Homework 5

Bivariate Statistics One-way ANOVA and Regression Analysis

Andri Setiyawan Benedikt Meyer Yosep Dwi Kristanto

November 22, 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

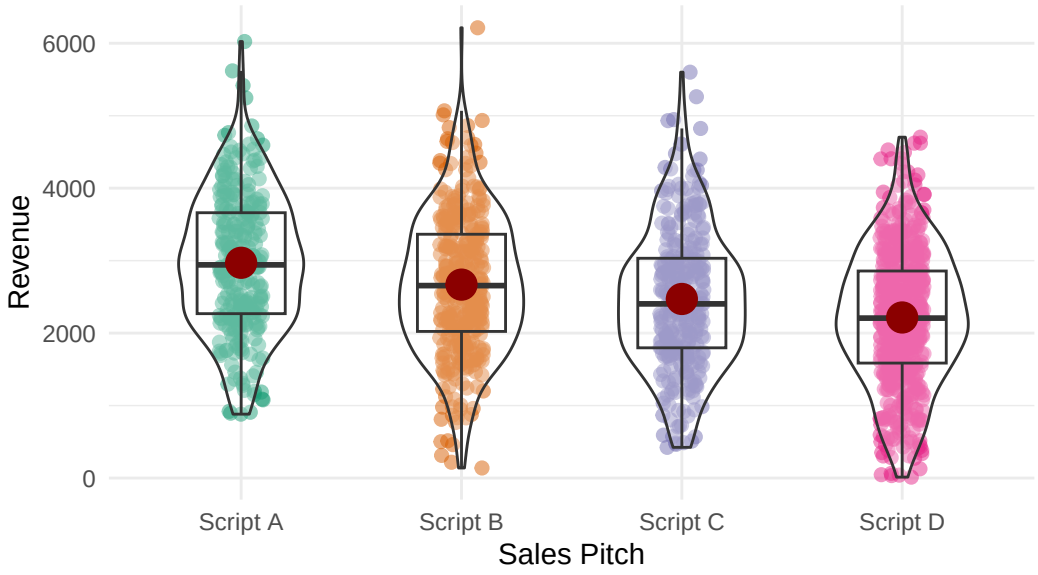


Figure 1: Distribution of `revenue` across `sales_pitch` (Script A, Script B, Script C, and Script D) as illustrated by violin and boxplots.

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.

