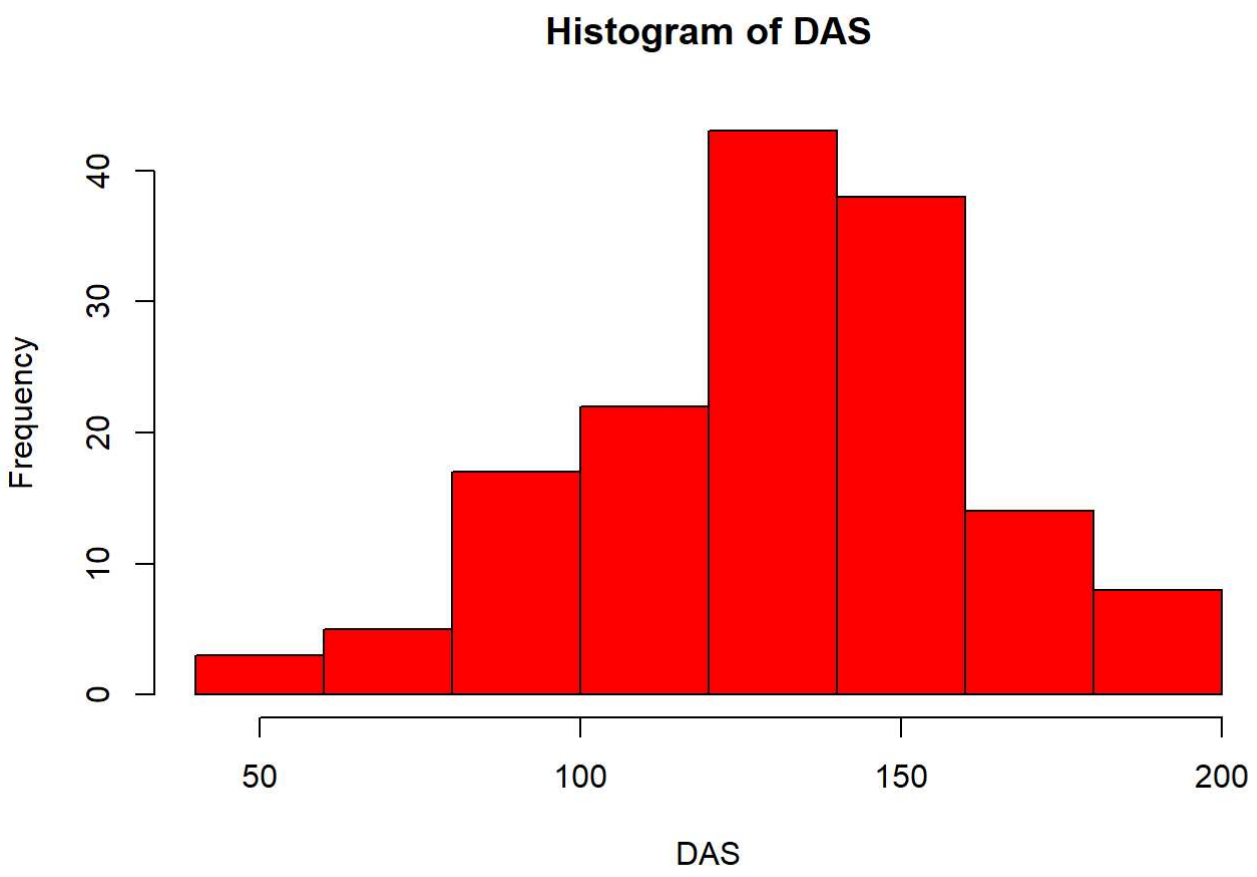
```
## [1] "Distribution type: normal"
```

Just as with the previous variable, `DAS` is normally distributed.

```
# Identify outliers of 'DAS'
Q1 <- quantile(adosleep$DAS, 0.25)
Q3 <- quantile(adosleep$DAS, 0.75)
IQR <- Q3 - Q1
outliers <- which(adosleep$DAS < (Q1 - 1.5 * IQR) | adosleep$DAS > (Q3 + 1.5 * IQR))
outliers <- adosleep[outliers, ]
head(outliers)
```

```
##      SOLacti     DBAS      DAS  Female   Stress      SSQ       MOOD
## 1:  5.396225 27.73053 55.04971   Males 38.26460 4.524677  2.5315833
## 2: 26.606449 60.38386 42.38063   Males 18.72021 2.090213 -0.3822468
## 3: 30.300058 55.80287 54.68933 Females 42.59759 4.861365  4.1892768
```

```
# Plot distribution of 'DAS'
hist(adosleep$DAS, main = "Histogram of DAS", xlab = "DAS", col = "red", border = "black")
```



