

# Homework 5

## Bivariate Statistics One-way ANOVA and Regression Analysis

Andri Setiyawan

Benedikt Meyer

Yosep Dwi Kristanto

December 3, 2024

### The Data Detective: Alex's Adventures in Numbers

*The following story was created with the assistance of ChatGPT, and the full prompt and conversation used to develop it can be accessed at <https://chatgpt.com/share/6740a6fc-d3f8-800a-a0dd-6a22eb523458>.*



Alex Martinez wasn't your average statistician. Where others saw endless rows of numbers, Alex saw stories waiting to be told. Armed with a laptop, a passion for problem-solving, and an almost magical ability to make numbers come alive, Alex split their time

between two very different worlds: a bustling sales company and a university brimming with curious students.

## Cracking the Sales Script Code

Alex's first stop was the boardroom of Dynamic Sales Solutions, a company known for experimenting with creative ways to connect with customers. Their latest challenge? Figuring out which of their four sales scripts—dubbed Friendly Chat (Script A), Persuasive Pitch (Script B), Fact Frenzy (Script C), and Casual Catch-Up (Script D)—was most effective.

"We've got all this data from the past month," said Carla, the sales manager, waving a stack of papers. "But we're stumped on which script is truly the best."

Alex grinned. "Let's let the data do the talking."

After hours of crunching numbers and running an analysis of variance (ANOVA), see Section 1, Alex gathered the team around. On the screen, a colorful bar chart appeared, with Script A towering over the rest.

"Here's what we found," Alex explained, pointing to the bars. "Script A—Friendly Chat—brought in the most revenue. It's not just a little better; it's significantly better. Scripts B and C did okay, but Script D—Casual Catch-Up—lagged behind."

Carla frowned. "Why is Friendly Chat so effective?"

"It's simple," Alex said, clicking to a slide that summarized the findings. "Friendly Chat focuses on understanding customer needs. When sales reps actively listen and respond to customer concerns, they build trust. People are more likely to buy when they feel heard."

The sales team buzzed with excitement. "So we should use Script A more often?" asked one rep.

"Exactly," Alex nodded. "But remember, even the best script only works if you genuinely care about your customers. Numbers can guide us, but human connection seals the deal."

Numbers can guide us, but human connection seals the deal.

## Unlocking Academic Success

After a quick lunch, Alex headed to their next challenge: a university seminar where students were eager to understand what made some of them excel in exams while others struggled.

Professor Tanaka greeted Alex warmly. "Our students are curious about how their habits impact their grades. Can you help demystify it for them?"

Alex nodded, pulling up a scatter plot and a 3D graph showing the results of their analysis, see Section 2. "We looked at two factors," Alex began, addressing the class. "First, how many hours you spend revising each week. Second, your A-level entry points—basically, your academic background when you started here. Together, these predict your exam scores pretty well."

A student raised her hand. “How much do study hours really matter?”

“Great question,” Alex replied. “For every extra hour you spend studying, your score increases by 0.48 points on average. That might not seem like much, but it adds up. For example, if you study five more hours, you’ll gain about 2.4 points. That could bump you from a B to an A!”

Another student asked, “What about A-levels? Are they more important than studying?” “Well, each additional A-level entry point adds about 1.99 points to your score,” Alex explained. “But here’s the good news: no matter where you start, studying consistently makes a difference. Think of your A-levels as a foundation. The stronger it is, the easier it is to build on. But even with a shaky foundation, hard work can help you climb higher.” The students leaned forward, clearly inspired.

“Let me put it this way,” Alex said with a smile. “Imagine you’re planting a garden. Your A-levels are the soil. The richer the soil, the better your plants will grow. But watering the plants—your study hours—matters just as much. If you take care of the garden, it will bloom.”

### **Inspiring the Future**

After the seminar, a student named Jamie approached Alex. “I always thought I wasn’t smart enough to do well,” Jamie admitted. “But now I see that my effort counts too.” Alex beamed. “Absolutely, Jamie. Numbers don’t lie: hard work and preparation make a huge difference. Keep at it, and you’ll surprise yourself.”

