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.

Table 1: Summary statistics of `revenue` in each `sales_pitch` in 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

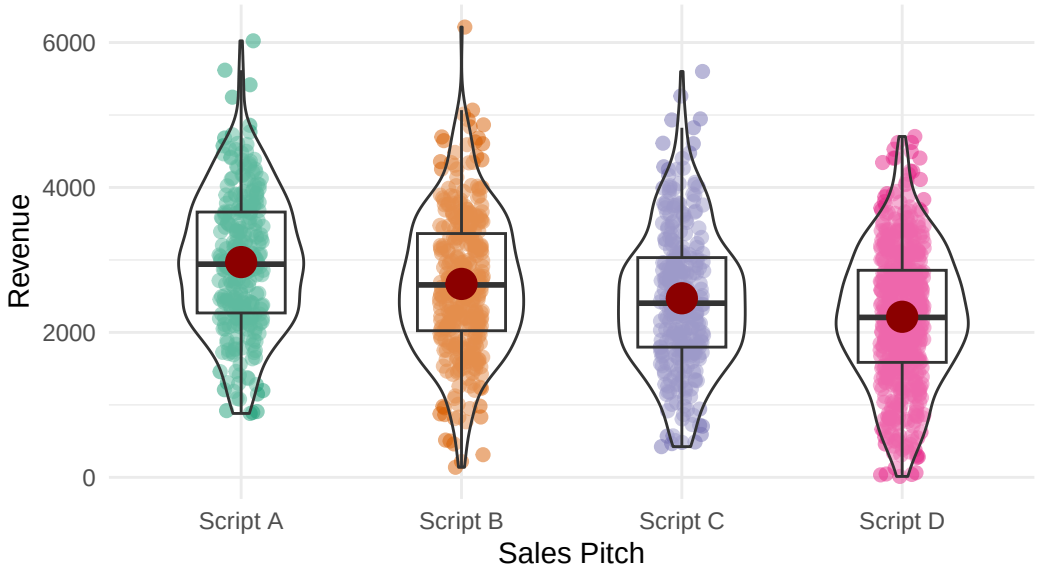


Figure 1: Distribution of `revenue` in each `sales_pitch` in telemarketing data

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

1 Telemarketing

1.1 Data Exploration

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.
- Here

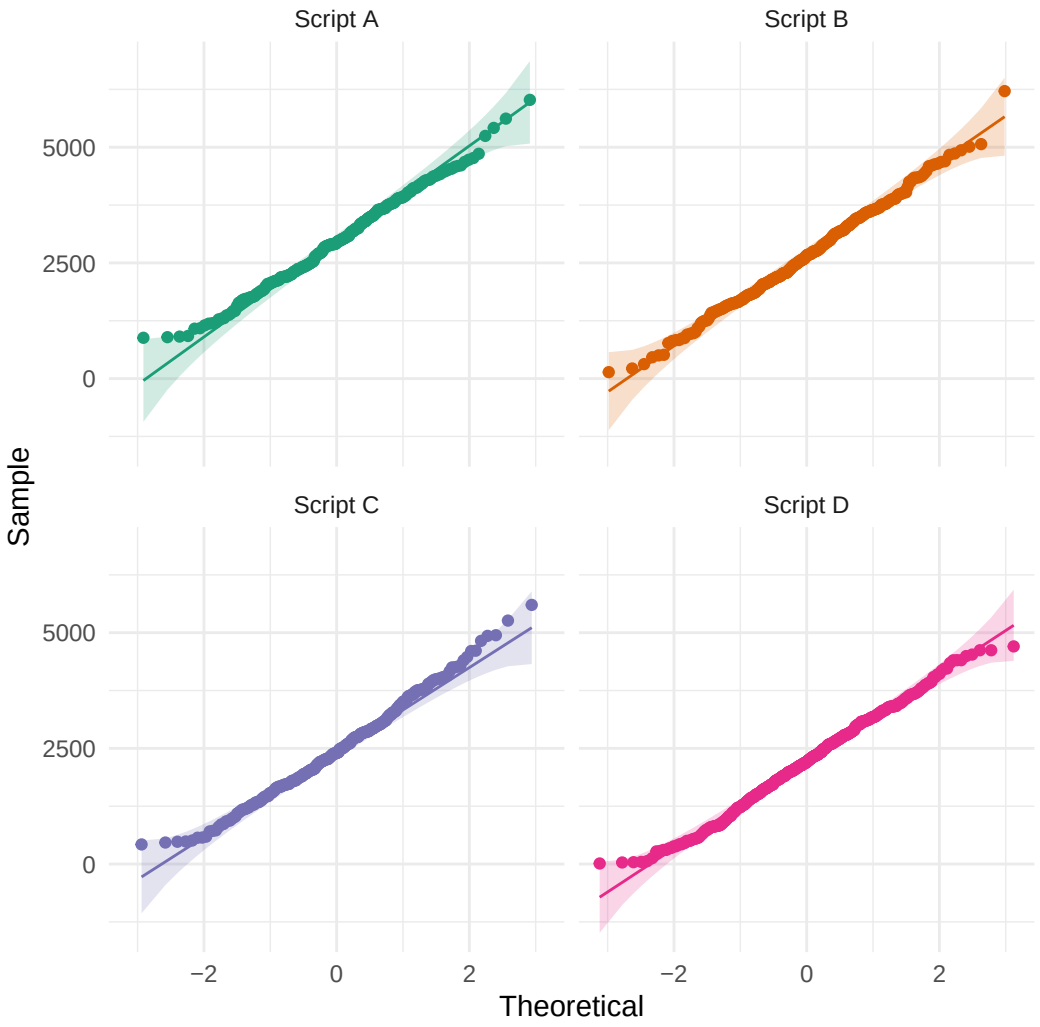


Figure 2: QQ plot of `revenue` across `sales_pitch`

Table 3: Results of Levene test for homogeneity of variance

df1	df2	statistic	p
3	1484	0.0924258	0.9642314

Table 4: ANOVA table for testing the `revenue` difference across `sales_pitch`

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

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

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

1.5 Testing for the significance of F

1.6 Interpreting F

1.7 Post-hoc test

2 Students' Performance

2.1 Data Exploration

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.