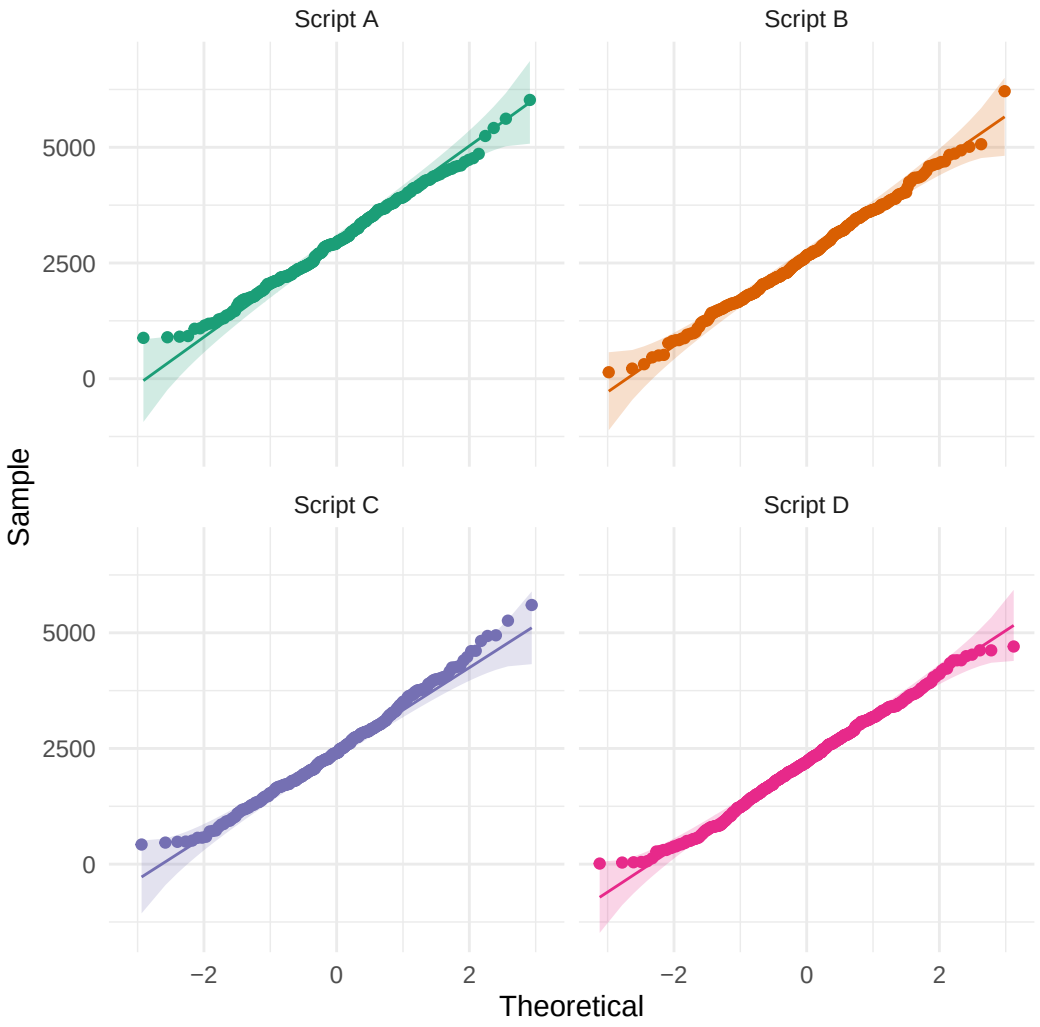


Figure 2: QQ plot of `revenue` across `sales_pitch`

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

Table 4 shows a p-value of 2.38×10^{-26}, which is less than .05. A visualization of the p-value is presented in Figure 3.

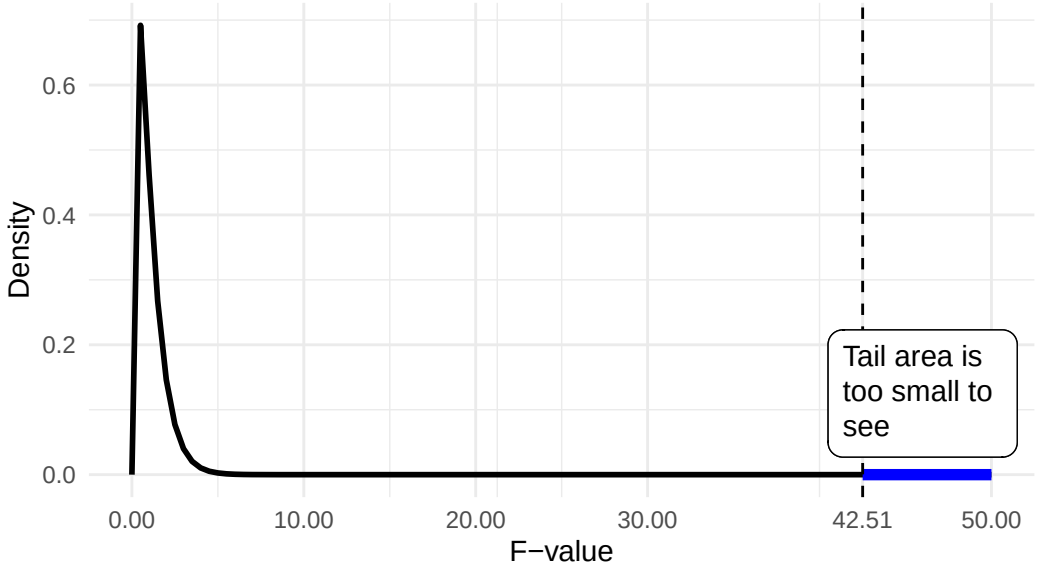


Figure 3: Theoretical F-distribution with degrees of freedom 3 and 1484, illustrating the tail area corresponding to the p-value

1.6 Interpreting F

Assuming the null hypothesis is true, i.e., that the average `revenue` is equal across all `sales_pitch` groups, the sample yields an F-statistic of 42.51 and a p-value of 2.4×10^{-26}, which is less than .05. As a result, we reject the null hypothesis, indicating that at least two groups in the `sales_pitch` have different average `revenue`.

1.7 Effect Size

The generalized eta-squared (ges) from Table 4 of 0.079 indicates that approximately 7.9% of the total variance in `revenue` can be attributed to differences across the `sales_pitch` groups. The value of 0.079 suggests a medium effect size, indicating that group differences in `sales_pitch` explain a meaningful but not overwhelming portion of the variance in `revenue`.