As Alex packed up for the day, they reflected on how different their two clients were but how similar their lessons had been. Whether it was choosing the best sales script or helping students excel, the secret lay in paying attention to the story behind the numbers. Back at home, Alex typed up a blog post summarizing the day’s work. “Data is more than just numbers,” they wrote. “It’s a tool to understand the world, solve problems, and make better decisions. Whether you’re running a business or studying for an exam, remember: effort, curiosity, and a little data can take you a long way.”

[Data is] a tool to *understand the world, solve problems, and make better decisions.*

The post ended with a simple but powerful message: “Hard work pays off. Numbers prove it.”

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

## 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.

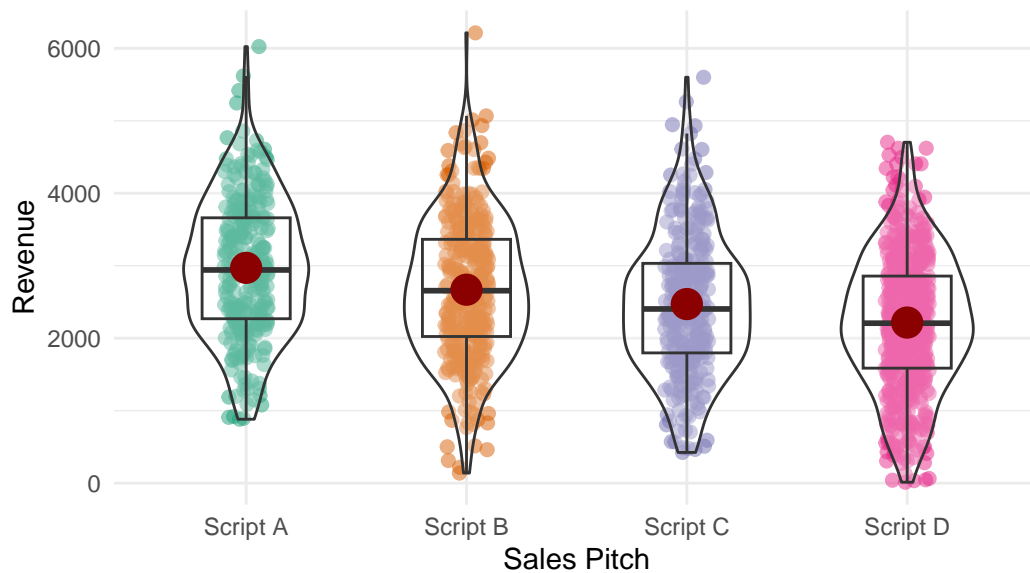


Figure 1: Distribution of **revenue** across **sales\_pitch** (Script A, Script B, Script C, and Script D) as illustrated by violin and boxplots.

