Homework 5

Bivariate Statistics One-way ANOVA and Regression Analysis

Andri Setiyawan   Benedikt Meyer   Yosep Dwi Kristanto

November 14, 2024

Problems

1) ANOVA: Launch SPSS and open the data file Telemarketing.sav

   Assume that in an attempt to maximize profits, a telemarketing company is conducting an experiment to determine which of four scripted sales pitches generates the best revenue. 1500 different telemarketing calls are randomly assigned to one of the four scripts, and the resulting revenue for each call is recorded.

   Run an appropriate ANOVA test for this research design.

2) Regression Analysis: Run a multiple regression analysis on the examrevison.sav dataset, pay particular attention to the 7 Regression diagnostics conditions. This data represents measures from students used to predict how they perform in an exam.

1 Telemarketing

1.1 Data Exploration

Table 1 shows that average revenue and variability differ across the four sales pitches, with Script A generating the highest average revenue and Script D the lowest.

The distribution of `revenue` across the four `sales_pitch` (Script A, Script B, Script C, and Script D) is visually summarized using violin plots combined with boxplots, as shown in Figure 1.

Table 1: Summary statistics for `revenue` (n, mean, and standard deviation) across different `sales_pitch` in the telemarketing data

sales_pitch	n	M	SD
Script A	279	2970.630	947.2344
Script B	351	2669.133	970.9186
Script C	305	2471.292	967.0648
Script D	553	2215.649	943.0035

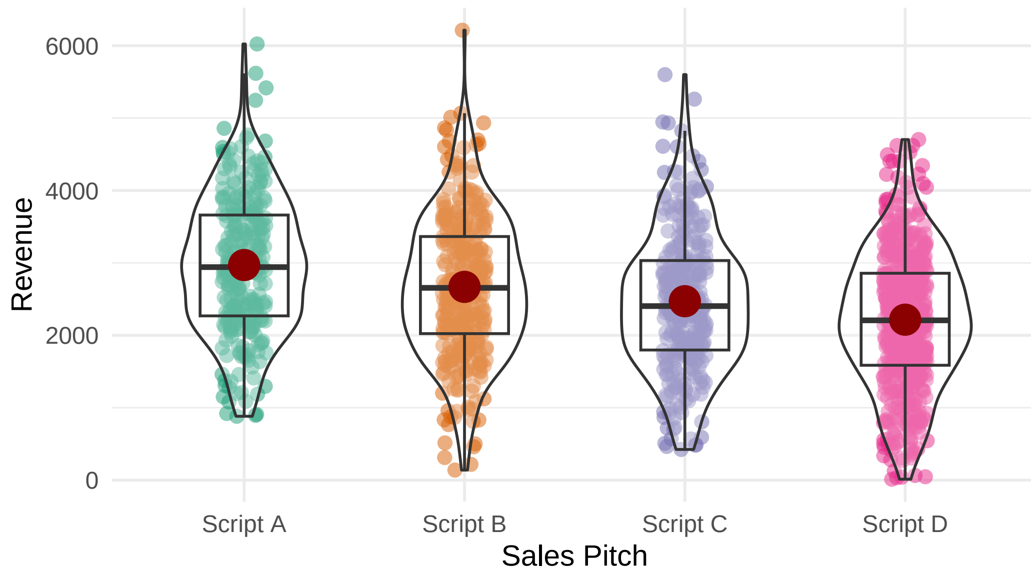


Figure 1: Distribution of `revenue` across `sales_pitch` (Script A, Script B, Script C, and Script D) as illustrated by violin and boxplots.

1.2 Assumption Checking

- The outcome variable, revenue, is measured on a ratio scale.
- The groups are mutually exclusive, with four distinct categories: Script A, Script B, Script C, and Script D.
- The grouping variable consists of four levels: Script A, Script B, Script C, and Script D.
- The QQ plots were used to assess the normality of `revenue` distributions for each `sales_pitch` (Script A, Script B, Script C, and Script D). See Figure 2.