Table 5: Pairwise t-test results comparing revenue between sales_pitch groups, with Bonferroni-adjusted p-values

Group 1	Group 2	t	df	p (Bonferroni-adj.)
Script A	Script B	3.924603	602.4695	0.000523
Script A	Script C	6.299952	579.2764	2.22e-09
Script A	Script D	10.870057	555.5645	2.8e-25
Script B	Script C	2.608597	641.9791	0.0495
Script B	Script D	6.920593	728.9095	3.16e-11
Script C	Script D	3.739169	613.5349	0.00109

1.8 Post-hoc test

Table 5 displays the pairwise t-test result with Bonferroni-adjusted p-values. It indicates the statistical differences between the groups. These results suggest that there are significant differences in average of `revenue` between each pair of `sales_pitch` groups, which are highlighted in the pairwise comparisons.

The results of the pairwise t-tests are visually represented in Figure 4.

1.9 Reporting the Results

A one-way analysis of variance (ANOVA) was conducted to examine the effect of sales pitch on revenue. The results revealed a significant difference in average revenue across the four sales pitch groups, $F(3, 1484) = 42.51$, $p < .001$, generalized eta-squared (ges) = .079, indicating that sales pitch had a medium effect on revenue.

Post-hoc comparisons were performed using pairwise t-tests with Bonferroni correction to identify specific group differences. The results indicated the following:

- The average revenue for Script A ($M = 2970.63$, $SD = 947.23$) was significantly higher than for Script B ($M = 2669.13$, $SD = 970.92$), $t(602.47) = 3.92$, $p < .001$; Script C ($M = 2471.29$, $SD = 967.06$), $t(579.28) = 6.30$, $p < .001$; and Script D ($M = 2215.65$, $SD = 943.00$), $t(555.56) = 10.87$, $p < .001$.
- Similarly, Script B had significantly higher revenue than Script C, $t(641.98) = 2.61$, $p = .0495$, and Script D, $t(728.91) = 6.92$, $p < .001$.
- Lastly, Script C had significantly higher revenue than Script D, $t(613.53) = 3.74$, $p = .00109$.

These results suggest that Script A consistently led to the highest revenue, while Script D resulted in the lowest revenue among the four groups.

