OSA Data Analyst

Jaemean Shine

2025-05-20

Question 1

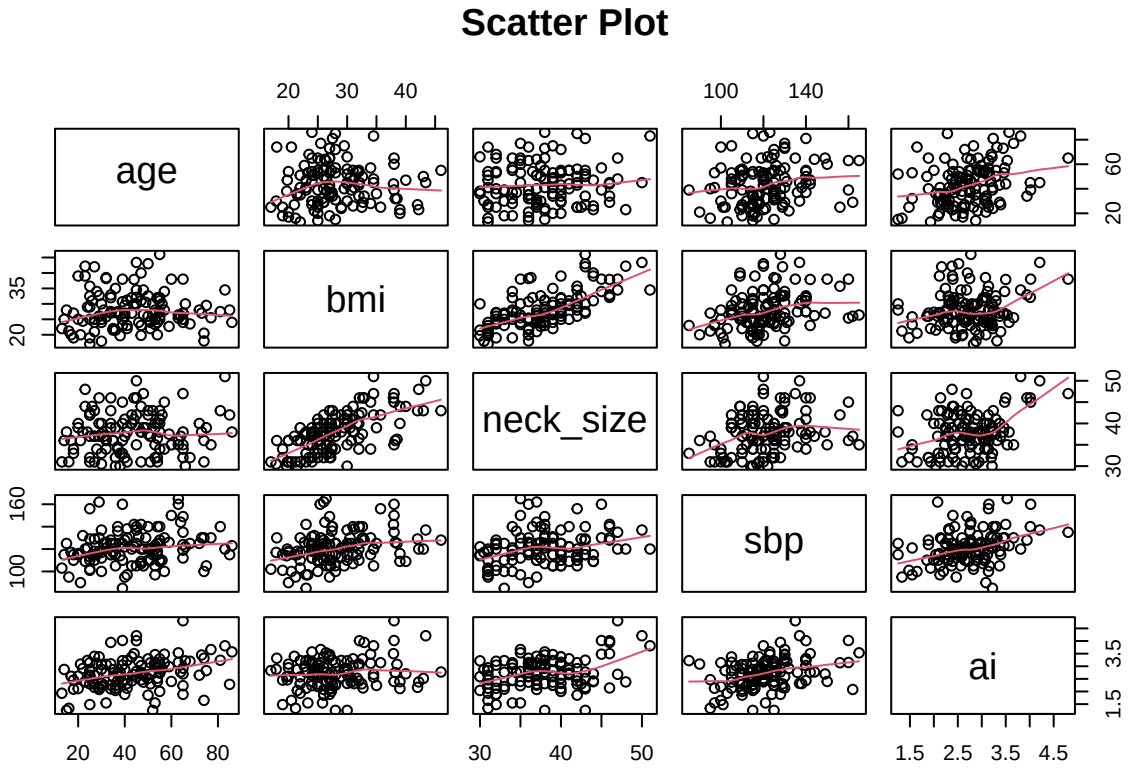
Load the data set and check the structure of data set

```
sleep <- read.csv("sleep.csv", header = TRUE)
head(sleep)
```

```
##   age bmi neck_size sbp       ai
## 1  31  28      37.0 125 2.251292
## 2  46  24      39.0 131 3.190476
## 3  30  24      37.5 112 3.068053
## 4  65  27      38.0 149 2.694627
## 5  54  20      31.0 118 2.939162
## 6  50  26      42.0 110 1.856298
```

Making visualisation to check correlation visually.

```
plot(sleep, main = "Scatter Plot", panel = panel.smooth)
```



Test Correlation through all variables

```
sleep.cor <- cor(sleep)
sleep.cor
```

```
##                 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
```

Correlation Analysis

The correlation matrix shows the pairwise relationships between the variables in the `sleep` dataset.

- Positive Correlations:
  - `ai` and `neck_size` (0.330): Moderate positive correlation.
  - `ai` and `sbp` (0.346): Moderate positive correlation.
  - `ai` and `age` (0.317): Moderate positive correlation.
  - `bmi` and `neck_size` (0.671): Strong positive correlation.
- Weak or No Clear Relationship:
  - `ai` and `bmi` (0.194): Very weak positive correlation.
  - `age` and `bmi` (0.022): Very weak correlation, indicating no significant relationship.