Variable `SSQ`

```
# Assess the distribution of 'SSQ'
dist <- testDistribution(adosleep$SSQ)
type <- dist$distr
paste("Distribution type:", type)
```

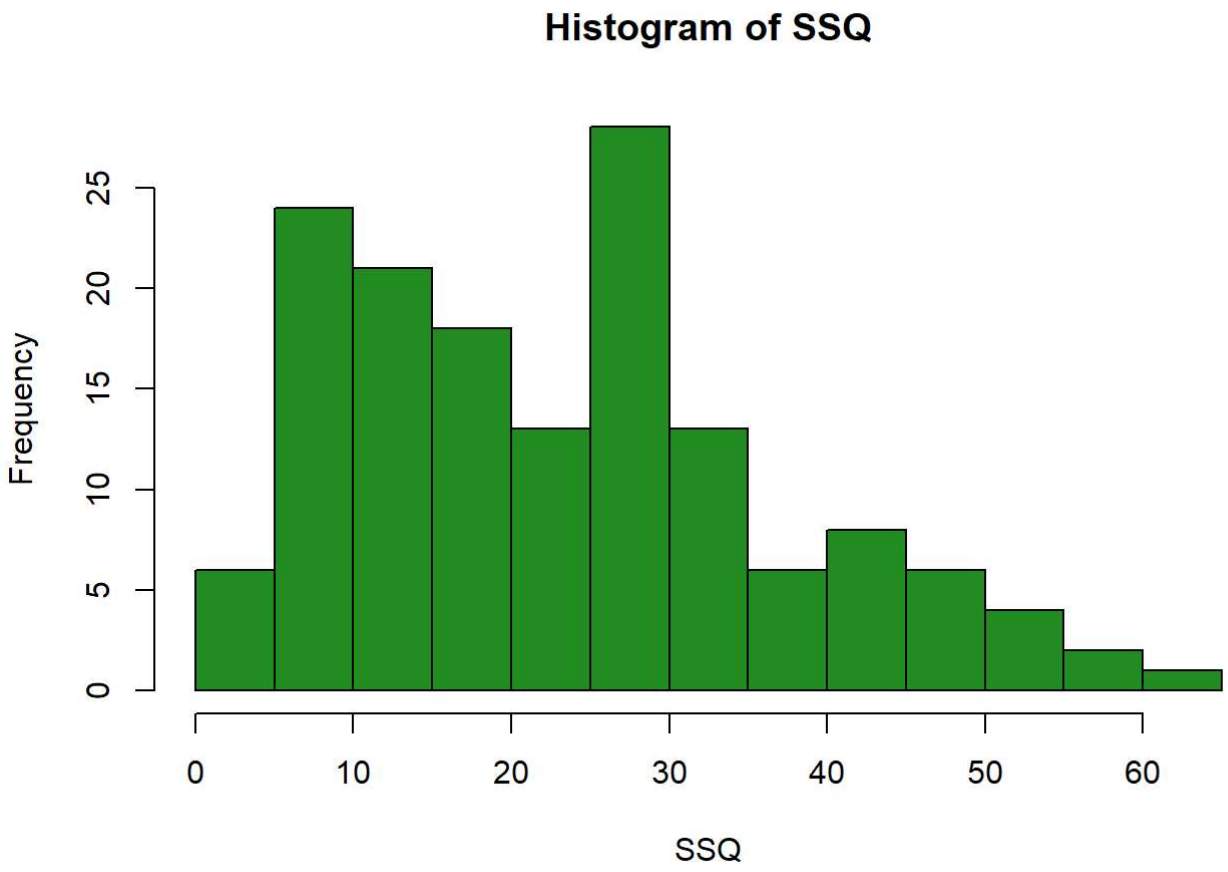
```
## [1] "Distribution type: normal"
```