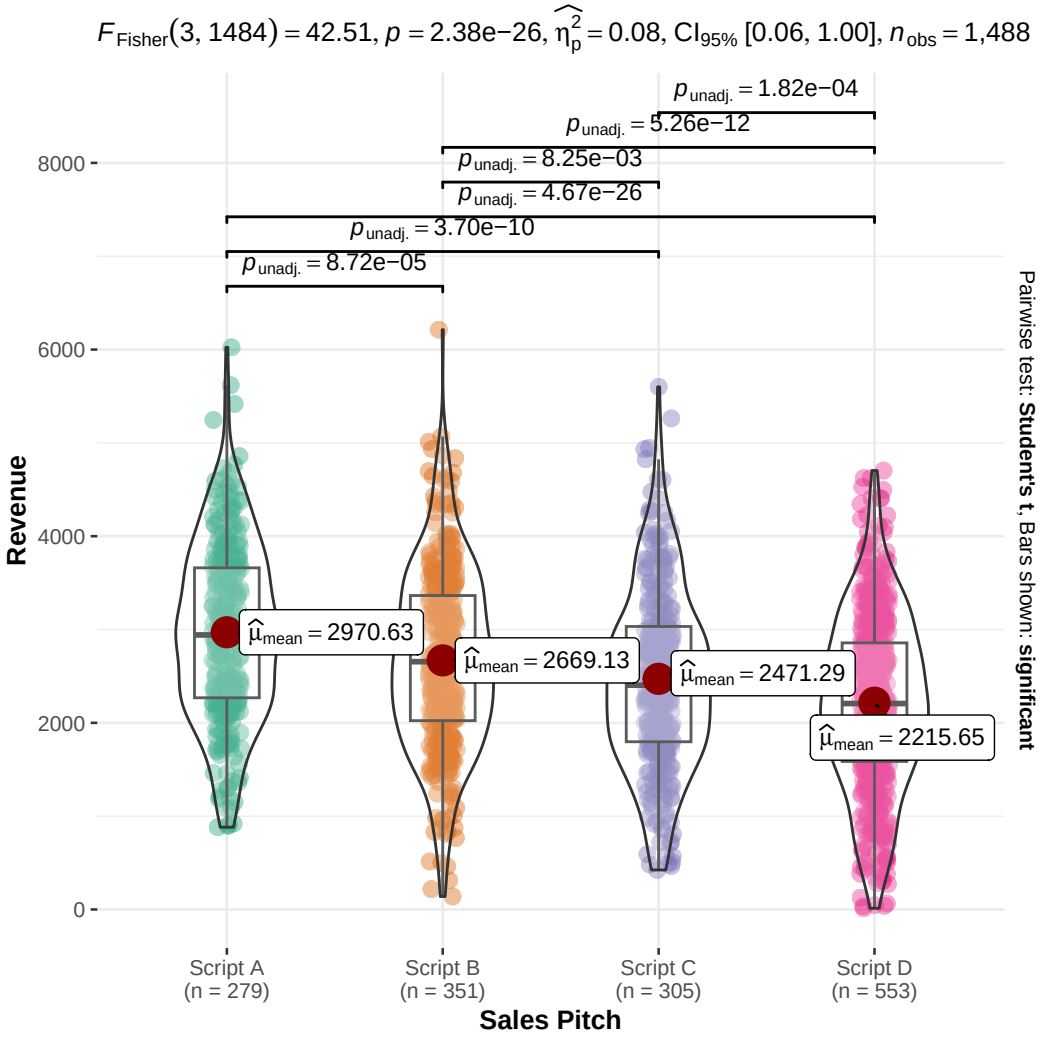


Figure 3: ANOVA and post-hoc test results for telemarketing data