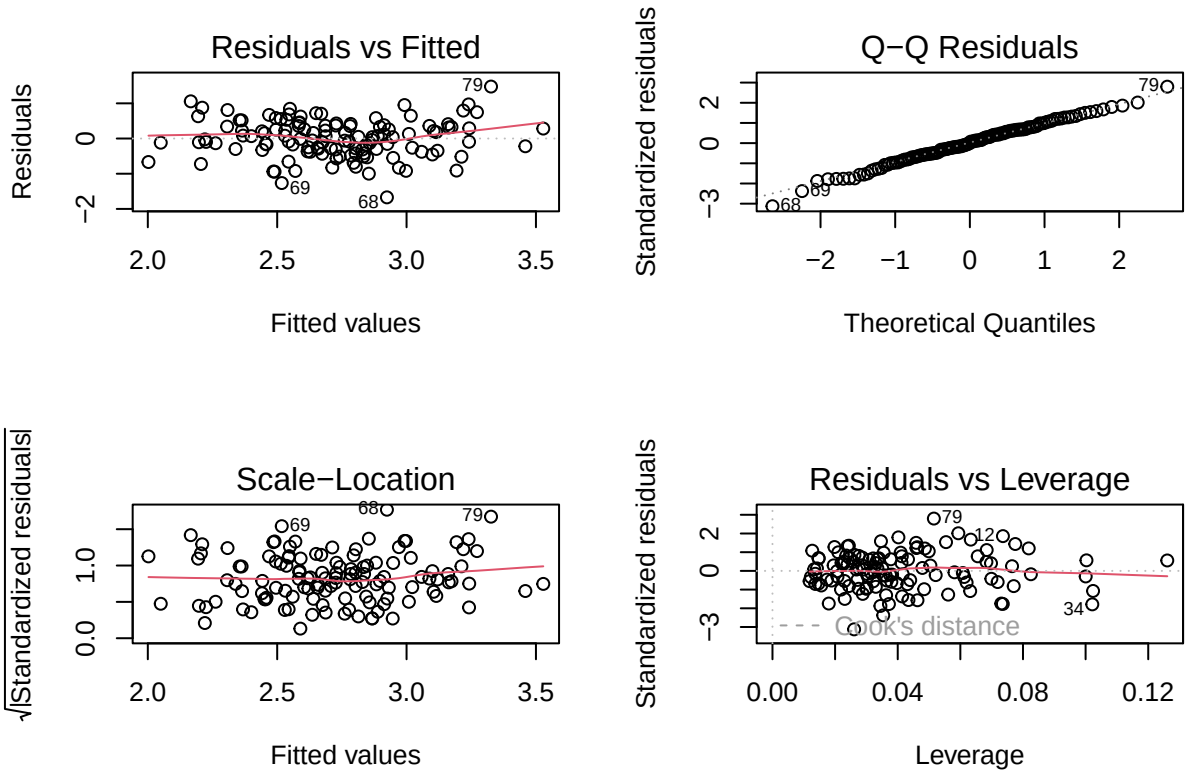
- Implications for Model Building:
  - Variables like `neck_size` and `sbp` might have a more substantial impact on `ai` than `bmi`

Fit a multiple linear regression mode

```
sleep.lm1 <- lm(ai ~ . , data = sleep)
```

Checking generall assumption for regression model

```
par(mfrow = c(2,2))
plot(sleep.lm1)
```



The "Residuals vs Fitted" plot suggests that the assumption of equal variance is reasonable.

The Q-Q plot shows that the residuals are approximately normally distributed, indicating that the normality assumption of the regression model is satisfied.

Thus, this model is good to go for further statistical test

Result of regression model 1

```r
summary(sleep.lm1)
```

```
##
## Call:
## lm(formula = ai ~ ., 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
```

According to the above regression summary, the p-value of bmi is 0.386, which indicates that it is not a statistically significant predictor in this model.