`SSQ` also follows a roughly normal distribution.

```
# Identify outliers of 'SSQ'
Q1 <- quantile(adosleep$SSQ, 0.25)
Q3 <- quantile(adosleep$SSQ, 0.75)
IQR <- Q3 - Q1
outliers <- which(adosleep$SSQ < (Q1 - 1.5 * IQR) | adosleep$SSQ > (Q3 + 1.5 * IQR))
outliers <- adosleep[outliers, ]
head(outliers)
```

```
##     SOLacti     DBAS       DAS  Female   Stress       SSQ      MOOD
## 1: 45.51697 84.50470 154.86215 Females 26.81655 14.024905 10.177421
## 2:  7.04748 49.65948  66.98932 Females 14.66991 -1.533016  3.162297
```

```
# Plot distribution of 'SSQ'
hist(adosleep$SOLacti, main = "Histogram of SSQ", xlab = "SSQ", col = "forestgreen", border = "b
lack")
```



Variable `MOOD`

```
# Assess the distribution of 'MOOD'
dist <- testDistribution(adosleep$MOOD)
type <- dist$distr
paste("Distribution type:", type)
```

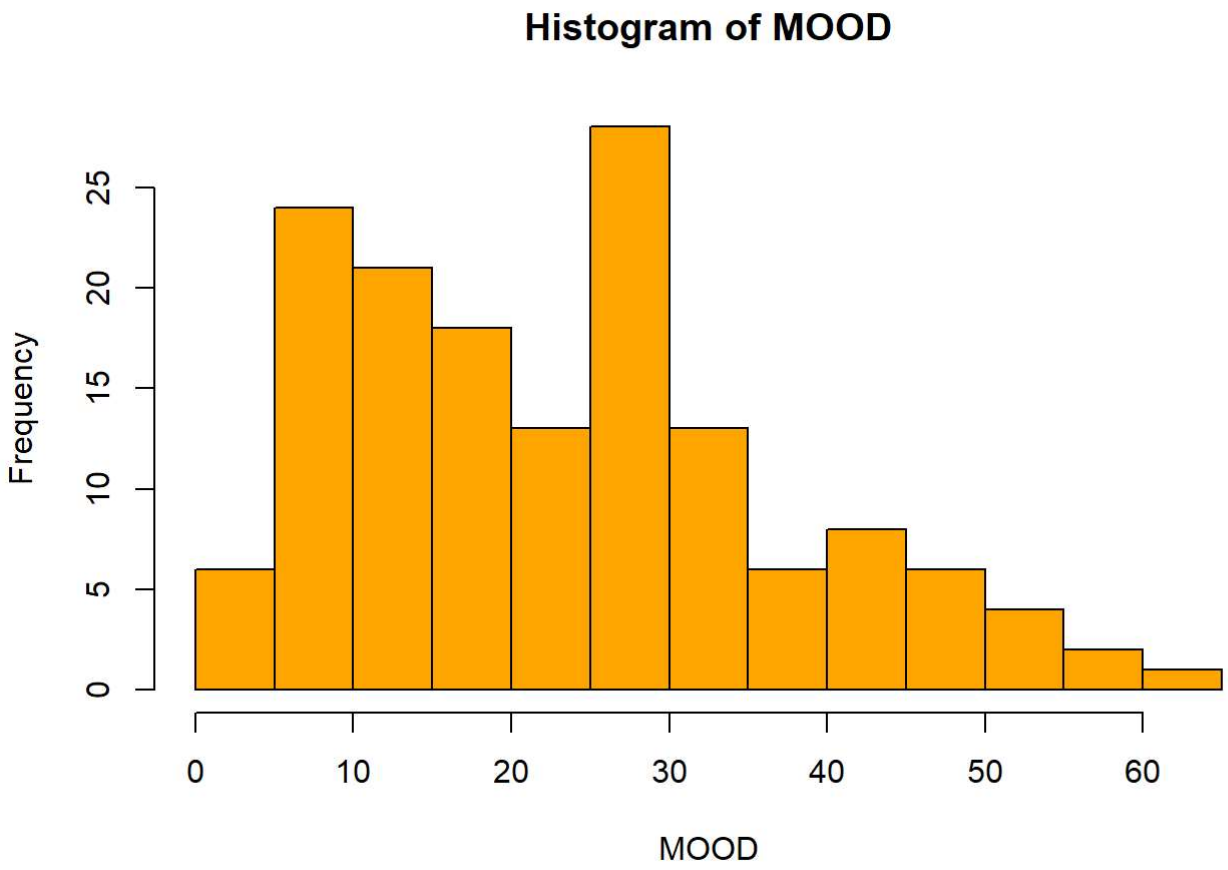
```
## [1] "Distribution type: normal"
```