## 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.

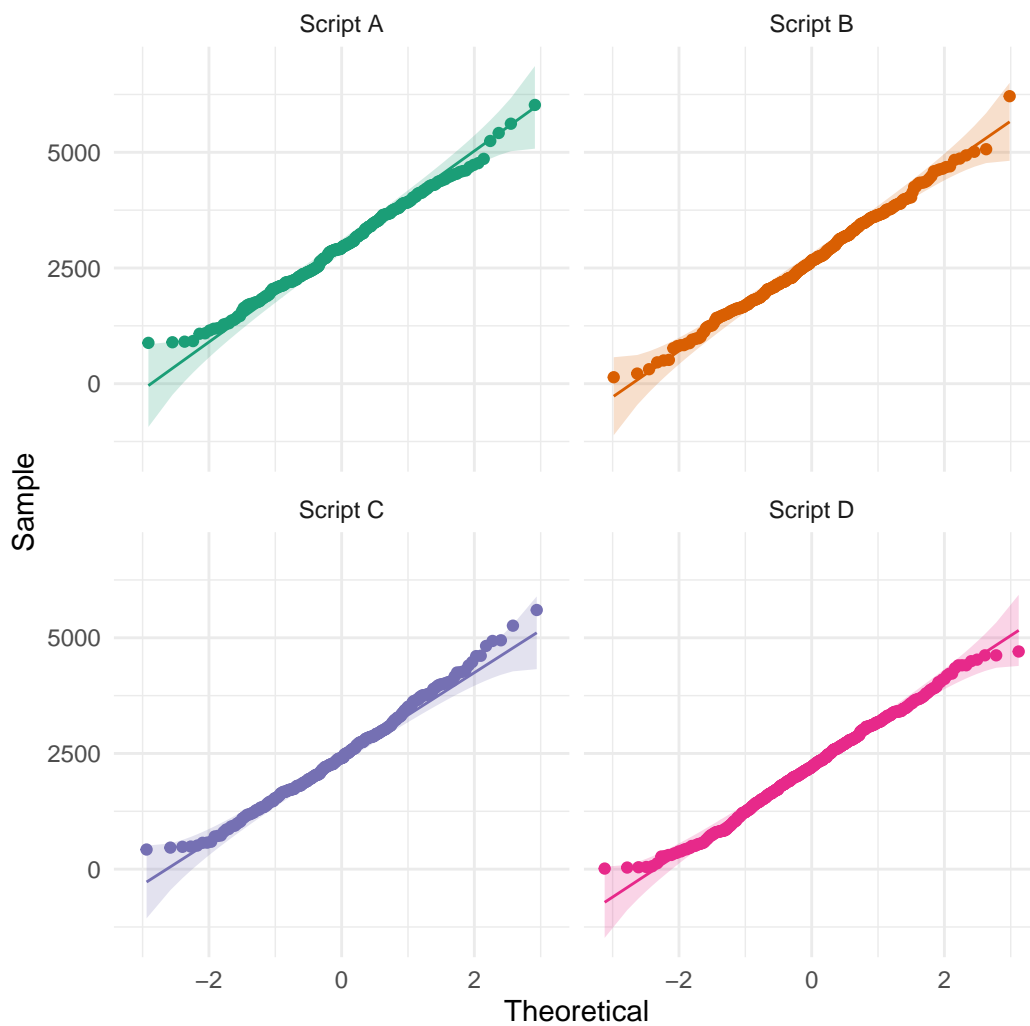


Figure 2: QQ plot of revenue across **sales\_pitch**

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  $2.38 \times 10^{-26}$ , which is less than .05. A visualization of the p-value is presented in Figure 3.

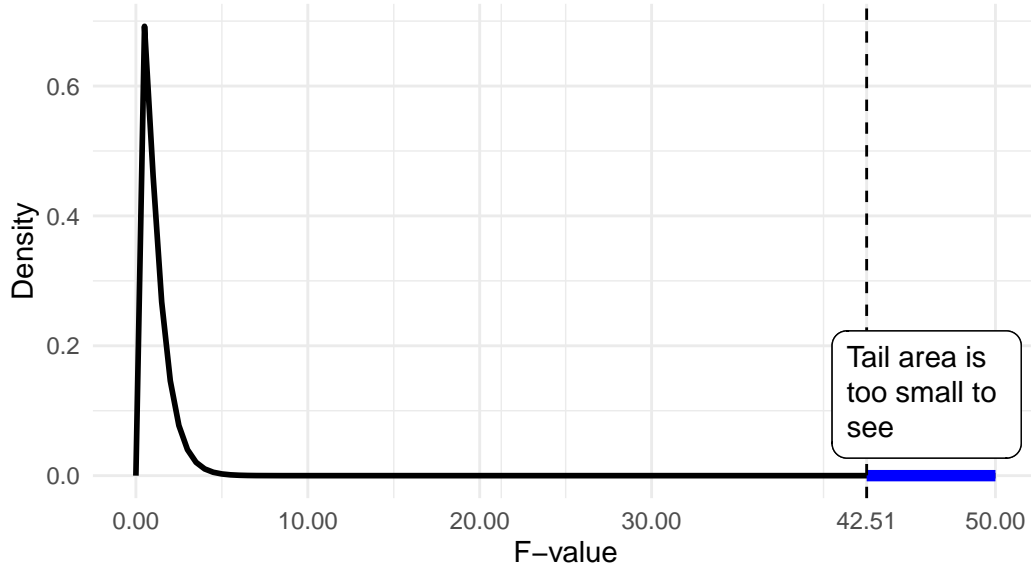


Figure 3: Theoretical  $F$ -distribution with degrees of freedom 3 and 1484, illustrating the tail area corresponding to the p-value

## 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  $2.4 \times 10^{-26}$ , 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 ( $\eta^2_{\text{ges}}$ ) 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 ( $\eta^2$ ) = .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.