The full regression model has an adjusted R^2 of 0.2213, meaning that approximately 22.13% of the variation in the ai can be explained by the predictors age, bmi, neck_size, and sbp.

confident level for neck_size variable

```r
sleeplm.cof <- confint(sleep.lm1, level = 0.95)
sleeplm.cof
```

```
##                    2.5 %     97.5 %
## (Intercept) -1.185688998 0.86687658
## age          0.002918023 0.01465899
## bmi         -0.032253641 0.01255021
## neck_size    0.012488924 0.06876571
## sbp          0.003178414 0.01725814
```

According to the 95% confidence interval table, the predictor "neck_size" appears to have the strongest positive effect on the response variable. It has the highest estimated coefficient (0.0406) among the predictors, and its 95% confidence interval ranges from 0.0125 to 0.0688, not including zero.

Mathematical multiple regression model

$$Y = \beta_0 + \beta_1 \cdot age_i + \beta_2 \cdot bmi_i + \beta_3 \cdot neck\_size_i + \beta_4 \cdot sbp_i + \epsilon_i$$

Hypothesis for this model

The null hypothesis is:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

The alternative hypothesis is:

$$H_1 : \text{At least one of } \beta_i \neq 0$$

Making ANOVA mode for validate regression model 1

```
sleep.aov <- anova(sleep.lm1)
sleep.aov
```

```
## 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
```

Null distribution formula

Under the null hypothesis, the test statistic follows an F-distribution:

$$F \sim F_{k,n-k-1}$$

Compute F critical value

In this model, $k = 4$ (number of predictors) and $n = 122$ (number of observations), so:

$$F \sim F_{4,117}$$

Find ANOVA regression model 1 F- value to compute P-value

```
summary(sleep.aov)
```

```
##        Df            Sum Sq          Mean Sq          F value
##  Min.   :  1.0   Min.   : 1.605   Min.   :0.2935   Min.   : 5.467
##  1st Qu.:  1.0   1st Qu.: 2.425   1st Qu.:1.6045   1st Qu.: 7.564
##  Median :  1.0   Median : 2.646   Median :2.4251   Median : 8.640
##  Mean   : 24.2   Mean   : 9.121   Mean   :2.3120   Mean   : 9.598
```

```
##  3rd Qu.:  1.0   3rd Qu.: 4.591   3rd Qu.:2.6460   3rd Qu.:10.673
##  Max.   :117.0   Max.   :34.337   Max.   :4.5911   Max.   :15.644
##                                                    NA's   :1
##      Pr(>F)
##  Min.   :0.0001314
##  1st Qu.:0.0024877
##  Median :0.0040400
##  Mean   :0.0073210
##  3rd Qu.:0.0088733
##  Max.   :0.0210727
##  NA's   :1
```

The F-value of overall model is 9.598.

##compute P-value

```
pf(9.598, 4, 117, lower.tail = FALSE)
```

```
## [1] 9.534954e-07
```

The p-value of this model is 9.53×10^7, which is smaller than the significance level of 0.05. Thus, we can reject the null hypothesis and conclude that the model is statistically significant.

in the contextual point, this indicates that at least one of the predictors has a statistically significant effect on the respondent. it means at least one of the predictors can impact on OSA.

Making model 2

```
sleep.lm2 <- lm(ai ~ age + neck_size + sbp, data = sleep)
summary(sleep.lm2)
```

```
##
## Call:
## lm(formula = ai ~ age + neck_size + sbp, data = sleep)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.65415 -0.35334  0.04008  0.37534  1.45627
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.066166   0.506509  -0.131  0.89629
## age          0.009007   0.002951   3.053  0.00280 **
## neck_size    0.032630   0.010831   3.013  0.00317 **
## sbp          0.009579   0.003475   2.757  0.00676 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5412 on 118 degrees of freedom
## Multiple R-squared:  0.2422, Adjusted R-squared:  0.2229
## F-statistic: 12.57 on 3 and 118 DF,  p-value: 3.452e-07
```

To optimise model, I cut the bmi variable, because it has highest P-value.

