STAT2170 Assignment

Ernie Leung (47234083)

2025-05-22

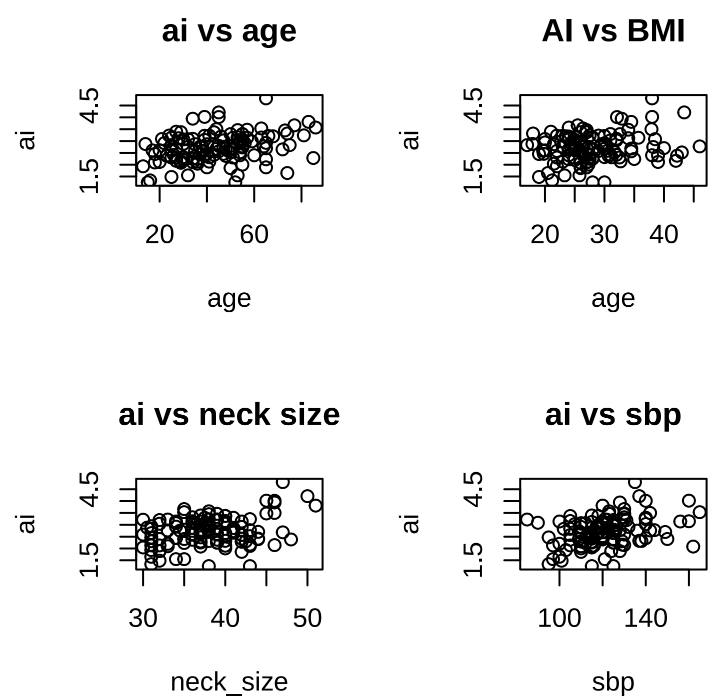
Contents

1 Question 1

```r
sleep <- read.csv("sleep.csv")
```

1.1 (a) Plot & Correlation Matrix of Data

```r
# Plots
par(mfrow = c(2, 2))
plot(sleep$age, sleep$ai, main = "ai vs age", xlab = "age", ylab = "ai")
plot(sleep$bmi, sleep$ai, main = "AI vs BMI", xlab = "age", ylab = "ai")
plot(sleep$neck_size, sleep$ai, main = "ai vs neck size", xlab = "neck_size", ylab = "ai")
plot(sleep$sbp, sleep$ai, main = "ai vs sbp", xlab = "sbp", ylab = "ai")
```




```r
# Correlation matrix
cor(sleep)
```

```
##                  age        bmi  neck_size       sbp        ai
## age       1.00000000 0.02192595 0.08255638 0.2012049 0.3172935
## bmi       0.02192595 1.00000000 0.67087306 0.3099451 0.1944877
## neck_size 0.08255638 0.67087306 1.00000000 0.2545203 0.3296021
## sbp       0.20120485 0.30994514 0.25452032 1.0000000 0.3464153
## ai        0.31729345 0.19448769 0.32960209 0.3464153 1.0000000
```

1.2 (b) Fitting Model & 95% Confidence Interval

```r
# Fit the full linear regression model
fm <- lm(ai ~ age + bmi + neck_size + sbp, data = sleep)
summary(fm)
```

```
##
## Call:
## lm(formula = ai ~ age + bmi + neck_size + sbp, data = sleep)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.67136 -0.32269  0.01491  0.35778  1.47595
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.159406   0.518207  -0.308  0.75893
## age          0.008789   0.002964   2.965  0.00367 **
## bmi         -0.009852   0.011312  -0.871  0.38557
## neck_size    0.040627   0.014208   2.859  0.00503 **
## sbp          0.010218   0.003555   2.875  0.00481 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5417 on 117 degrees of freedom
## Multiple R-squared:  0.2471, Adjusted R-squared:  0.2213
## F-statistic: 9.598 on 4 and 117 DF,  p-value: 9.54e-07
```

Next, we extract a **95% confidence interval** to estimate impact of `neck_size` on `ai`.

```r
confint(fm, "neck_size", level = 0.95)
```

```
##                2.5 %     97.5 %
## neck_size 0.01248892 0.06876571
```

The coefficient for neck_size tells us how the arousal index (ai) is expected to change when neck size increases by 1 cm, in this case it is statistically significant (p = 0.003), which indicates that neck size effects the arousal index.