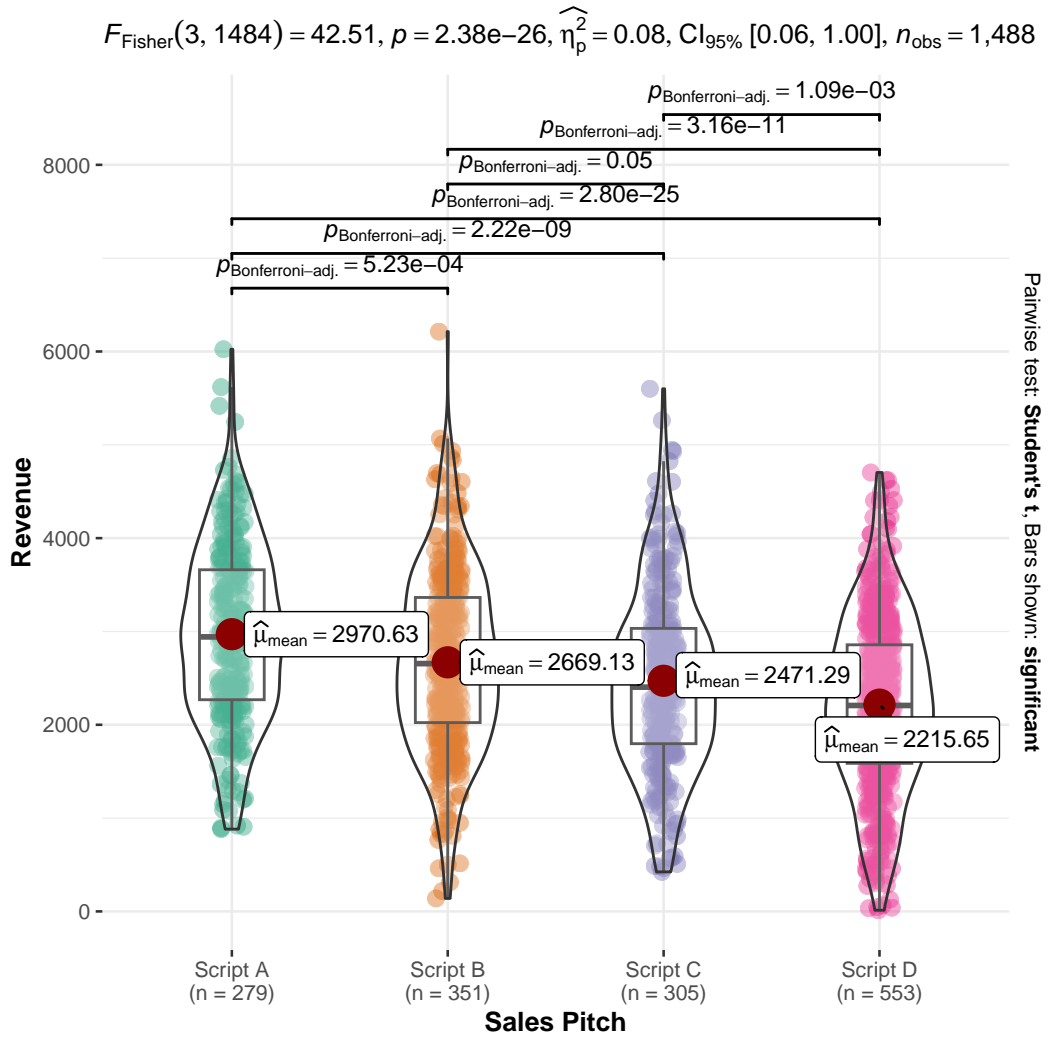


Figure 4: Violin plots and boxplots showing the distribution of revenue across the sales\_pitch groups, with results of pairwise t-tests indicating differences between groups.