this new model indicate that

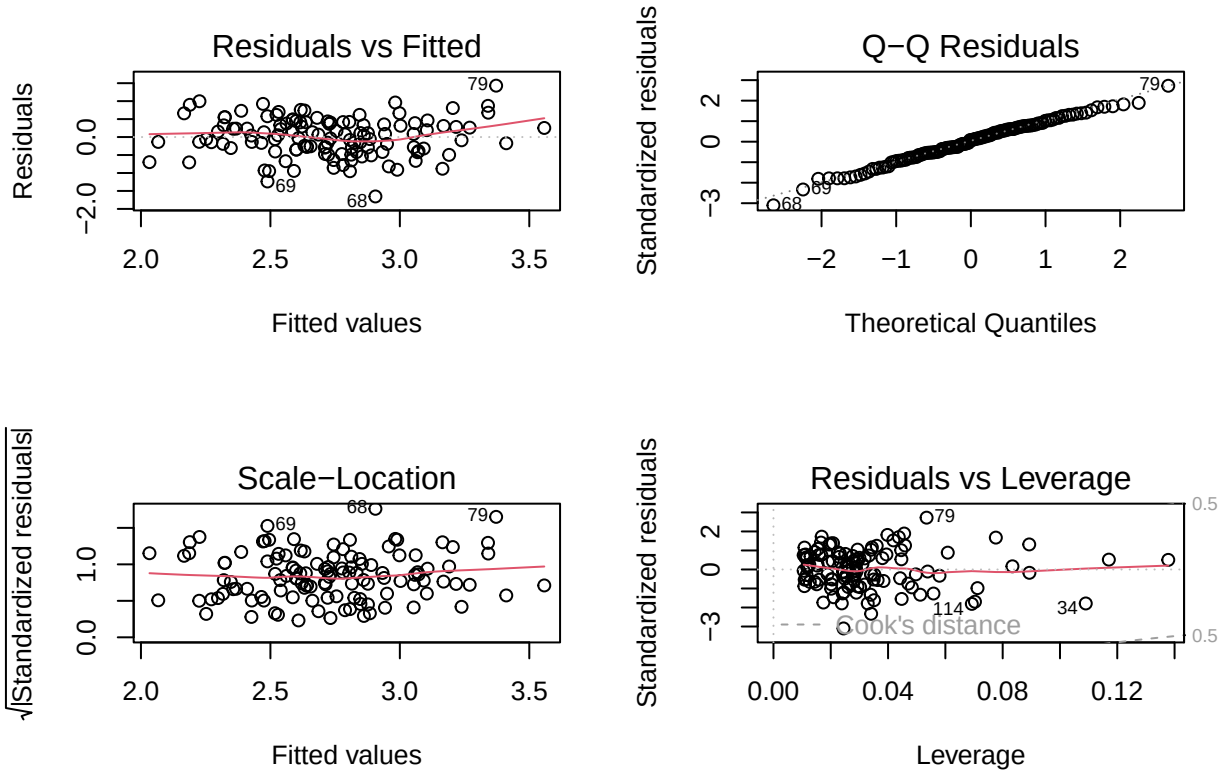
Making model 3

```
sleep.lm3<- lm(ai ~ age + I(neck_size^2) + I(sbp^2), data = sleep)
summary(sleep.lm3)
```

```
##
## Call:
## lm(formula = ai ~ age + I(neck_size^2) + I(sbp^2), data = sleep)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.65293 -0.34117  0.04357  0.38552  1.43040
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.122e+00  2.670e-01   4.203 5.15e-05 ***
## age            8.976e-03  2.940e-03   3.053  0.00280 **
## I(neck_size^2) 4.372e-04  1.383e-04   3.161  0.00200 **
## I(sbp^2)       3.844e-05  1.369e-05   2.807  0.00585 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5395 on 118 degrees of freedom
## Multiple R-squared:  0.2468, Adjusted R-squared:  0.2276
## F-statistic: 12.89 on 3 and 118 DF,  p-value: 2.435e-07
```

Checking the general assumption for ANOVA Regression Model 3

```
par(mfrow = c(2,2))
plot(sleep.lm3)
```



Resitual equal variance and normality assumption are satisfied for model 3.