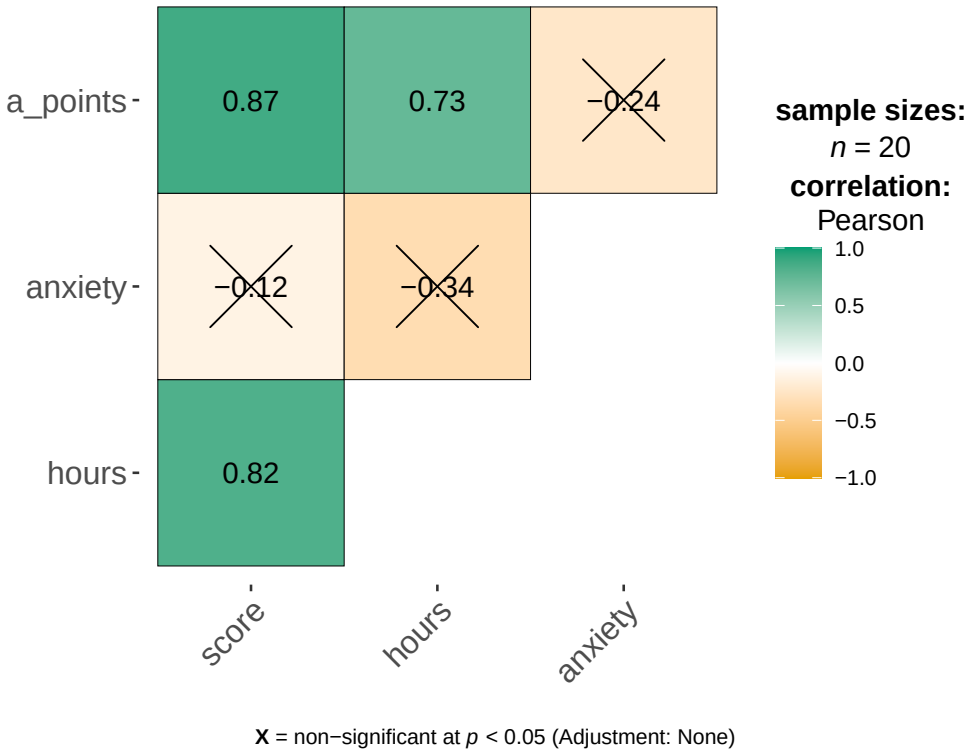
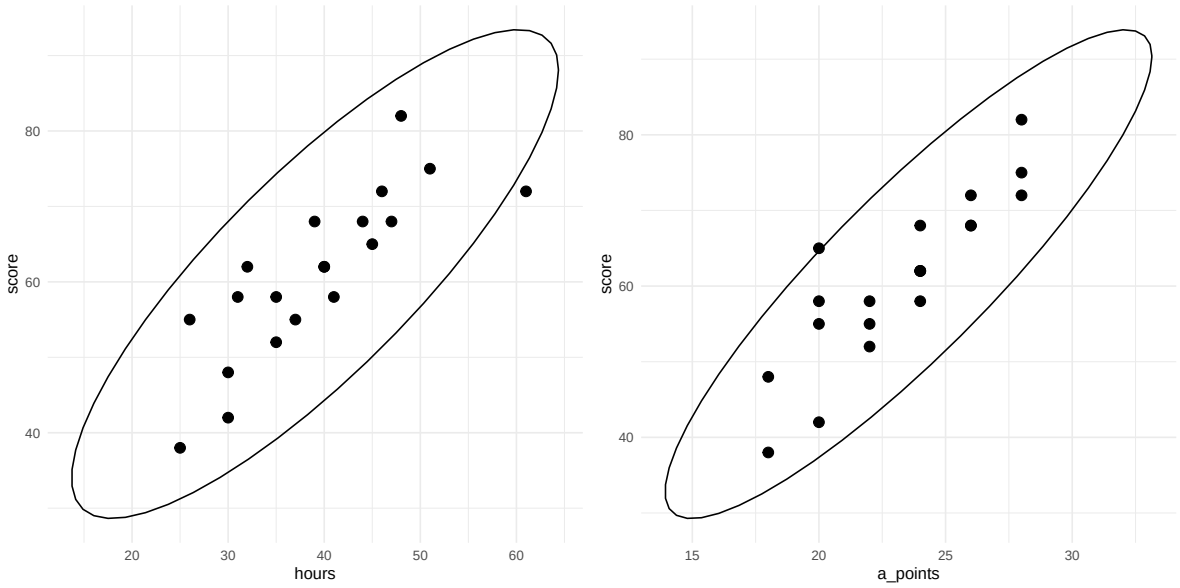