## 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` ( $r = -0.12$ ).

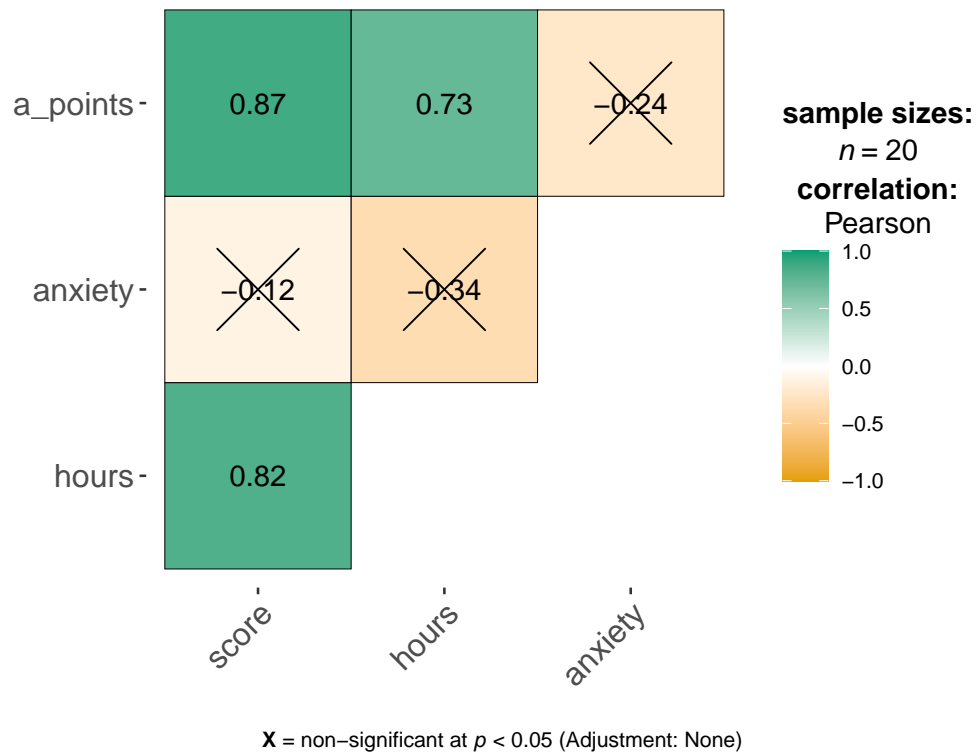


Figure 5: Correlation matrix among all variables in exam data

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.