`MOOD` is also normally distributed.

```
# Identify outliers of 'MOOD'
Q1 <- quantile(adosleep$MOOD, 0.25)
Q3 <- quantile(adosleep$MOOD, 0.75)
IQR <- Q3 - Q1
outliers <- which(adosleep$MOOD < (Q1 - 1.5 * IQR) | adosleep$MOOD > (Q3 + 1.5 * IQR))
outliers <- adosleep[outliers, ]
head(outliers)
```

```
##     SOLacti     DBAS      DAS  Female   Stress      SSQ     MOOD
## 1: 20.86599 97.27421 150.5658 Females 52.64846 10.01529 12.13602
```

```
# Plot distribution of 'MOOD'
hist(adosleep$SOLacti, main = "Histogram of MOOD", xlab = "MOOD", col = "orange", border = "blac
k")
```



Task 4

Instructions: Examine the bivariate correlations between study variables: SSQ, MOOD, Stress, SOLacti, DAS, DBAS. Plot a heatmap depicting the correlations table (use the plot function and appropriate theme).

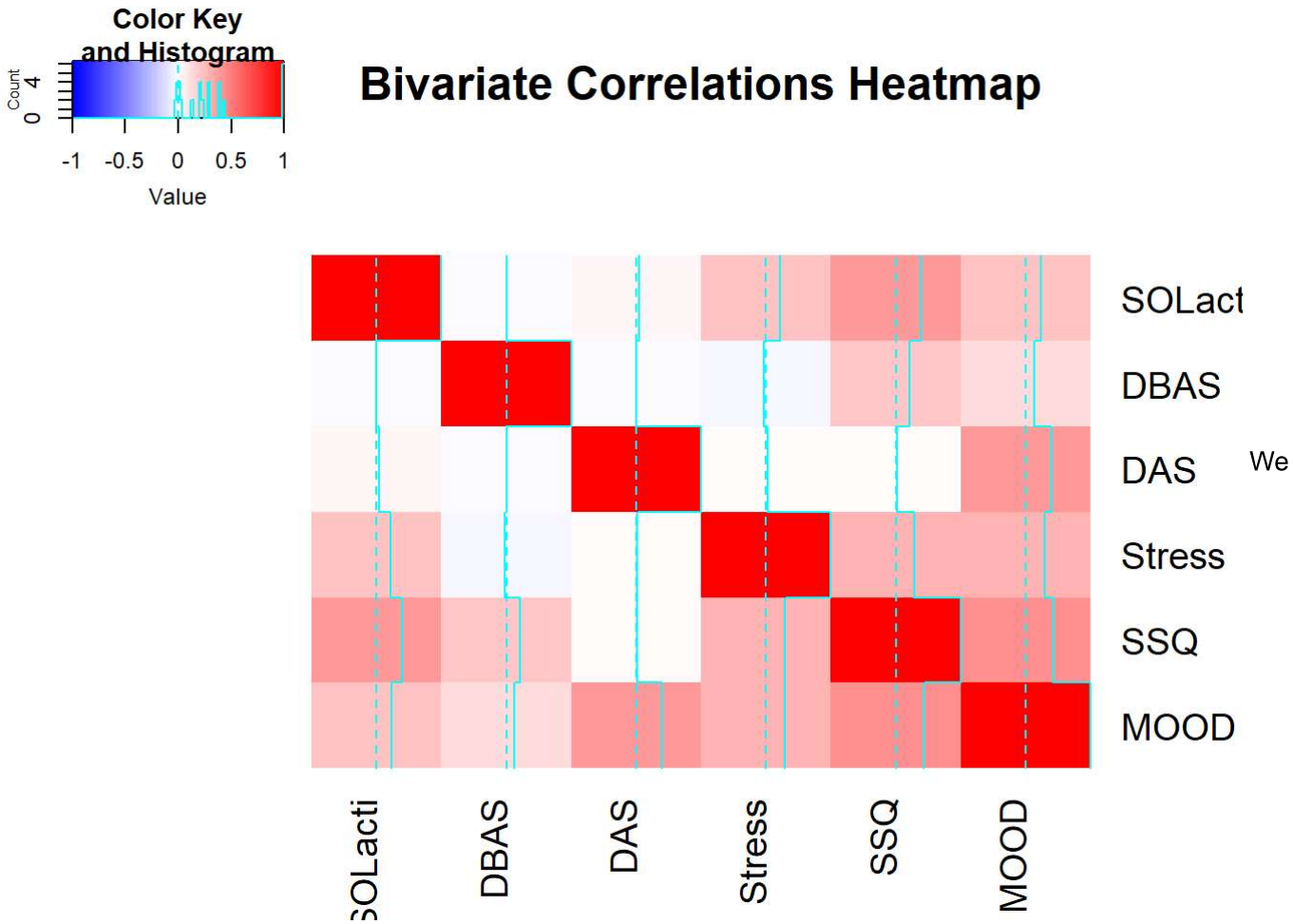
Next, we will examine the correlations between the study variables `SOLacti`, `DBAS`, `DAS`, `Stress`, `SSQ`, and `MOOD`. This can be done using a heatmap.

```
# Get subset of data containing only necessary variables
study_variables <- adosleep[, c("SOLacti", "DBAS", "DAS", "Stress", "SSQ", "MOOD")]

# Generate correlation matrix
corr_matrix <- cor(study_variables)
corr_matrix <- round(corr_matrix, digits=2)
corr_matrix <- round(corr_matrix, digits=3)
print(corr_matrix)
```

```
##         SOLacti  DBAS  DAS Stress  SSQ MOOD
## SOLacti    1.00  0.00 0.04   0.22 0.39 0.23
## DBAS       0.00  1.00 0.00  -0.03 0.21 0.13
## DAS        0.04  0.00 1.00   0.02 0.02 0.40
## Stress     0.22 -0.03 0.02   1.00 0.29 0.29
## SSQ        0.39  0.21 0.02   0.29 1.00 0.43
## MOOD       0.23  0.13 0.40   0.29 0.43 1.00
```

```
# Generate heatmap
heatmap.2(corr_matrix, main="Bivariate Correlations Heatmap", col=colorRampPalette(c("blue","whi
te","red"))(100), margins=c(5,5), notecol="black", Rowv=NULL, Colv=NULL, symkey=FALSE, key=TRUE)
```