Figure 4: Correlation matrix among all variables in exam data

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 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
```

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.

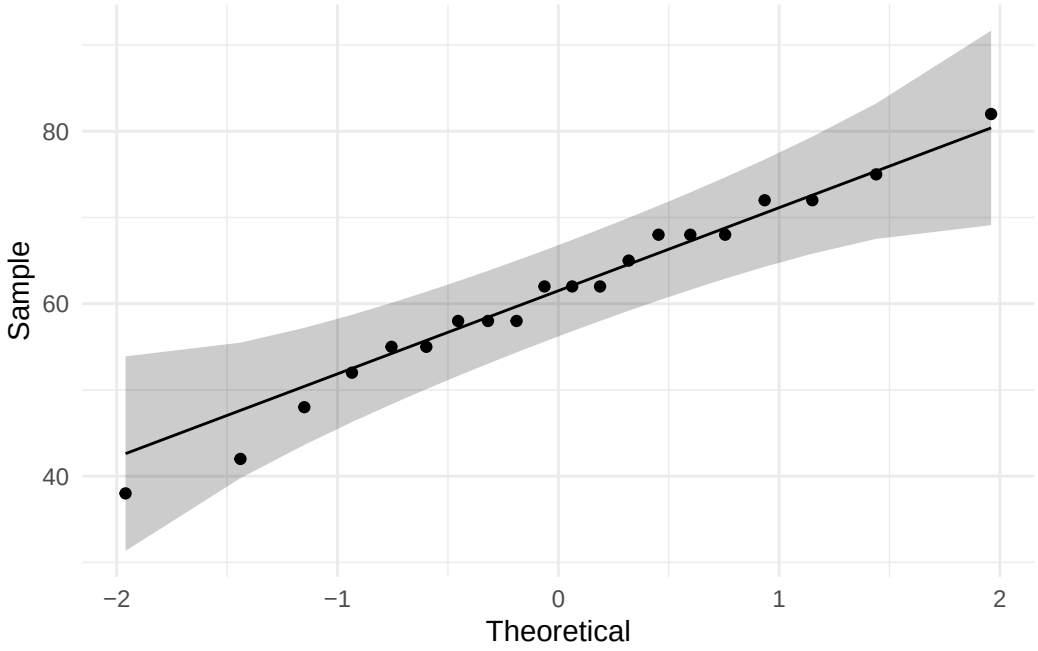