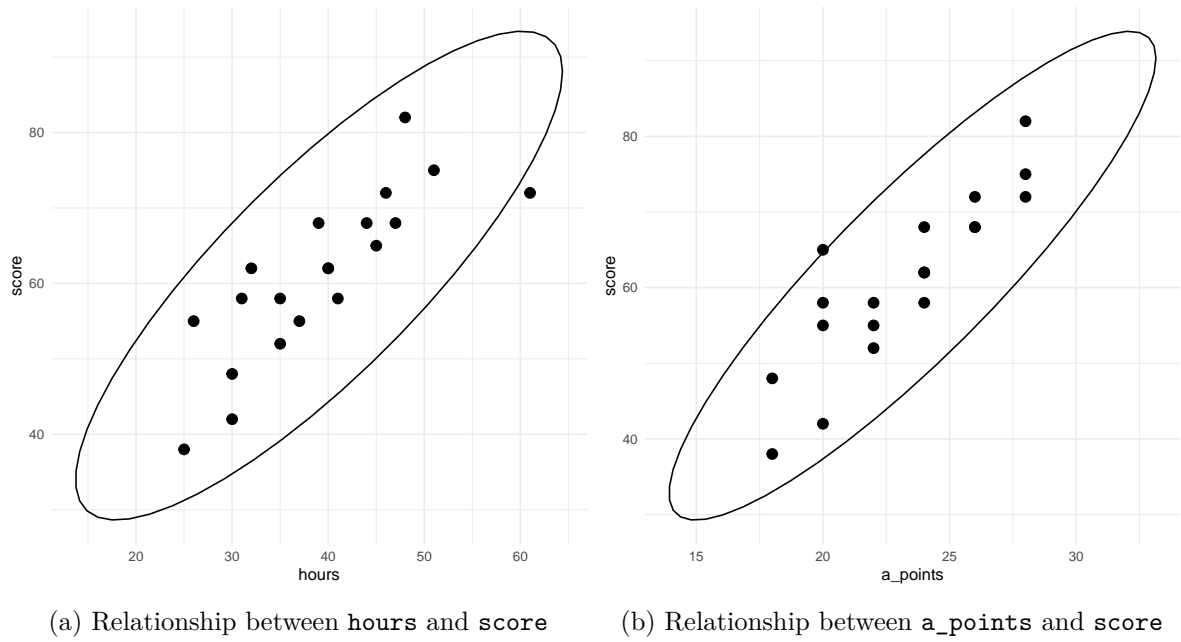


Figure 6: Linearty verification

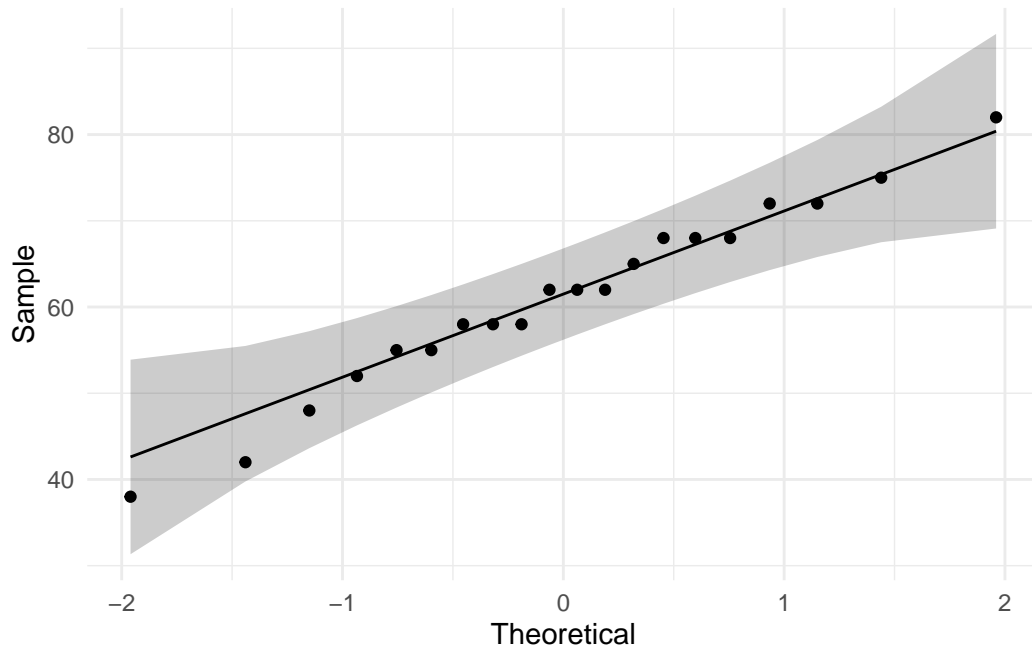


Figure 7: Assessing normality for `score`

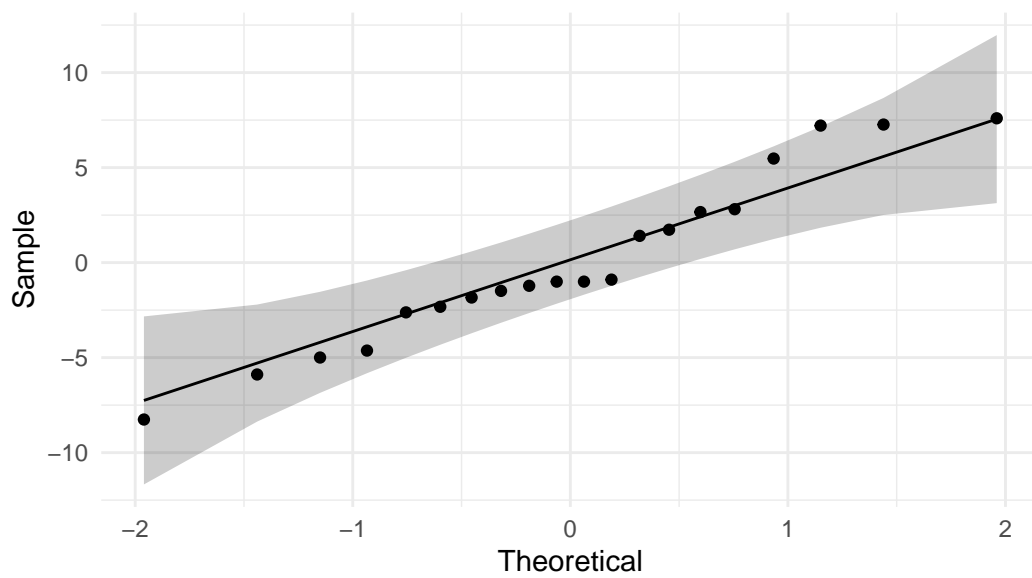


Figure 8: Assessing the normality of residuals

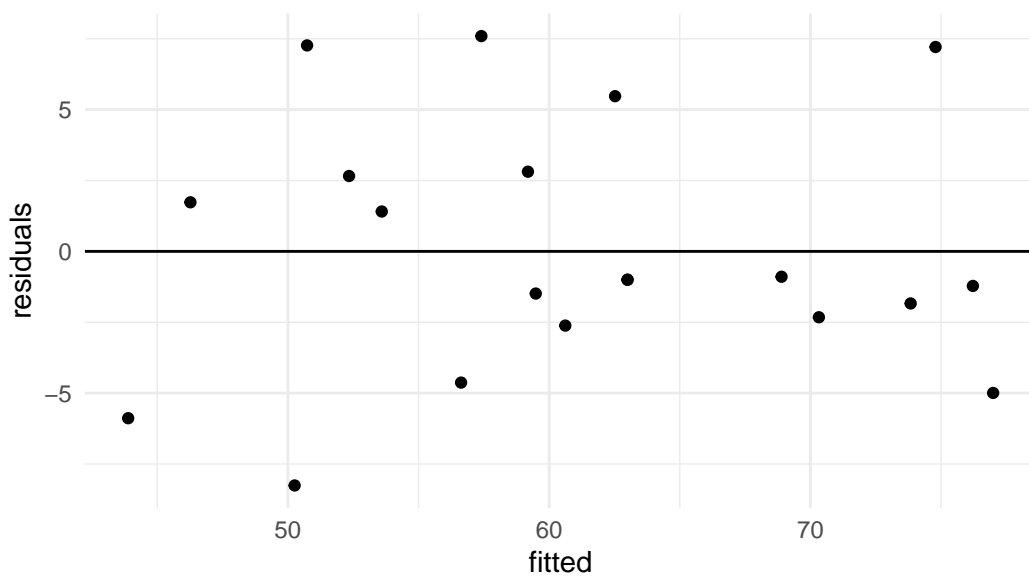


Figure 9: Assessing homoscedasticity of residuals

**Independence of residuals:** We use the Durbin-Watson test to assess whether the residuals of the model are independent. The results are as follows.

```
lag Autocorrelation D-W Statistic p-value
1      -0.6168051      3.222206  0.002
Alternative hypothesis: rho != 0
```

The Durbin-Watson test yielded a statistic of 3.22 ( $p = .002$ ) and a lag-1 autocorrelation of  $-.617$ , indicating significant negative autocorrelation in the residuals. This suggests the assumption of independent residuals is violated. However, given that the data is not time-series and the autocorrelation is relatively mild, we will proceed with the regression analysis.

**Outlier detection:** Figure 10 displays the Cook's distance plot, which is used to assess influential data points. No points had a Cook's distance greater than 1, indicating that all observations exerted minimal influence on the regression model, suggesting that the results are stable and not unduly affected by any single data point.