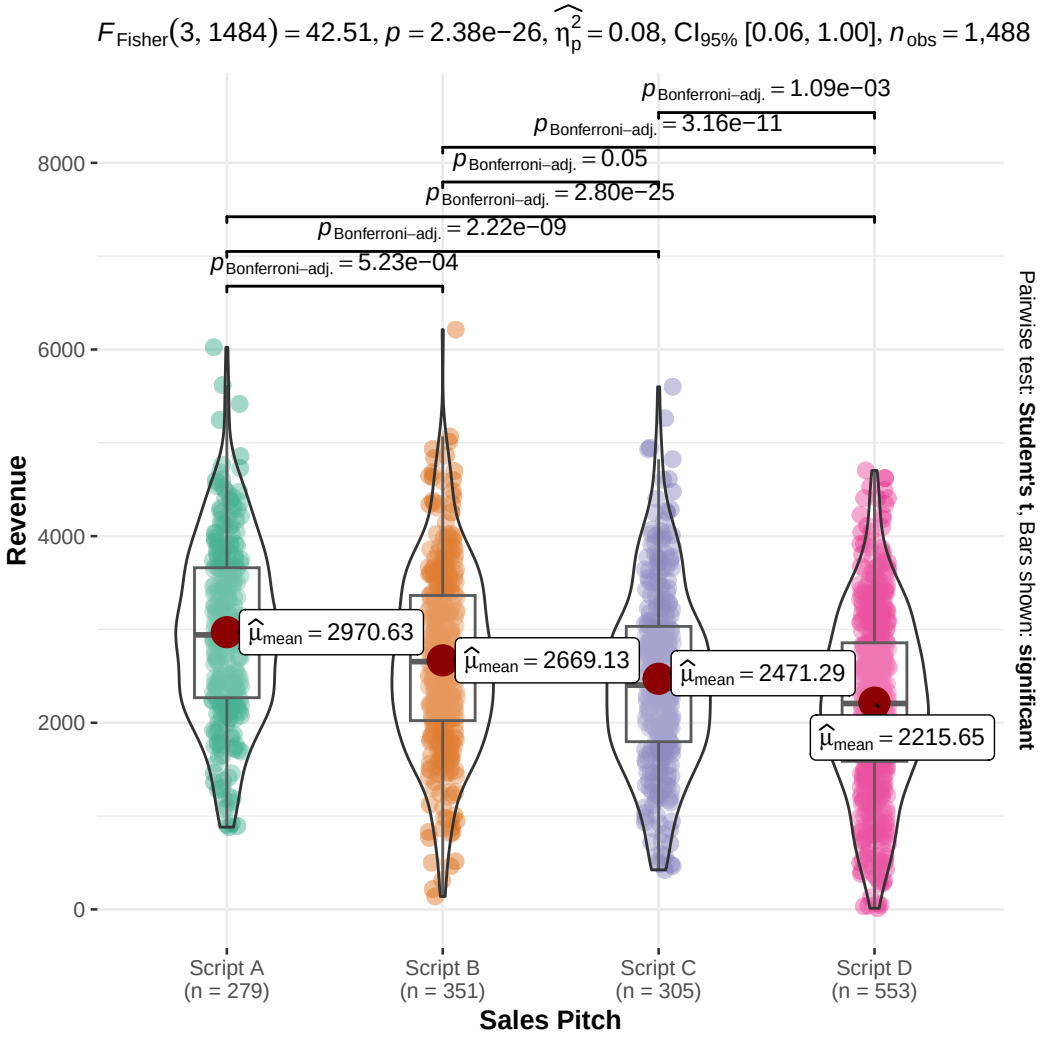


Figure 4: Violin plots and boxplots showing the distribution of revenue across the sales_pitch groups, with results of pairwise t-tests indicating differences between groups.

2 Students' Performance

2.1 Data Exploration

Figure 5 presents the relationships between `score`, `hours`, `anxiety`, and `a_points` in the exam dataset using Pearson correlation coefficients. The strongest positive correlation is observed between `score` and `a_points` (r = 0.87), followed by the correlation between `score` and `hours` (r = 0.82). A weaker, negative correlation is found between `hours` and `anxiety` (r = -0.34), as well as between `score` and `anxiety` (r = -0.12).

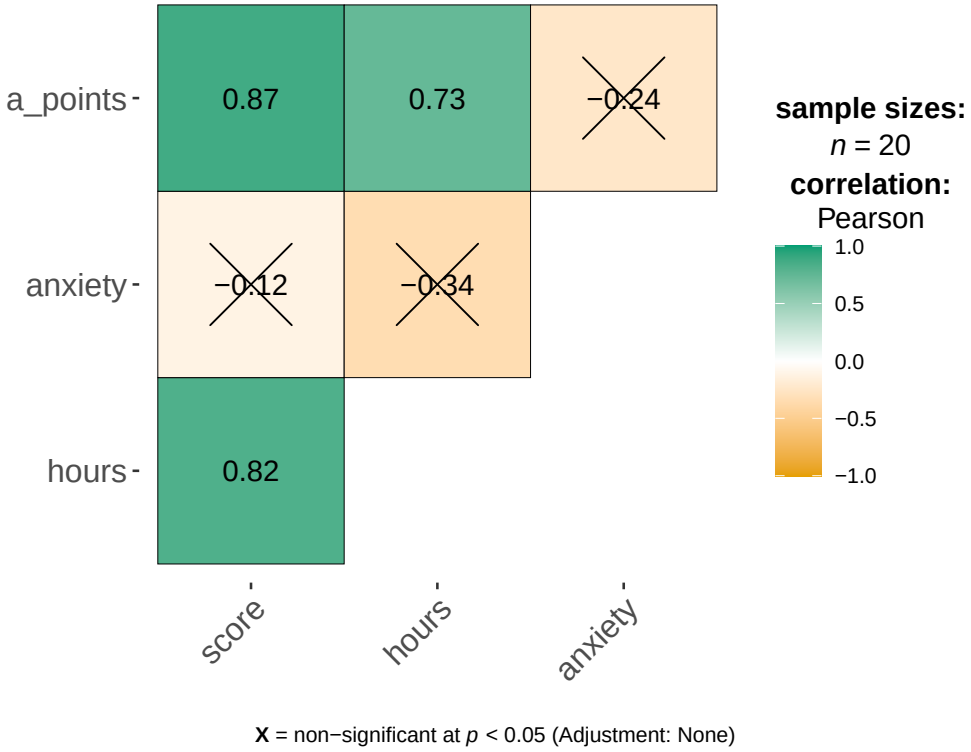


Figure 5: Correlation matrix among all variables in exam data

Table 6 presents the results of regression analyses with `score` as the dependent variable and combinations of `hours`, `anxiety`, and `a_points` as independent variables.