We see that the strongest correlations occur between `MOOD` and `DAS`, `MOOD` and `SSQ`, `MOOD` and `Stress`, and `SSQ` and `Stress`.

Task 5

Create a basic table of graphs of descriptive statistics using the egltable function. Standardize the predictors to get standardized estimates (as in the article) using `as.vector(scale(variable_name))`.

Predictors: `SOLacti`, `DBAS`, `DAS`, `Stress`.

```
# Display descriptive statistics
egltable(data = adosleep,
         vars = c("SOLacti", "DBAS", "DAS", "Stress", "SSQ", "MOOD")
)
```

```
##                    M (SD)
## 1: SOLacti  23.33 (13.60)
## 2:    DBAS  72.10 (23.88)
## 3:     DAS 130.57 (30.45)
## 4:  Stress  32.84 (10.92)
## 5:     SSQ    6.18 (3.00)
## 6:    MOOD    4.53 (2.49)
```

```
# Standardize predictors
standardized_adosleep <- adosleep
standardized_adosleep$SOLacti <- as.vector(scale(standardized_adosleep$SOLacti))
standardized_adosleep$DBAS <- as.vector(scale(standardized_adosleep$DBAS))
standardized_adosleep$DAS <- as.vector(scale(standardized_adosleep$DAS))
standardized_adosleep$Stress <- as.vector(scale(standardized_adosleep$SOLacti))

# Display descriptive statistics of standardized predictors to confirm successful standardizatio
n
egltable(data = standardized_adosleep,
         vars = c("SOLacti", "DBAS", "DAS", "Stress")
)
```

```
##                   M (SD)
## 1: SOLacti  0.00 (1.00)
## 2:    DBAS -0.00 (1.00)
## 3:     DAS -0.00 (1.00)
## 4:  Stress  0.00 (1.00)
```

Task 6

Fit three different models and compare them. * Model 1: Just the covariates * Model 2: Model 1 + main constructs of interest without interactions * Model 3: Model 2 + add the hypothesized interaction between subjective sleep quality and global dysfunctional beliefs

```
# Fit model 1 (covariates only)
model1 <- lm(MOOD ~ SOLacti + DBAS + Female + Stress, data = standardized_adosleep)

# Fit model 2 (main construct of interest DAS added)
model2 <- lm(MOOD ~ DAS + SOLacti + DBAS + Female + Stress, data = standardized_adosleep)

# Fit model 3 (interaction term between SSQ and DAS added)
model3 <- lm(MOOD ~ DAS + SOLacti + DBAS + Female + Stress + SSQ:DAS, data = standardized_adosle
ep)
```

Task 7

Combine the results of the three models into one table using the screenreg() function. Note the asterisk to the right of the threshold p-values and the errors in parentheses.