Figure 6: Assessing normality for `score`

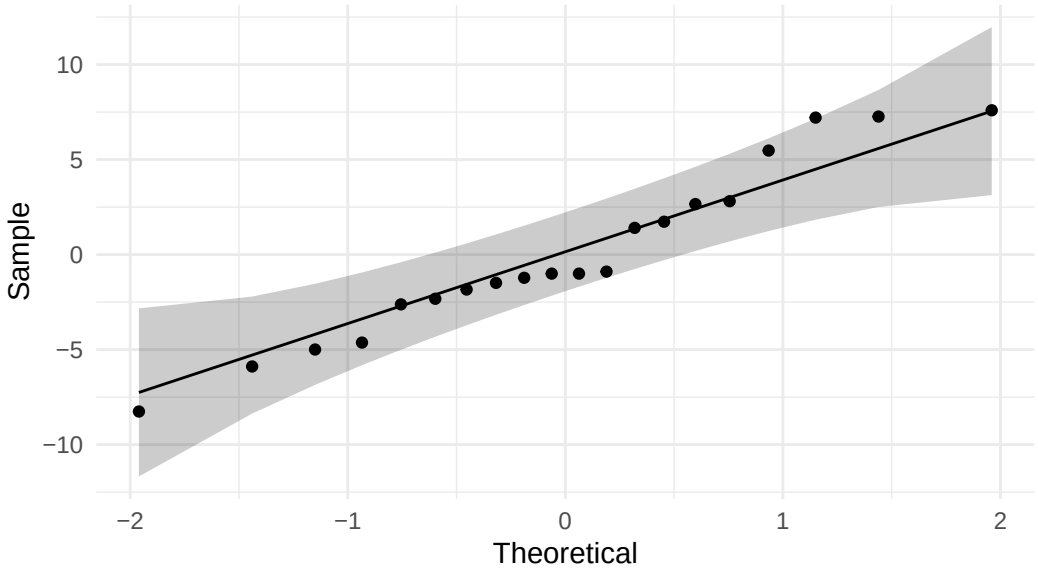


Figure 7: Assessing the normality of residuals

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

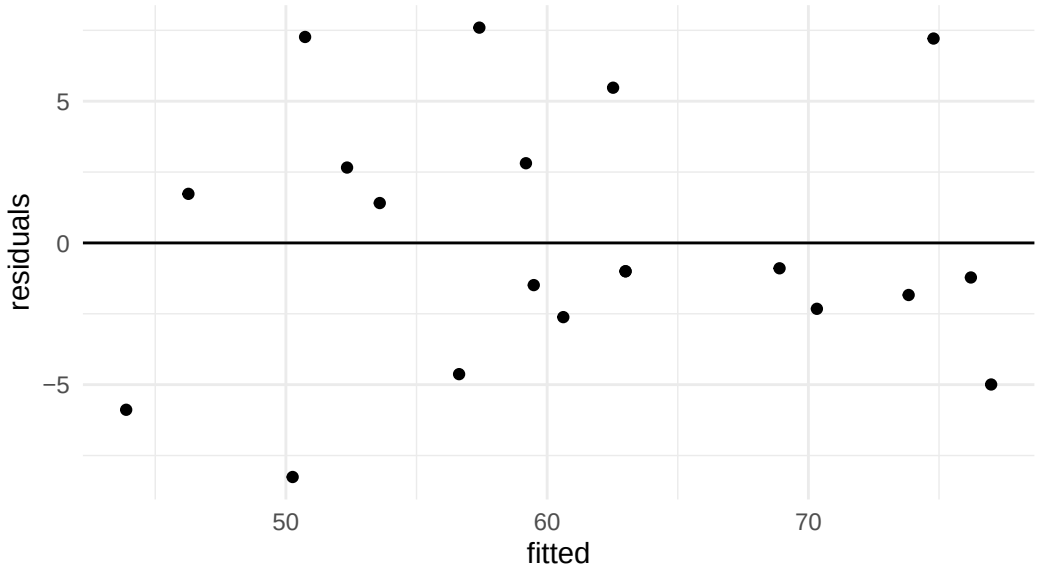


Figure 8: Assessing homoscedacity of residuals