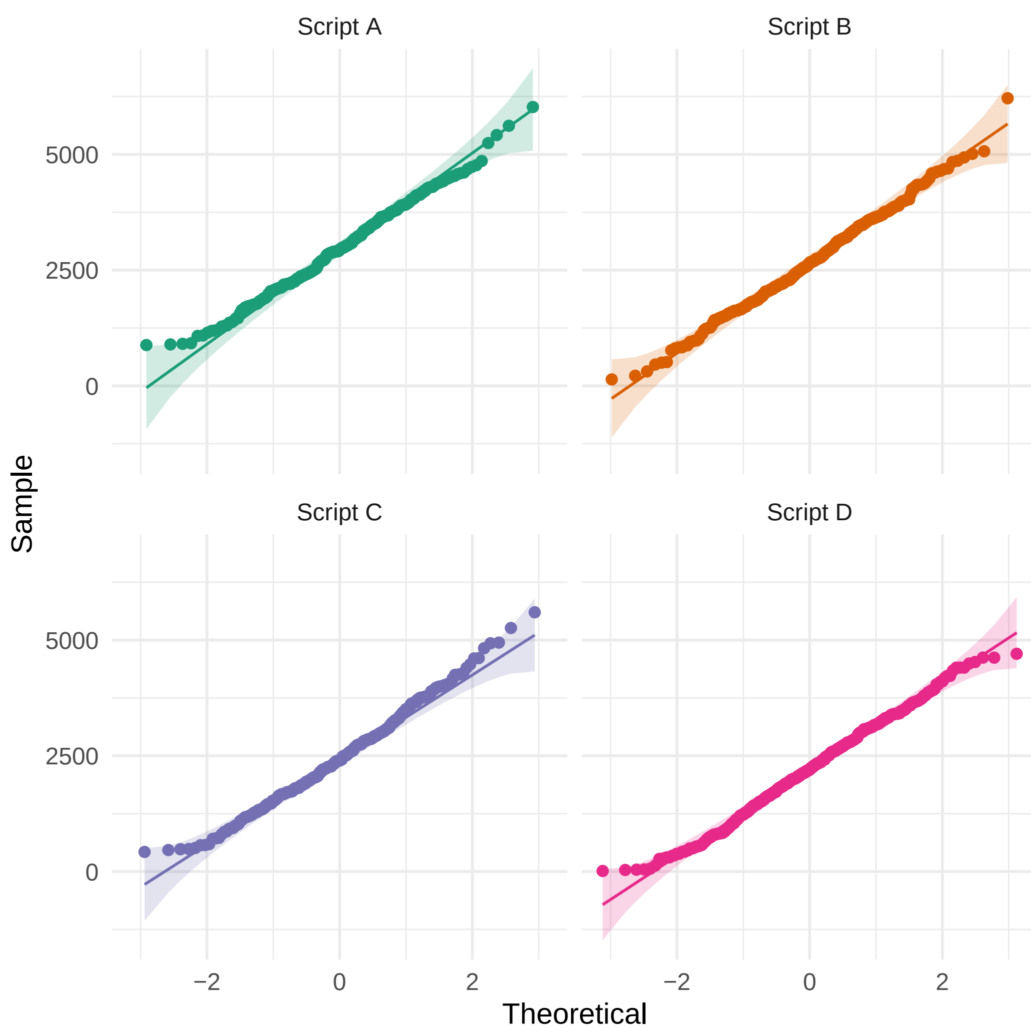


Figure 2: QQ plot of `revenue` across `sales_pitch`

Table 2: Shapiro-Wilk test of normality for `revenue` across `sales_pitch`

sales_pitch	variable	statistic	p
Script A	revenue	0.9936389	0.2874851
Script B	revenue	0.9960223	0.5244825
Script C	revenue	0.9913526	0.0708032
Script D	revenue	0.9949141	0.0647445

Table 3: Results of Levene test for homogeneity of variance

df1	df2	statistic	p
3	1484	0.0924258	0.9642314

From Figure 2, it appears that the `revenue` for each `sales_pitch` is likely drawn from a normally distributed population. This observation is supported by the Shapiro-Wilk test results presented in Table 2.