Note: Attempting to install the `screenreg` package led to this warning:

```
Warning in install.packages :   package ‘screenreg’ is not available for this version of R
```

Thus, I opted to use a package with similar functionality, `stargazer`.

```
# Display regression results side-by-side
thresholds <- c(0.05, 0.01, 0.001)
regression_results <- stargazer(model1, model2, model3, type="text", p.auto = FALSE, p.threshold
s = thresholds)
```

```
##
## ================================================================================================
##                                              Dependent variable:
##                     ---------------------------------------------------------------------------
##                                                      MOOD
##                               (1)                     (2)                      (3)
## ------------------------------------------------------------------------------------------------
## DAS                                                1.031***                 0.653*
##                                                     (0.176)                 (0.386)
##
## SOLacti                      0.568***               0.525***                 0.523***
##                              (0.194)                (0.175)                  (0.175)
##
## DBAS                          0.333*                0.330*                   0.350**
##                              (0.194)                (0.175)                  (0.176)
##
## FemaleFemales                1.029***               1.216***                 1.211***
##                              (0.388)                (0.352)                  (0.351)
##
## Stress
##
##
## DAS:SSQ                                                                      0.068
##                                                                             (0.061)
##
## Constant                     3.957***               3.854***                 3.853***
##                              (0.289)                (0.261)                  (0.261)
##
## ------------------------------------------------------------------------------------------------
## Observations                   150                    150                      150
## R2                            0.114                  0.284                    0.290
## Adjusted R2                   0.096                  0.264                    0.266
## Residual Std. Error    2.364 (df = 146)       2.132 (df = 145)         2.131 (df = 144)
## F Statistic         6.256*** (df = 3; 146) 14.394*** (df = 4; 145) 11.774*** (df = 5; 144)
## ================================================================================================
## Note:                                                              *p<0.1; **p<0.05; ***p<0.01
##
## =================
## 0.050 0.010 0.001
## -----------------
```

Task 8

Since higher scores on subjective sleep quality indicate poorer sleep quality, assess which model (if any) shows that overall worse sleep quality and overall dysfunctional attitudes are significantly associated with more negative mood ($p < .001$).

IE: Which model (if any) shows that `SOLacti` and `DBAS` are significantly associated with `MOOD` ?

We see from the results that `SOLActi` is significantly associated with `MOOD` in all 3 models ($P < 0.001$), but `DBAS` does not cross the $P < 0.001$ threshold in any of the models.

Task 9

Ensure that the models are appropriate. Check the variance inflation factors (using the vif function) and the distribution of residuals (using the testdistr function). Describe and interpret your findings.

```
# Check for aliased coefficients in model 1
print(alias(model1))
```

```
## Model :
## MOOD ~ SOLacti + DBAS + Female + Stress
##
## Complete :
##        (Intercept) SOLacti DBAS FemaleFemales
## Stress 0           1       0    0
```

```
# Check for aliased coefficients in model 2
print(alias(model2))
```

```
## Model :
## MOOD ~ DAS + SOLacti + DBAS + Female + Stress
##
## Complete :
##        (Intercept) DAS SOLacti DBAS FemaleFemales
## Stress 0           0   1       0    0
```

```
# Check for aliased coefficients in model 3
print(alias(model3))
```

```
## Model :
## MOOD ~ DAS + SOLacti + DBAS + Female + Stress + SSQ:DAS
##
## Complete :
##        (Intercept) DAS SOLacti DBAS FemaleFemales DAS:SSQ
## Stress 0           0   1       0    0             0
```

It appears that `SOLacti` has an aliased coefficient in all 3 models, thus preventing VIF calculation, so we will remove it and check VIF.

```
# Refit models without 'SOLacti'
model1_noalias <- lm(MOOD ~ DBAS + Female + Stress, data = standardized_adosleep)
model2_noalias <- lm(MOOD ~ DAS + DBAS + Female + Stress, data = standardized_adosleep)
model3_noalias <- lm(MOOD ~ DAS + DBAS + Female + Stress + SSQ:DAS, data = standardized_adoslee
p)
```

```
# Check model 1 VIF
print(car::vif(model1_noalias))
```

```
##     DBAS   Female   Stress
## 1.000094 1.000321 1.000251
```