Comparison model explnation

```
compare.model.r <- c("Model 1" = summary(sleep.lm1)$adj.r.squared, "Model 2" = summary(sleep.lm2)$adj.r
print(compare.model.r)
```

```
##   Model 1   Model 2   Model 3
## 0.2213170 0.2229104 0.2276094
```

The table and chart above show that Model 3 has the highest adjusted R^2, indicating that it provides the best explanatory power for the response variable. Furthermore, it does not violate the general assumptions of a regression model.

Conclusion

Model 3:

$$Y = \beta_0 + \beta_1 \cdot age_i + \beta_3 \cdot neck\_size_i^2 + \beta_4 \cdot sbp_i^2 + \epsilon_i$$

is selected as the best model because it has the highest adjusted$R^2(0.228)$ among the models considered.

Question 2

```
energy <- read.csv("energy.csv", header = TRUE)
head(energy)
```

```
##   range menu consumption
## 1     1    1   10.344389
## 2     1    2    8.410001
## 3     2    1   10.647400
## 4     2    2    8.574652
## 5     3    1    8.509676
## 6     3    2    6.138515
```

checking combination of variables

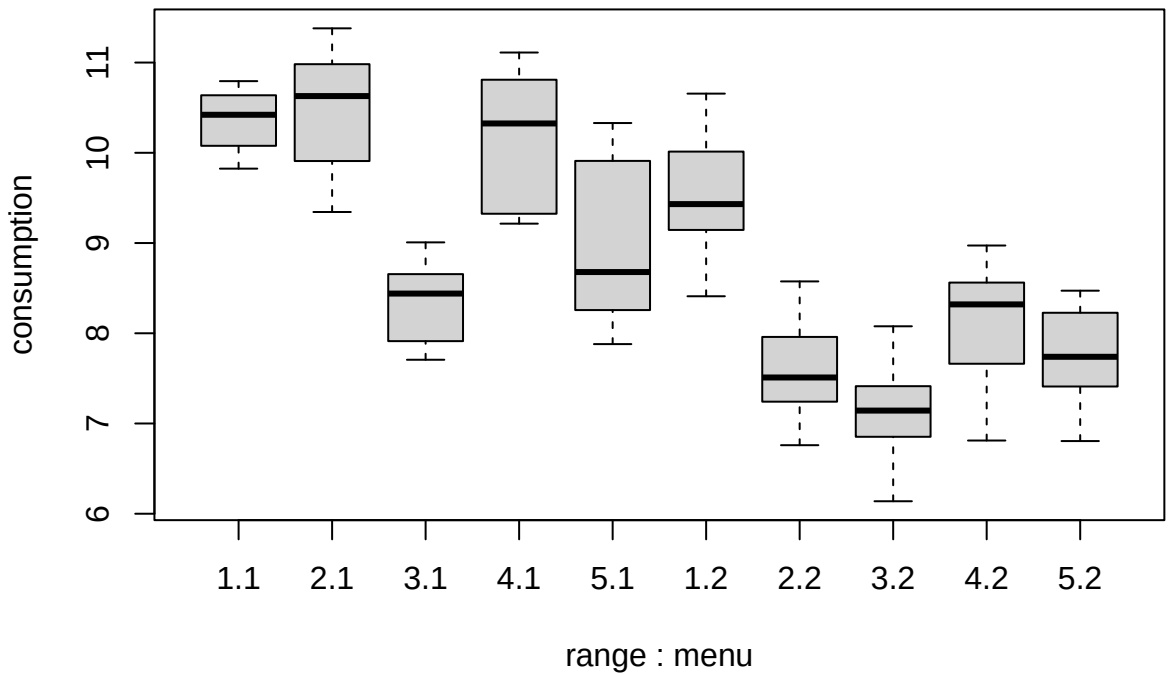
```
table(energy$range, energy$menu)
```

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

The design is balanced because each combination of range (1–5) and menu (1–2) contains the same number of observations (8).