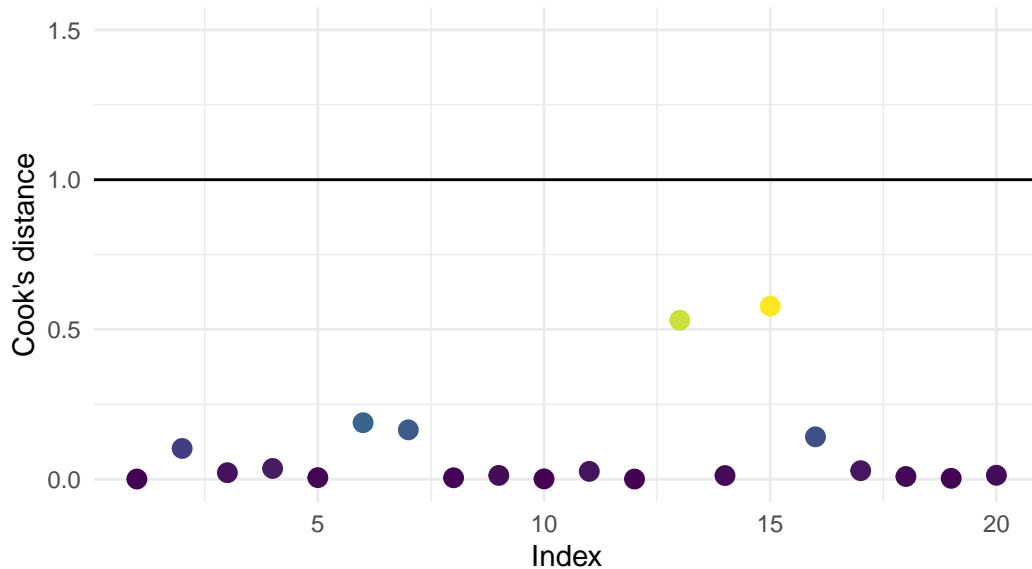


Figure 10: Plot of Cook's distance for each observation in the regression model. The dashed line represents a threshold of Cook's distance = 1, above which data points may be considered influential and potentially affect the regression results.

## 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

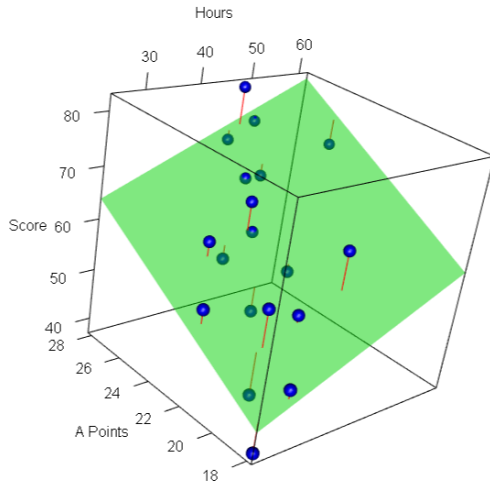
F-statistic: 42.09 on 2 and 17 DF, p-value: 2.604e-07

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` ( $R\text{-squared} = 0.832$ ).

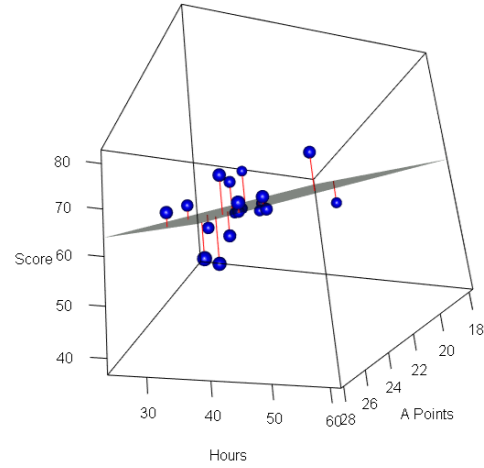
Figure 11 presents the multiple regression model with `hours` and `a_points` as independent variables and `score` as the dependent variable. The surface represents the predicted score across different combinations of `hours` and `a_points`. The red line segments in the plot depict the residuals, which represent the vertical distance between the observed data points and the corresponding predicted values on the surface.

## 2.5 Reporting the Results

A multiple regression analysis was conducted to examine whether hours spent revising (`hours`) and A-level entry points (`a_points`) predicted exam scores (`score`). The results showed that the overall model was significant,  $F(2, 17) = 42.09$ ,  $p < .001$ , and explained 83.2% of the variance in exam scores ( $R^2 = .832$ , adjusted  $R^2 = .812$ ).



(a)



(b)

Figure 11: 3D visualization of the multiple regression model examining the relationship between **hours** and **a\_points** as predictors of **score**

Individually, both predictors significantly contributed to the model. For hours spent revising,  $b = .48$ ,  $t(17) = 2.70$ ,  $p = .015$ , indicating that for each additional hour spent revising, exam scores increased by an average of .48 points. For A-level entry points,  $b = 1.99$ ,  $t(17) = 4.00$ ,  $p < .001$ , suggesting that each additional entry point was associated with an average increase of 1.99 points in exam scores. The intercept was not statistically significant,  $b = -3.93$ ,  $t(17) = -0.48$ ,  $p = .635$ .

The residual standard error was 4.75, indicating the average deviation of observed scores from predicted scores. These findings underscore the importance of study hours and prior academic achievement in predicting exam performance.