The VIF values for all 3 predictors are very close to 1, indicating no strong correlation with other predictors.

```
# Check model 1 residual distribution
dist <- testDistribution(residuals(model1))
type <- dist$distr
paste("Residual distribution type:", type)
```

```
## [1] "Residual distribution type: normal"
```

The residuals for model 1 are normally distributed, so this assumption of linear regression is not violated.

```
# Check model 2 VIF
print(car::vif(model2_noalias))
```

```
##      DAS     DBAS   Female   Stress
## 1.009850 1.000104 1.008522 1.002001
```

As in the last model, all predictors have VIFs close to 1, indicating no multicollinearity.

```
# Check model 2 residual distribution
dist <- testDistribution(residuals(model2))
type <- dist$distr
paste("Residual distribution type:", type)
```

```
## [1] "Residual distribution type: normal"
```

```
# Check model 3 VIF
print(car::vif(model3_noalias))
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

```
##      DAS     DBAS   Female   Stress  DAS:SSQ
## 4.877905 1.010875 1.008684 1.002096 4.874467
```

In model 3, we see that `DAS` and `DAS:SSQ` have high VIFs, though this is because the latter involves the former, so we would expect this. Thus, there are no unexpected problems.

```
# Check model 3 residual distribution
dist <- testDistribution(residuals(model3))
type <- dist$distr
paste("Residual distribution type:", type)
```

```
## [1] "Residual distribution type: normal"
```

All 3 models have normally-distributed residuals, meaning they are likely to be appropriate. Further tests for homoscedasticity and independence of residuals would further bolster the credibility of the models.

Task 10

Refit the model on raw (i.e., nonstandardized data).

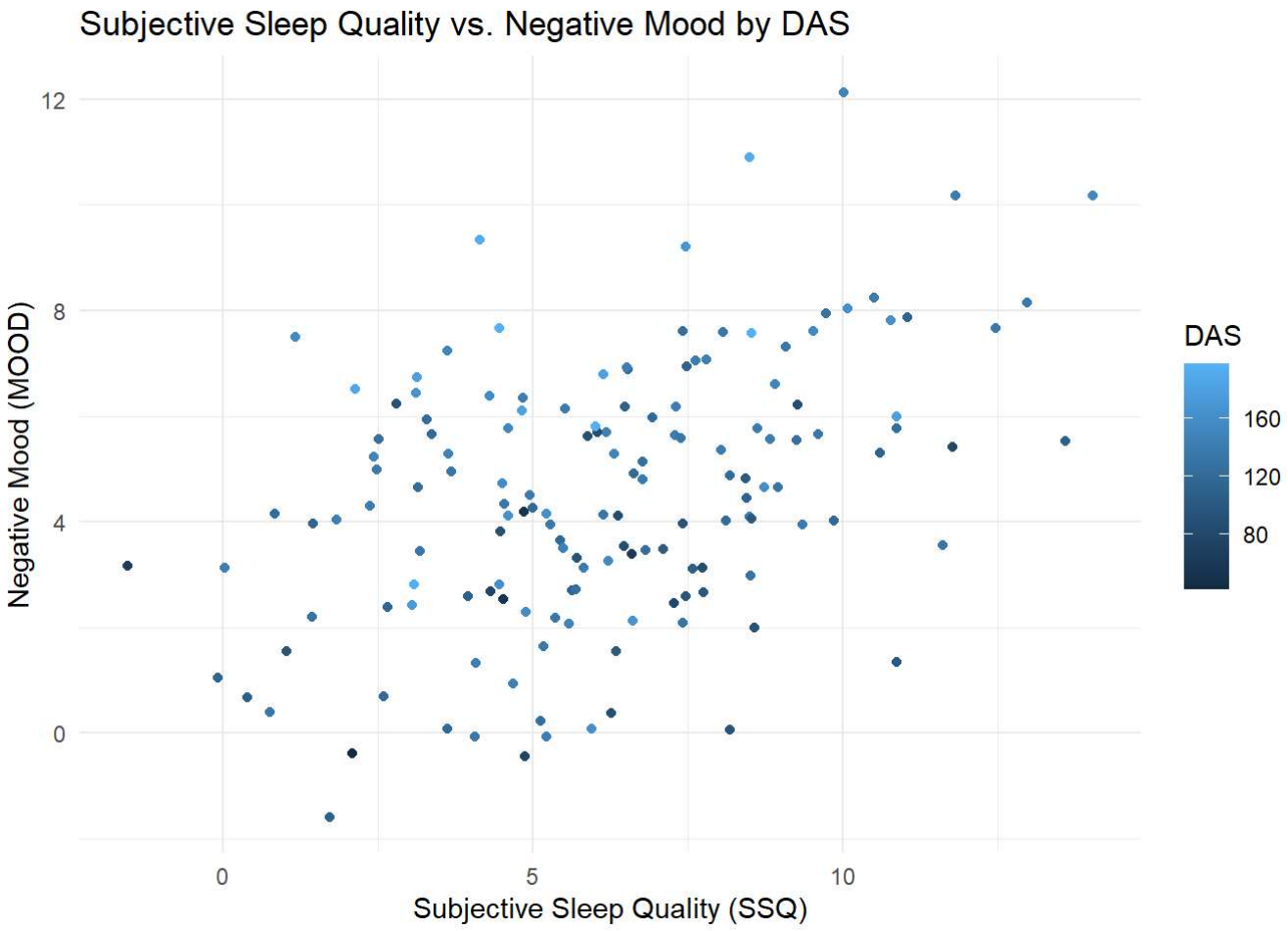
```
# Refit model 3 on nonstandardized data
model3 <- lm(MOOD ~ DAS + SOLacti + DBAS + Female + Stress + SSQ:DAS, data = adosleep)
summary(model3)
```