The p-value in Table 2 is greater than .05 suggests that the `revenue` in each `sales_pitch` follows a normal distribution.

- Table 3 presents the results of Levene's test for homogeneity of variances of `revenue` across the different `sales_pitch` groups. Since the p-value is greater than .05, it suggests that the assumption of equal variances is met.

1.3 Hypotheses

H_0: The average `revenue` is equal across all `sales_pitch` groups.

H_1: At least one pair of `sales_pitch` groups has a different average `revenue`.

1.4 Calculating the F statistic

The ANOVA results in Table 4 show an F-value of 42.505, testing the difference in average `revenue` across the `sales_pitch` groups.

Table 4: ANOVA table testing the difference in average `revenue` across `sales_pitch` groups.

Effect	DFn	DFd	F	p	p<.05	ges
sales_pitch	3	1484	42.505	0	*	0.079

1.5 Testing for the significance of F

1.6 Interpreting F

1.7 Post-hoc test

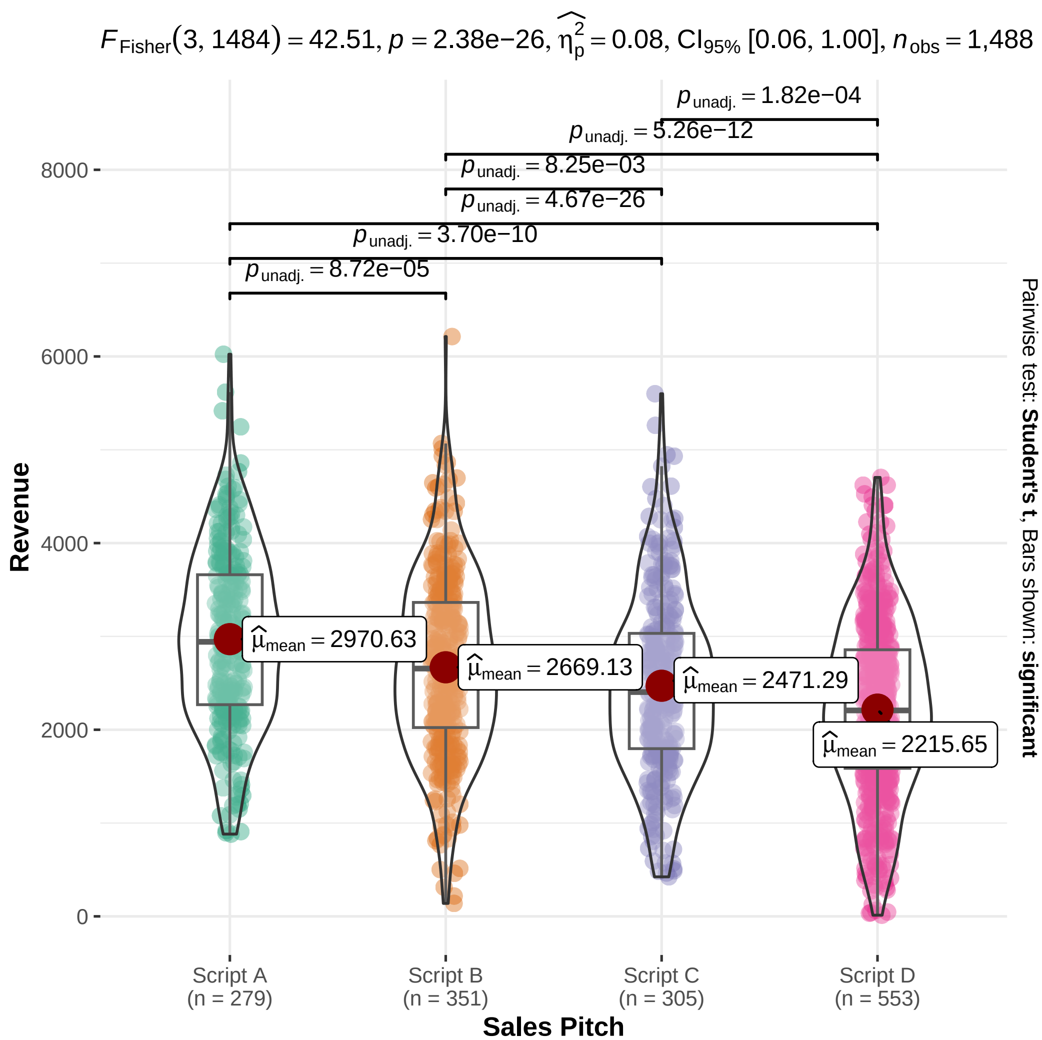


Figure 3: ANOVA and post-hoc test results for telemarketing data

2 Students' Performance

2.1 Data Exploration

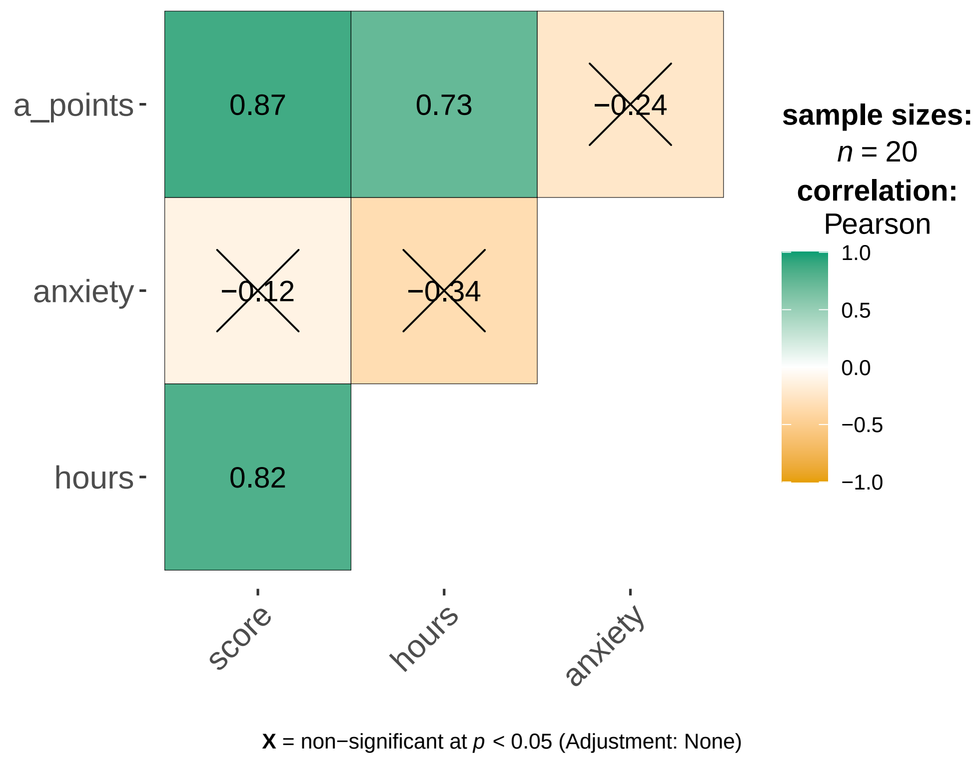


Figure 4: Correlation matrix among all variables in exam data

2.2 Hypotheses

H_0: All regression coefficients are equal to zero (except the intercept).

H_1: At least one of the regression coefficients is not equal to zero.