We choose `hours` and `a_points` as the independent variables for further analysis because they individually and together demonstrate strong predictive power for score. The model with these two variables accounts for 83% of the variance and has a high F-statistic, suggesting they are reliable predictors without the redundancy of including less impactful variables like `anxiety`.

Table 6: Summary table of regression analyses for all possible combinations of independent variables (`hours`, `anxiety`, and `a_points`) predicting `score` in the `exam` dataset

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

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`).

Linearity: Figure 6 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.

Measurement and normality of dependence variable: The dependence variable, i.e. `score`, is ratio. From the Figure 7, we can assume that `score` sample is from normally distributed population.

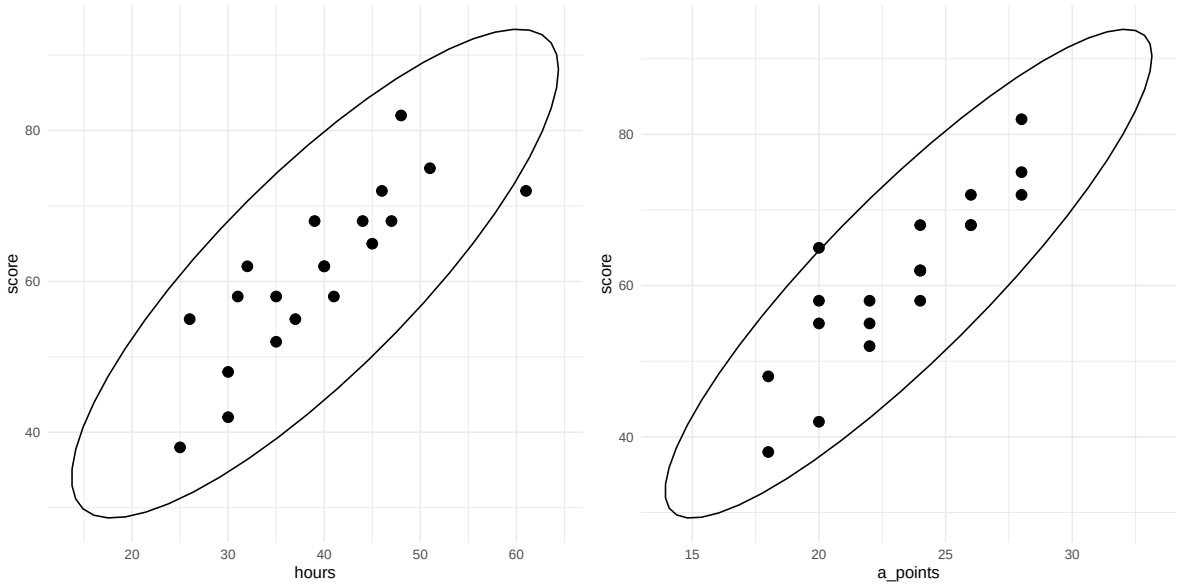
Absence of multicollinearity: Here is the correlation coefficient between `hours` and `a_points`.

```
[1] 0.7317732
```

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

Normal distribution of residuals: Figure 8 shows that the residuals are normally distributed.

Homoscedasticity: Figure 9 shows that the residuals have equal variance across dependence variable.



