PO91Q - Fundamentals in Quantitative Research Methods

Mock Exam 2025/26 - Solutions

Dr Florian Linke
Florian.Linke@warwick.ac.uk

1 | Block A

Level 1 (each question 4 points)

1. `header=TRUE` tells R that the first row of the CSV file contains the names of the columns.
    - This allows the resulting data frame to use meaningful variable names (e.g., crime_rate, drive_primary) instead of default names like V1, V2, etc.
    - If `header=FALSE`, R treats the first row as data rather than column names, which can misalign the dataset and require manual renaming of columns.

---

2. `tidyverse` or `dplyr`.

---

3. The code `simd2 <-filter(simd, crime_rate < 400)` selects rows from the `simd` dataset where the `crime_rate` is less than 400. It creates a new dataset simd2 containing only these filtered observations.
    - In plain terms: “Keep only the observations where the crime rate is below 400 and store them in `simd2`.”

Level 2 (each question 6 points)

4. If we change to `right = TRUE` and whilst preserving the same category assignments we had with `right = FALSE`, we need to shift the breakpoints slightly downward, because:
    - With `right = FALSE`: a value of 2.79 would be assigned to the next, “Avergae” category.
    - When we change this to `right = TRUE`, a value of 2.79 would now fall into the “Short” category. To move it back to “Average” we need to adjust the cut-off to slightly below this value

```
simd2 %>%
mutate(drive_cat = cut(drive_primary,
                       breaks = c(0, 2.79 - 1e-8, 3.572 - 1e-8, 30),
                       labels = c("Short", "Average", "Long"),
                       right = TRUE,
                       include.lowest = TRUE)) -> simd2
```

5. The output shows the result of a one-sample test in R.
    - The one-sample t-test compares the mean crime_rate in `simd2` to the hypothesized value of 200.
    - The sample mean is 177.65, which is less than 200.
    - The t-statistic is -16.103, indicating the sample mean is many standard errors below 200.
    - The p-value is less than 2.2e-16, meaning the result is highly statistically significant.
    - The 95% confidence interval for the true mean ranges from 174.92 to 180.37, which does not include 200.
    - Interpretation: There is extremely strong evidence that the average crime rate in the filtered dataset is lower than 200. The difference is unlikely to be due to random sampling.

---

6. The average increase in the standardised mortality ratio is 0.215 for every additional unit of alcohol-related admissions to hospital.

---

7. The average standardised mortality ratio in a data zone with a low rate of alcohol-related admissions to hospital is 74.014.

---

8. Very likely because the recoding into a binary dummy takes away nuance that pushes the model fit in Model 1 higher.

Level 3 (each question 8 points)

9. In Subsection 7.6, the sample `simd_sample$drive_primary` is used in a one-sample t-test with $\mu = 5$ and a very large negative t-statistic ($t = -97.62$).

    - The t-statistic formula is:

    $$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

    - Rearranging for the sample mean:

    $$\bar{x} = t \cdot \frac{s}{\sqrt{n}} + \mu$$

    - Given:
        - $t = -97.62$
        - $s = 1.34018$
        - $n = 4500$
        - $\mu = 5$

    - Calculate the standard error:

    $$se = \frac{1.34018}{\sqrt{4500}} \approx \frac{1.34018}{67.082} \approx 0.01997$$

- Calculate the sample mean:

$$\bar{x} = -97.62 \times 0.01997 + 5 \approx -1.949 + 5 \approx 3.051$$

- The sample mean of `simd$drive_primary` ≈ 3.051

---

10. Mean and median are quite close together, so there is no indication of a strong skew. This gives no reason for a non-linear transformation. But a scatter plot would give us more information to answer the question more confidently.

2 | Block B

1. With a smaller effect size of interest, a higher sample size is required to achieve the same level of statistical power. — True

    - If the target effect size is smaller, you need larger n to make effect $\cdot \sqrt{n}$ large enough to detect it at the same power.
    - In practice n scales approximately with $1/(\text{effect size})^2$ for simple tests, so halving the effect size requires about four times the sample size to hold power constant.