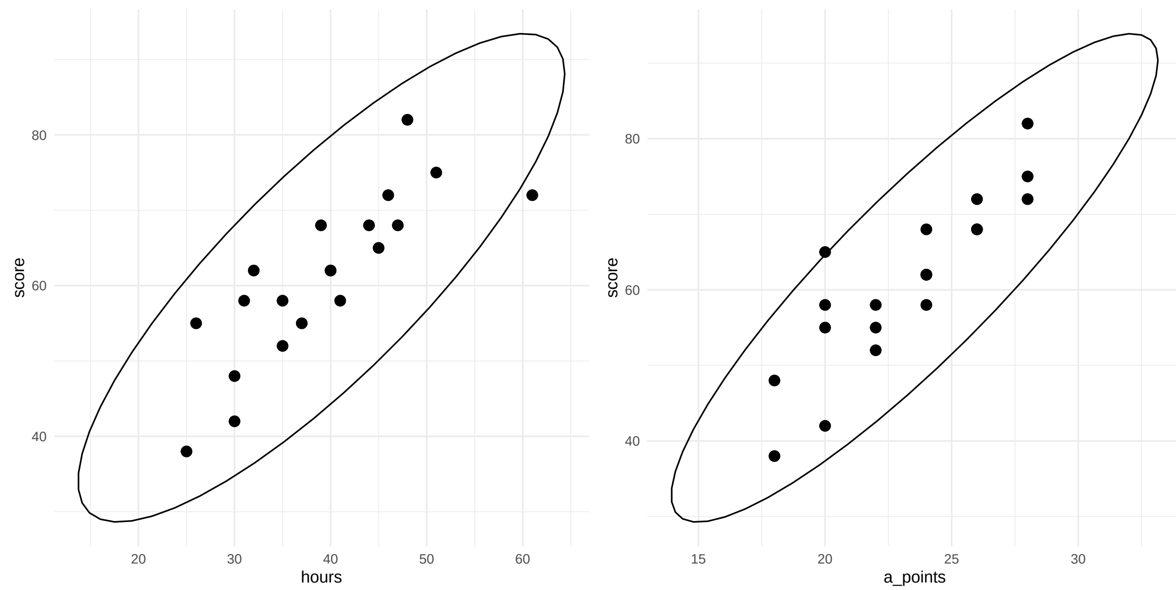
2.3 Assumption Checking

Correct specification of the model: It is make sense to predict students' performance score (`score`) with how long they spent on revision (`hours`) and their A-level entry points (`a_points`).

Table 5: Results of regression analysis on `score`

dependent_variable	independent_variables	F_statistic	p_value	R_squared	df	df_res
score	hours	37.2247122	0.0000092	0.6740590	1	18
score	anxiety	0.2554643	0.6193864	0.0139939	1	18
score	a_points	56.9216004	0.0000006	0.7597489	1	18
score	hours, anxiety	20.1343463	0.0000328	0.7031537	2	17
score	hours, a_points	42.0890633	0.0000003	0.8319795	2	17
score	anxiety, a_points	28.3368905	0.0000039	0.7692531	2	17
score	hours, anxiety, a_points	32.8112701	0.0000005	0.8601812	3	16

Linearity: Figure 5 shows relationships between `score`, `hours` and `a_points`. From the figures, we can see that the relationship between `hours`, `a_points`, and `score` are linear.



(a) Relationship between `hours` and `score` (b) Relationship between `a_points` and `score`

Figure 5: Linearty verification

Measurement and normality of dependence variable: The dependence variable, i.e. `score`, is ratio. From the Figure 6, we can assume that `score` sample is from normally distributed population.

Absence of multicollinearity: Here is the correlation coefficient between `hours` and `a_points`.

```
[1] 0.7317732
```