1.3 (c) Mathematical Model

The multiple linear regression model for this study is given by:

$$\mathrm{ai}_i = \beta_0 + \beta_1 \cdot \mathrm{age}_i + \beta_2 \cdot \mathrm{bmi}_i + \beta_3 \cdot \mathrm{neck\_size}_i + \beta_4 \cdot \mathrm{sbp}_i + \varepsilon_i$$

Where:

- ai_i: logged arousal index (ai) for the response variable i^{th} patient
- β_0: intercept, the expected arousal index when all predictors are 0

1.4 (c) Hypothesis for the Overall F-Test

To test whether the predictors (age, bmi, neck_size, sbp) are associated with the response variable `ai`, we hypothesize:

- **Null Hypothesis H_0**:

  $$\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

  Meaningg none of the predictors have a linear relationship with `ai`.

- **Alternative Hypothesis H_1**:

  $$\text{At least one } \beta_j \neq 0 \quad \text{for } j = 1, 2, 3, 4$$

  Meaning at least one predictor is linearly related to `ai`.

1.5 (c) ANOVA Table for the Full Model

```
fm <- lm(ai ~ age + bmi + neck_size + sbp, data = sleep)

anova(fm)
```

```
## Analysis of Variance Table
##
## Response: ai
##            Df Sum Sq Mean Sq F value    Pr(>F)
## age         1  4.591  4.5911 15.6440 0.0001314 ***
## bmi         1  1.605  1.6045  5.4674 0.0210727 *
## neck_size   1  2.646  2.6460  9.0159 0.0032731 **
## sbp         1  2.425  2.4251  8.2633 0.0048069 **
## Residuals 117 34.337  0.2935
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1.6 (c) Null Distribution of the Test Statistic

The test statistic used in the overall regression F test follows an F distribution under the null hypothesis:

$$F \sim F_{4,122-4-1} = F_{4,117}$$

1.7 (c) P-Value

We compute the p-value associated with the overall F test using the `pf()` function:

```
# F-Statistic and degrees of freedom
f_stat <- summary(fm)$fstatistic
fvalue <- f_stat[1]
df1 <- f_stat[2]
df2 <- f_stat[3]

# P-Value
pf(fvalue, df1, df2, lower.tail = FALSE)
```

```
##      value
## 9.5398e-07
```

1.8 (c) Conclusion

Statistical Conclusion:

Since the p-value is very small (typically < 0.05), we reject the null hypothesis (H_0). This result suggests that at least one of the predictors (age, BMI, neck size, or sbp) has a significant relationship with the arousal index (ai).

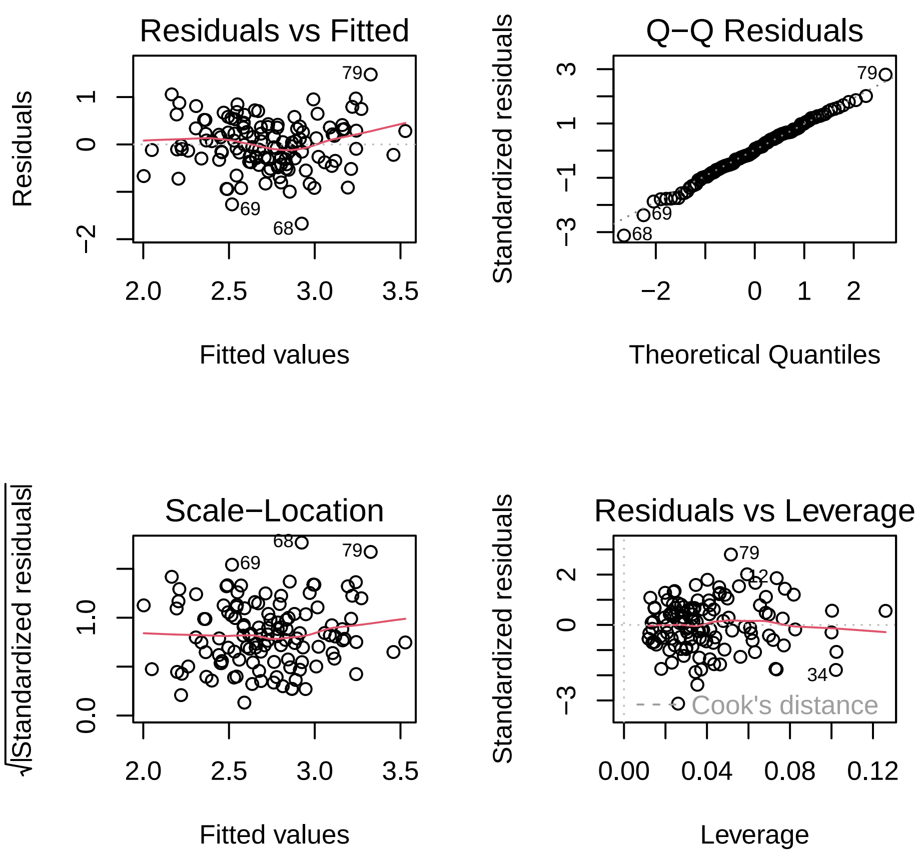
Contextual Conclusion:

There is sufficient evidence to conclude that one or more of the variables are significant predictors of the arousal index in patients suspected of having Obstructive Sleep Apnoea.

1.9 (d) Model Validation

We check the standard assumptions of the linear regression model:

```r
# Plots
par(mfrow = c(2, 2))
plot(fm)
```



Based on these plots, the assumptions (linearity, normality) of linear regression appear to be reasonably satisfied. Therefore, the full regression model is appropriate for explaining variation in the arousal index.

1.10 (e) R^2 Value

We extract the R^2 value from the full model to assess how well the model explains variation in the response variable.

```
# R-squared value
summary(fm)$r.squared
```

```
## [1] 0.2470586
```

1.11 (f) Finding the best multiple regression model

We compare models by examining their adjusted R^2 values and the significance of individual predictors.

```
fm <- lm(ai ~ age + bmi + neck_size + sbp, data = sleep)
adjr2_full <- summary(fm)$adj.r.squared
reduced_model <- lm(ai ~ age + bmi + neck_size, data = sleep)
reduced_adjr2 <- summary(reduced_model)$adj.r.squared

c(
  "Full model Adjusted R^2" = adjr2_full,
  "Reduced model Adjusted R^2" = reduced_adjr2
)
```

```
##    Full model Adjusted R^2 Reduced model Adjusted R^2
##                  0.2213170                  0.1733864
```

```
final_model <- if (reduced_adjr2 > adjr2_full) reduced_model else fm
print(summary(final_model))
```

```
##
## Call:
## lm(formula = ai ~ age + bmi + neck_size + sbp, data = sleep)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.67136 -0.32269  0.01491  0.35778  1.47595
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.159406   0.518207  -0.308  0.75893
## age          0.008789   0.002964   2.965  0.00367 **
## bmi         -0.009852   0.011312  -0.871  0.38557
## neck_size    0.040627   0.014208   2.859  0.00503 **
## sbp          0.010218   0.003555   2.875  0.00481 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.5417 on 117 degrees of freedom
## Multiple R-squared:  0.2471, Adjusted R-squared:  0.2213
## F-statistic: 9.598 on 4 and 117 DF,  p-value: 9.54e-07
```

1.12 (g) Comments on R^2 and Adjusted R^2

```r
r2_full <- summary(fm)$r.squared
adjr2_full <- summary(fm)$adj.r.squared
r2_final <- summary(final_model)$r.squared
adjr2_final <- summary(final_model)$adj.r.squared

c(
  "Full model R^2" = r2_full,
  "Full model Adjusted R^2" = adjr2_full,
  "Final model R^2" = r2_final,
  "Final model Adjusted R^2" = adjr2_final
)
```

```
##           Full model R^2  Full model Adjusted R^2          Final model R^2
##                0.2470586                0.2213170                0.2470586
## Final model Adjusted R^2
##                0.2213170
```

This small decrease in both values shows that `sbp` contributes very minor to the model, and is insigificant. The adjusted R^2 which takes into the factor of model complexity decreased only slightly suggesting that the Adjusted R^2 model still performs reasonably well. Therefore, adjusted R^2 provides a better basis for comparing models with different numbers of predictors.

2 Question 2