box plot

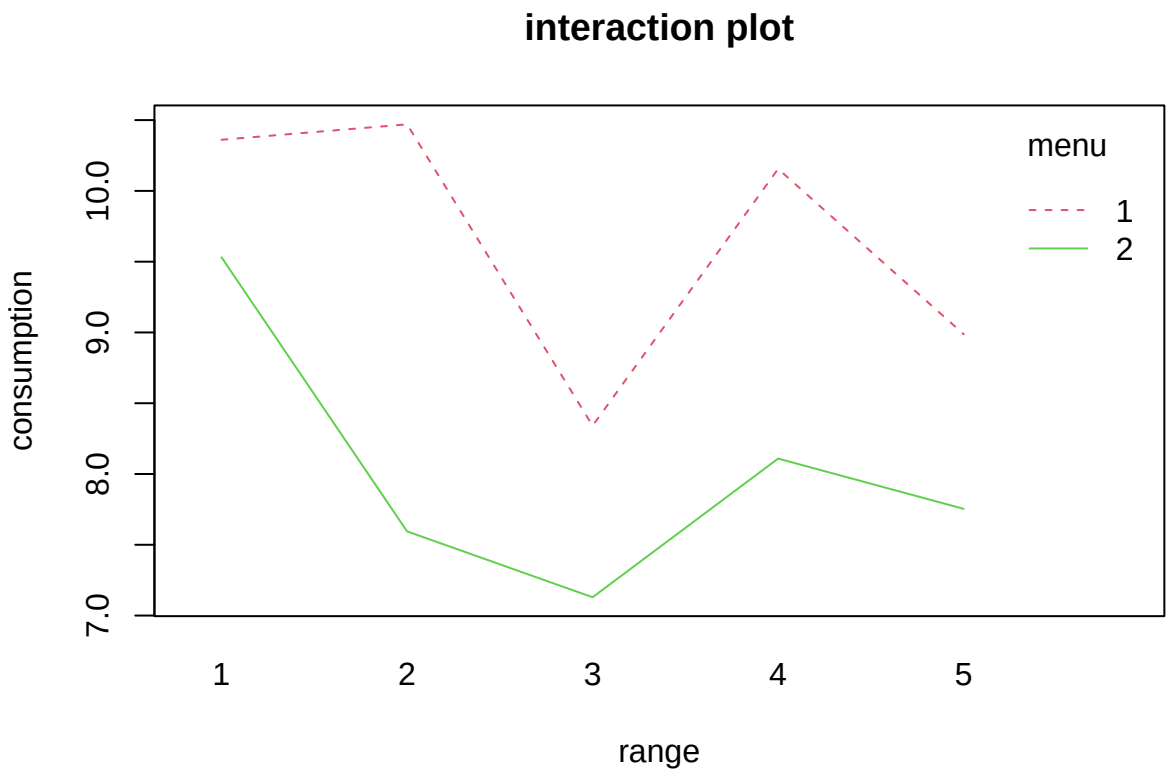
```
boxplot(consumption ~ range*menu, data  = energy)
```



The above boxplot shows that Menu 1 consistently leads to higher energy consumption than Menu 2 across all ranges. Also, ranges 2 and 4 tend to consume more energy than range 3.

interaction plot

```
interaction.plot(x.factor = energy$range, trace.factor = energy$menu, response = energy$consumption, xla
```



The interaction plot shows that the lines for Menu 1 and Menu 2 are not perfectly parallel, particularly for ranges 2 and 4. This suggests a potential interaction effect between range and menu, where the effect of menu on consumption may vary across ranges. Further statistical testing is required to determine whether this interaction is significant.

Regression formula

$$Yij = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

γ = interaction between α and β

Hypotheses

main effect of range:

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0$$

$$H_1 : \text{at least one of } \alpha_i \neq 0$$

Main effect of menu:

$$H_0 : \beta_1 = \beta_2 = 0$$

$$H_1 : \text{at least one of } \beta_j \neq 0$$

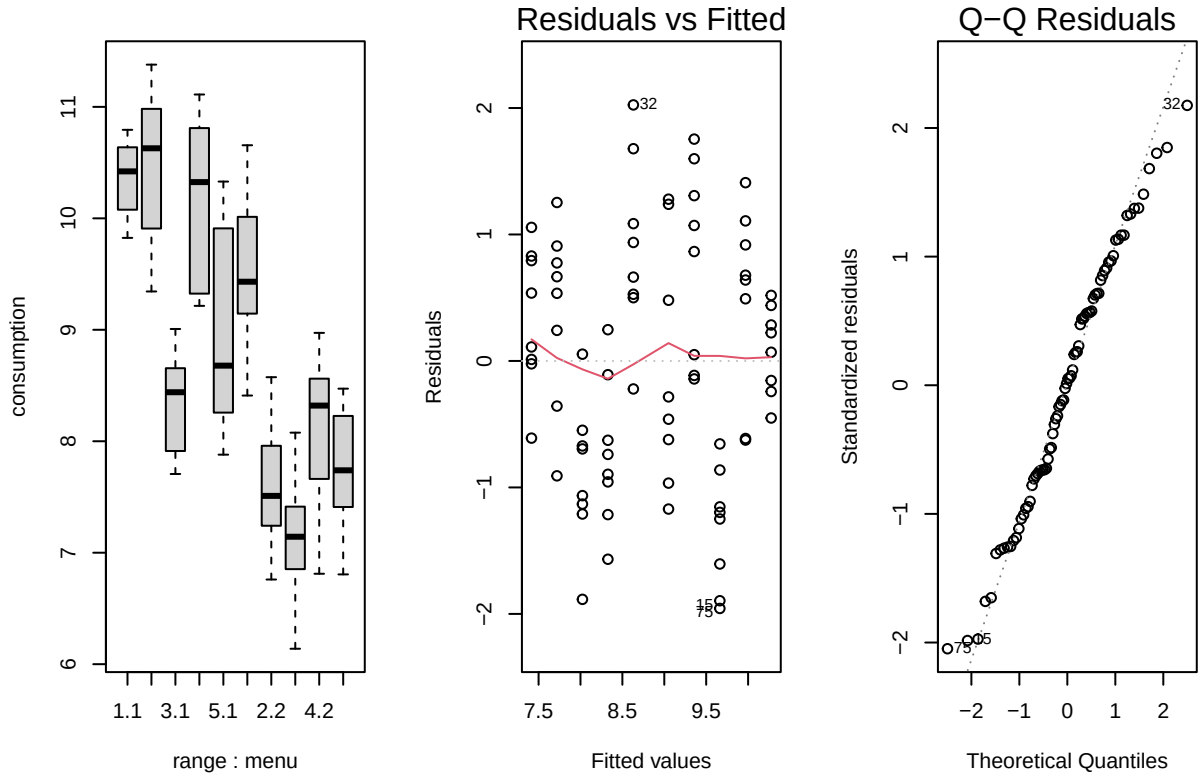
ANOVA balanced design test

```
energy.aov.bt <- aov(consumption ~ range*menu, data = energy)
summary(energy.aov.bt)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## range         1  14.89   14.89   15.91 0.000152 ***
## menu          1  53.77   53.77   57.45    7e-11 ***
## range:menu    1   0.00    0.00    0.00 0.988205
## Residuals    76  71.14    0.94
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Checking the ANOVA model general assumption.

```
par(mfrow = c(1,3))
boxplot(consumption ~ range + menu, data = energy)
plot(energy.aov.bt, which = 1:2)
```



Result of general assumption test

the plot above show that the variability of variables and the structure of residuals. The second "Residual vs Fitted" plot indicate that the residuals are evenly spread by suggesting the equal variance assumption meet.

The third "Q-Q Residuals" plot suggest the residuals follow the normal distribution, satisfying normality assumption. Thus, this model satisfies all general assumption of ANOVA test.

Result of ANOVA test

The ANOVA table shows that the p-value for the interaction term `range:menu`(γ_{ij}) is 0.988, which is much greater than the significance level of 0.05. This indicates that the interaction between `range` and `menu` is not statistically significant. Also, we can reject the null hypothesis for two predictors except for γ, since the P-value of two predictors is smaller than significance level(0.05). this result means two predictors would impact on respondent respectively.

Conclusion

The ANOVA results show that both range and menu have statistically significant effects on energy consumption ($p < 0.001$), however their interaction is not significant ($p = 0.988$). Diagnostic plots confirm that the model assumptions of normality and equal variance are satisfied. Therefore, we conclude that the choice of range and menu independently influences electricity consumption, but there is no evidence of interaction between them in this context.