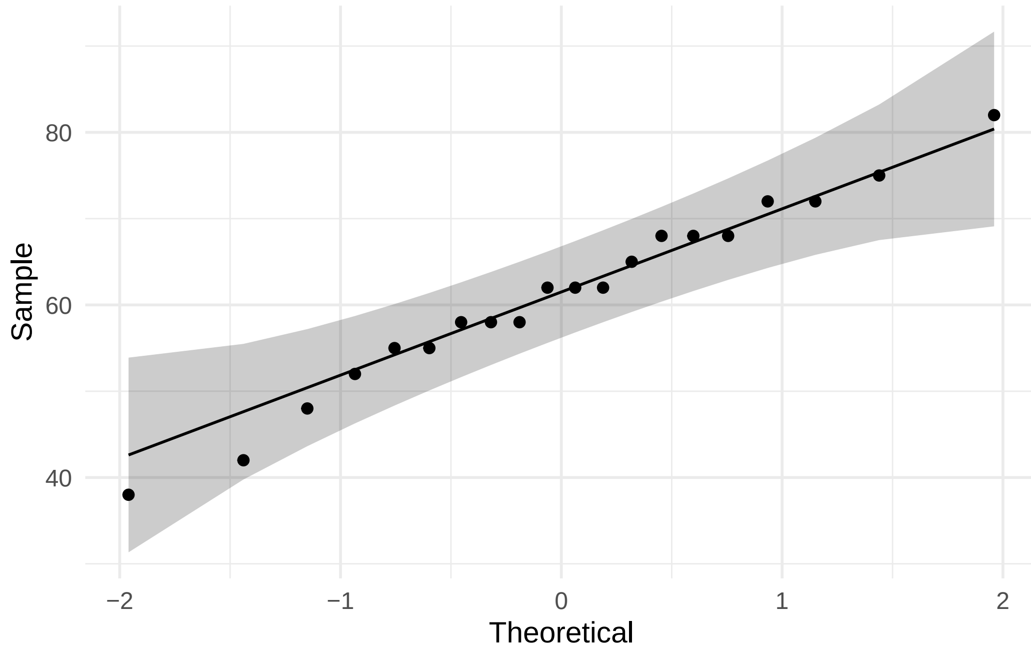


Figure 6: Assessing normality for `score`

Since the correlation coefficient is less than .8, we infer that there is no multicollinearity between the independent variables.

Normal distribution of residuals: Figure 7 shows that the residuals are normally distributed.

Homoscedasticity: Figure 8 shows that the residuals have equal variance across dependence variable.