```r
energy <- read.csv("energy.csv")
```

2.1 (a) Balanced vs Unbalanced Design

A balanced design means that each combination of the factor (`range` & `factor`) have the **same amout of observations**. Whilst the latter does not.

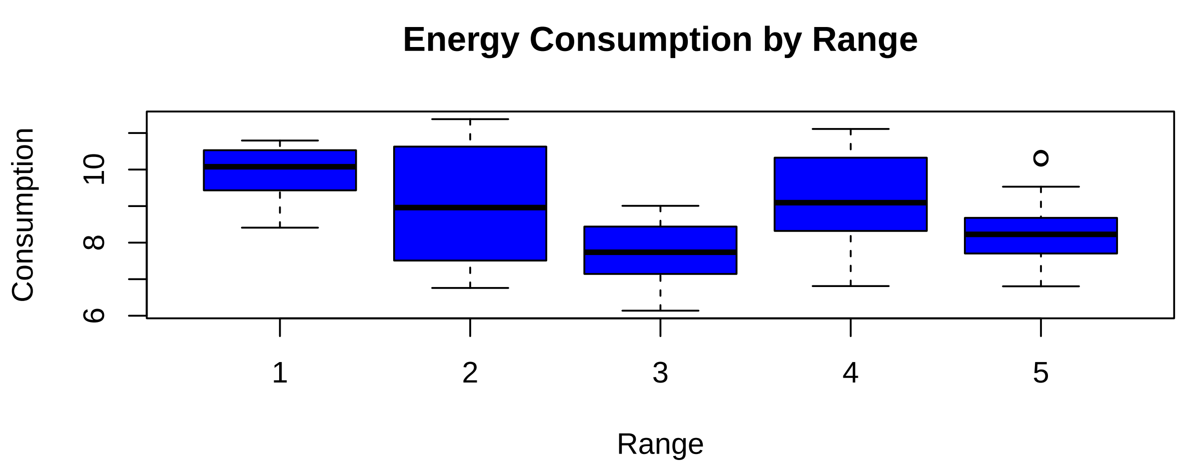
```r
# Table to check the num of observations per group
table(energy$range, energy$menu)
```

```
##
##     1 2
##   1 8 8
##   2 8 8
##   3 8 8
##   4 8 8
##   5 8 8
```

2.2 (b) Preliminary Graphs

We construct two plots to examine how consumption varies by range and menu. ### Plot 1: Boxplot of Consumption by Range

```r
boxplot(consumption ~ range, data = energy,
        main = "Energy Consumption by Range",
        xlab = "Range",
        ylab = "Consumption",
        col = "blue")
```



2.3 (c) Full Interaction Model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

Where:

- Y_{ijk}: observed response (energy consumption) for the k^{th} replicate under the i^{th} range and j^{th} menu
- μ: overall mean energy consumption
- α_i: effect of the i^{th} **range** (for $i = 1, 2, 3, 4, 5$)
- β_j: effect of the j^{th} **menu** (for $j = 1, 2$)
- $(\alpha\beta)_{ij}$: **interaction effect** between the i^{th} range and j^{th} menu
- ε_{ijk}: random error term
$$\varepsilon_{ijk} \sim N(0, \sigma^2)$$

This model tests the main effects of `range` and `menu` and the interaction between them.

2.4 (d) Analysing the Data

We use a two-way ANOVA model with `range` and `menu` to see if there is a significant effect on energy consumption.

2.4.1 Hypotheses

We test the following hypotheses:

- **Main effect of range**

  $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_5 = 0$
  $H_1 :$ At least one $\alpha_i \neq 0$

- **Main effect of menu**

  $H_0 : \beta_1 = \beta_2 = 0$
  $H_1 :$ At least one $\beta_j \neq 0$

- **Interaction effect**

  $H_0 : (\alpha\beta)_{ij} = 0$ for all i, j
  $H_1 :$ At least one interaction term $\neq 0$

2.4.2 Performing the Analysis

```
energy$range <- factor(energy$range) #specify factors
energy$menu <- factor(energy$menu)

aov(consumption ~ range * menu, data = energy)
```

```
## Call:
##    aov(formula = consumption ~ range * menu, data = energy)
##
## Terms:
##                    range     menu range:menu Residuals
## Sum of Squares  44.58041 53.77218   10.78096  30.66651
## Deg. of Freedom        4        1          4        70
##
## Residual standard error: 0.661886
## Estimated effects may be unbalanced
```

2.4.3 Conclusion

The ANOVA results show that the interaction between range and menu is **NOT significant** (where $p > 0.05$) so we must interpret the effects, of which:

- The `range` has a significant effect on energy consumption ($p < 0.01$), while the `menu` does not ($p > 0.05$).

Therefore, energy consumption depends on the range used, regardless of menu.