---

2. The p-value and the α-level are the same thing. — False

    - The p-value is a data-dependent quantity: the probability, under the null hypothesis, of observing data as extreme or more extreme than what was observed.
    - The α-level is a pre-specified cut-off (e.g., 0.05) that the researcher chooses before the test; it sets the maximum tolerated probability of a Type I error.
    - You compare p to α (reject if $p \leq \alpha$), but p and α are conceptually and functionally distinct: one is observed evidence, the other is a decision threshold.

---

3. Adjusted R^2 ($\bar{\text{R}}^2$) can be negative. — True

    - Unlike R^2, which is bounded between 0 and 1, $\bar{R}^2$ adjusts for the number of predictors relative to the sample size.
    - When the model fits poorly — that is, when adding predictors does not improve explanatory power enough to offset the penalty — $\bar{R}^2$ can fall below 0.
    - A negative $\bar{R}^2$ indicates that the model fits the data worse than a horizontal line at $\bar{y}$ (the mean-only model).

3 | Block C

i	x_i	y_i	$y - \bar{y}$	$(y - \bar{y})^2$	$x - \bar{x}$	$(x - \bar{x})^2$	$(y - \bar{y})(x - \bar{x})$
1	4	10	2.60	6.7600	-5.80	33.6400	-15.0800
2	7	13	5.60	31.3600	-2.80	7.8400	-15.6800
3	8	9	1.60	2.5600	-1.80	3.2400	-2.8800
4	11	5	-2.40	5.7600	1.20	1.4400	-2.8800
5	19	0	-7.40	54.7600	9.20	84.6400	-68.0800
MEAN	9.8	7.4					
SUM			0.00	101.2000	0.00	130.8000	-104.6000

Slope

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{-104.6}{130.8} = -0.799694 \approx -0.800$$

Intercept

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 7.40 - (-0.799694)(9.80) = 15.2370 \approx 15.237$$

Sample regression function

$$\hat{y}_i = 15.237 - 0.800\, x_i$$

Block D

1. Non-Linear Transformation of Variables

- **What it is:** A non-linear transformation applies a mathematical function (e.g., log, $\sqrt{\ }$, x^2, $1/x$) to a variable to change its scale or relationship with other variables. These transformations can make non-linear relationships more linear or stabilize variance.
- **Example (worked):** Imagine a dataset where income (x) and spending (y) are related, but the effect of income on spending gets smaller as income grows. The scatterplot curves upward, not straight. Taking $\log(x)$ and $\log(y)$ straightens that curve into a line, allowing a linear regression to capture the pattern accurately.
- **Why it matters:** Transformations can improve model fit and make coefficients interpretable in percentage or elasticity terms.
- **What it is not:** It is not the same as fitting a "non-linear model" in parameters; rather, it's a linear model estimated on transformed variables.

---

2. Statistical Power

- **What it is:** Statistical power is the probability of correctly rejecting a false null hypothesis (i.e., detecting an effect if it exists). It equals $1 - \beta$, where β is the probability of a Type II error.
- **Example (worked):** Suppose a researcher tests whether a new teaching method increases student test scores. The null hypothesis is that the mean score is 70, and the alternative is that it is higher. If the true mean with the new method is 75, and the standard deviation is 10, taking a sample of 25 students gives $se = 10/\sqrt{25} = 2$. Using a significance level $\alpha = 0.05$, the critical value for rejecting H_0 is approximately 73.92. The probability that the sample mean exceeds 73.92 given the true mean 75 is the power.
- **Why it matters:** Higher power reduces the risk of Type II errors and ensures that studies are more likely to detect meaningful effects.
- **What it is not:** Power is not the probability that the null hypothesis is true or false; it is a property of the test given a true alternative effect size.