2.4 Model

```
Call:
lm(formula = score ~ hours + a_points, data = exam)

Residuals:
   Min     1Q Median     3Q    Max 
-8.258 -2.398 -1.001  2.697  7.595 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

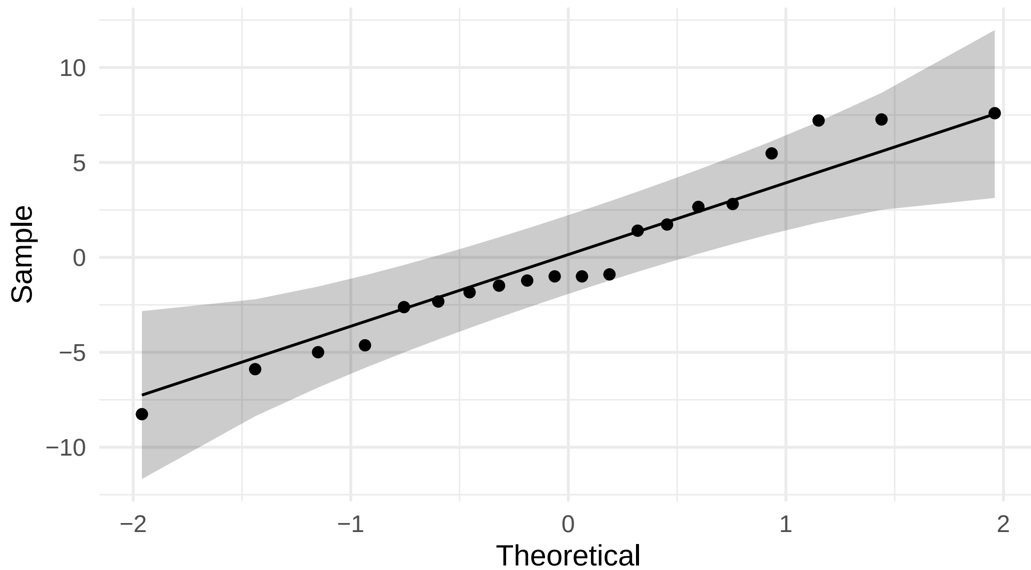


Figure 7: Assessing the normality of residuals

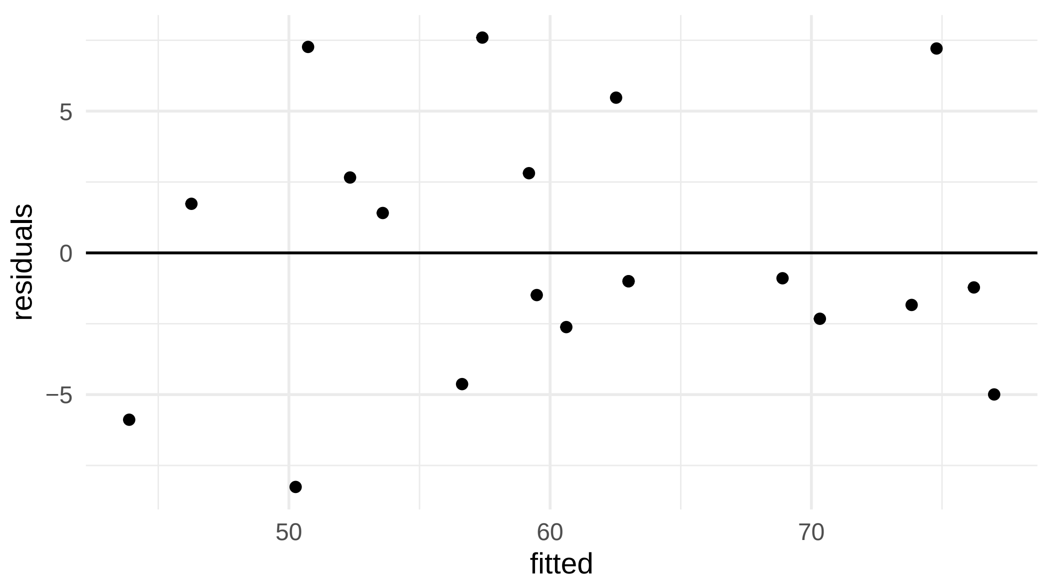
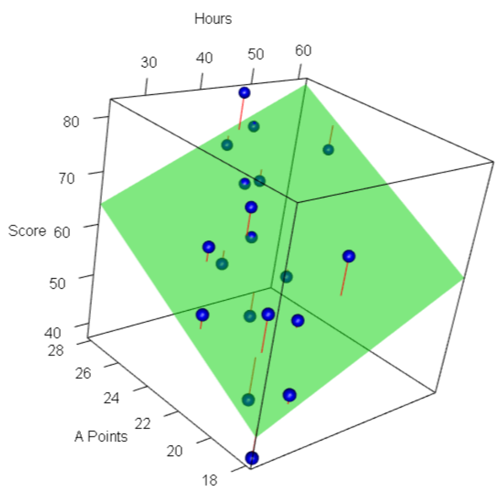


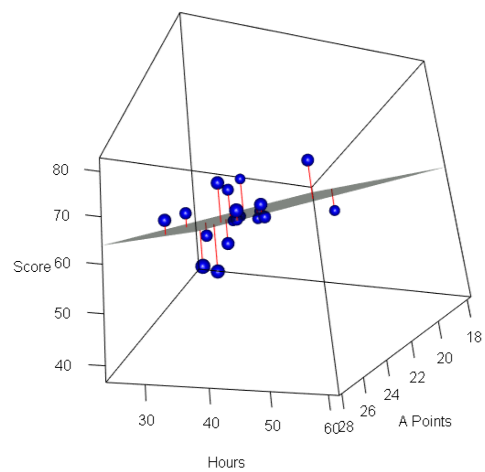
Figure 8: Assessing homoscedacity of residuals

```
(Intercept)  -3.9251     8.1143  -0.484 0.634754
hours         0.4765     0.1762   2.703 0.015069 *
a_points      1.9945     0.4990   3.997 0.000933 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.751 on 17 degrees of freedom
Multiple R-squared:  0.832, Adjusted R-squared:  0.8122
F-statistic: 42.09 on 2 and 17 DF,  p-value: 2.604e-07
```



(a)



(b)

Figure 9: Visualization of the 3D model