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

Figure 6: Linearty verification

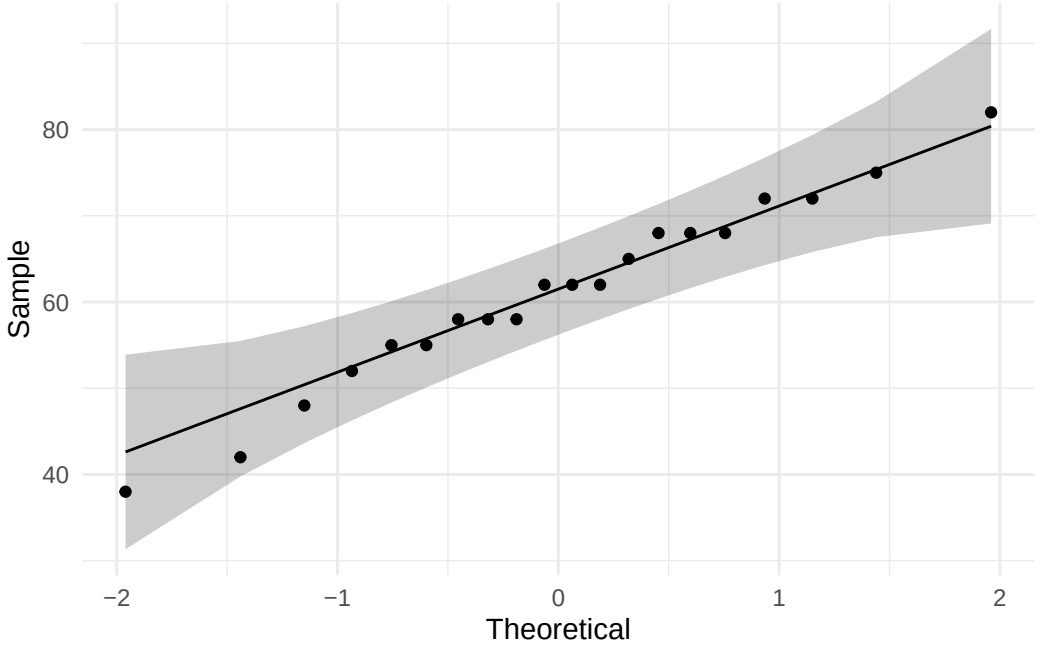


Figure 7: Assessing normality for `score`

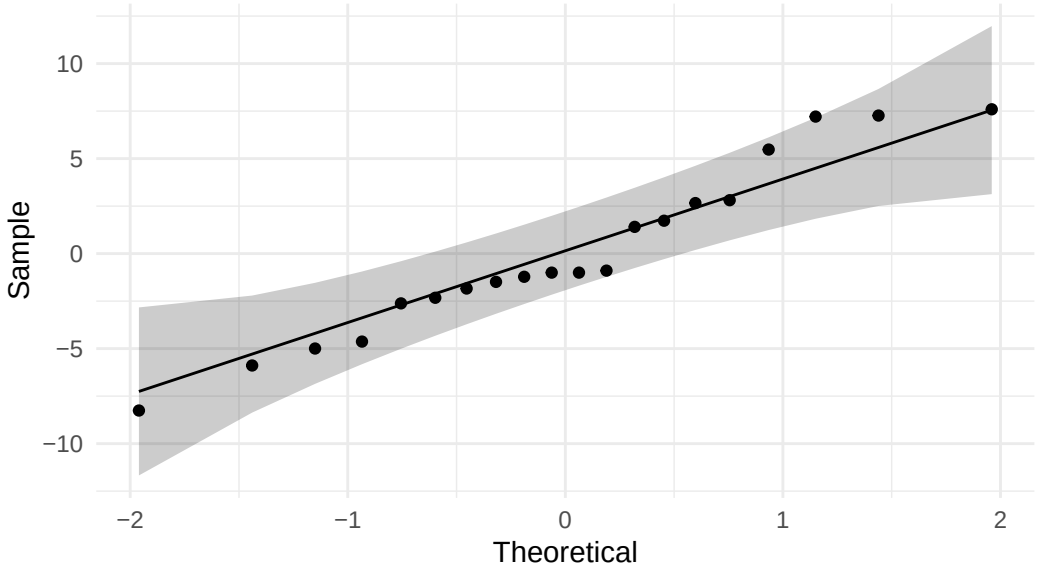


Figure 8: Assessing the normality of residuals

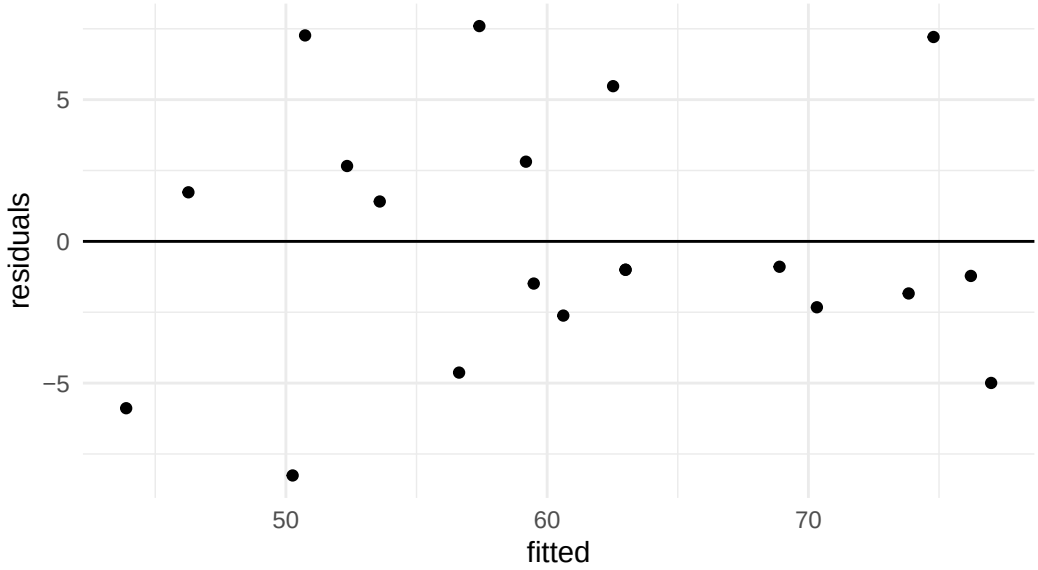


Figure 9: Assessing homoscedacity of residuals

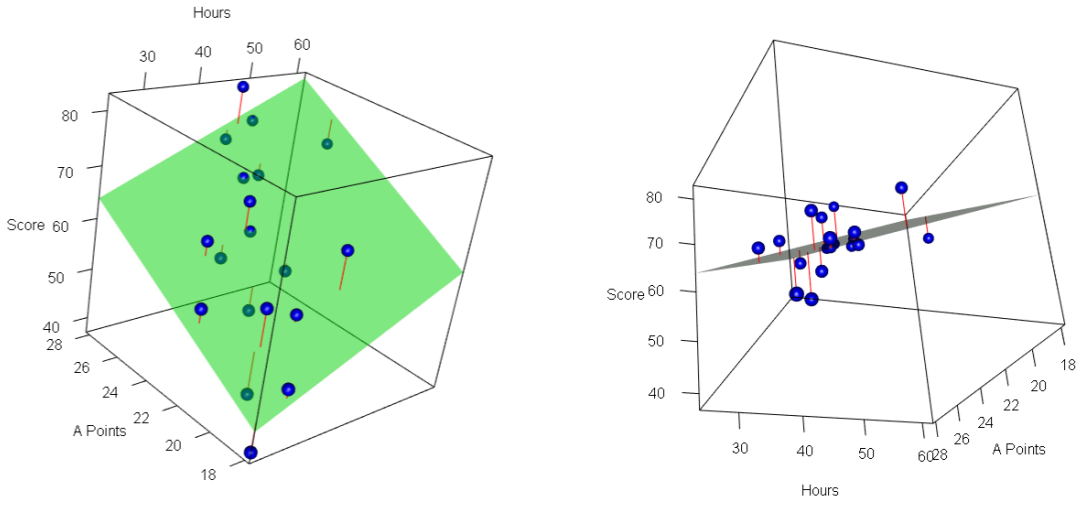
2.4 Model

Below is the result of a multiple regression analysis examining the relationship between `hours` and `a_points` as predictors of `score`. The regression model was statistically significant, with an F-statistic of 42.09 and a p-value of 2.604e-07, indicating that the model as a whole explains a significant portion of the variation in the dependent variable, `score`.

```
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|)
(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
```

The coefficient for `hours` was .4765 (t = 2.703, p = .0151), indicating that for each additional hour, the `score` is expected to increase by .4765, holding `a_points` constant. The coefficient for `a_points` was 1.9945 (t = 3.997, p = .000933), suggesting that for each additional point in `a_points`, the `score` is expected to increase by 1.9945, holding `hours` constant. The residual standard error was 4.751, and the model explained 83.2% of the variance in score (R-squared = 0.832).

Figure 10 presents the multiple regression model with `hours` and `a_points` as independent variables and `score` as the dependent variable. The surface represents the predicted score across different combinations of `hours` and `a_points`. The red line segments in the plot depict the residuals, which represent the vertical distance between the observed data points and the corresponding predicted values on the surface.



(a) (b)

Figure 10: 3D visualization of the multiple regression model examining the relationship between `hours` and `a_points` as predictors of `score`