```
##
## Call:
## lm(formula = MOOD ~ DAS + SOLacti + DBAS + Female + Stress +
##     SSQ:DAS, data = adosleep)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8362 -1.0120  0.0646  1.1927  4.2557
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.8583252  1.0240707  -2.791 0.005970 **
## DAS            0.0211712  0.0056684   3.735 0.000270 ***
## SOLacti        0.0054505  0.0124748   0.437 0.662825
## DBAS           0.0074084  0.0066203   1.119 0.264996
## FemaleFemales  1.2441343  0.3112540   3.997 0.000102 ***
## Stress         0.0441916  0.0147991   2.986 0.003325 **
## DAS:SSQ        0.0022518  0.0004424   5.090 1.11e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.885 on 143 degrees of freedom
## Multiple R-squared:  0.4483, Adjusted R-squared:  0.4251
## F-statistic: 19.36 on 6 and 143 DF,  p-value: < 2.2e-16
```

Task 11

Use ggplot to visualize the relations between subjective sleep quality and negative mood, and the relation of poor sleep quality and mood of vulnerable adolescents with higher levels of dysfunctional attitudes. If you completed the previous steps correctly, the following R code should plot the graph:

```
# Plot SSQ vs. MOOD with DAS color
ggplot(adosleep, aes(x = SSQ, y = MOOD, color=DAS)) +
  geom_point() +
  labs(
    x = "Subjective Sleep Quality (SSQ)",
    y = "Negative Mood (MOOD)",
    title = "Subjective Sleep Quality vs. Negative Mood by DAS"
  ) +
  theme_minimal()
```



We see a fairly strong positive linear relationship between subjective sleep quality and negative mood, so it is safe to say that this data indicates that an individual who reports a higher subjective sleep quality will be more likely to suffer from low mood. Additionally, we see that lighter-colored points are more common at higher negative mood levels, indicating that `DAS` has a noticeable effect on an individual's negative mood.

Task 12

Review the original objective of the analysis and ensure you were able to address the objectives, produce answers, and back up your claims with relevant calculations.

Study objectives: School terms and vacations represent naturally occurring periods of restricted and extended sleep opportunities. A cognitive model of the relationships among objective sleep, subjective sleep, and negative mood was tested across these periods, with sleep-specific (i.e., dysfunctional beliefs and attitudes about sleep)

and global (i.e., dysfunctional attitudes) cognitive vulnerabilities as moderators.

Objective checklist:

- Modeled relationships between objective sleep, subjective sleep, and negative mood? ✓
- Included dysfunctional attitudes and cognitive vulnerabilities in the analysis? ✓

All study objectives have been addressed, and all claims have been backed up with calculations and visualizations.