Analysis of the observational plan

---

25 marks

1 Problem

Recall the physical laboratory involving the plane. (See the file background.pdf for more information.)

The equation of the plane can be written as

$$y = \alpha + \beta x + \gamma z$$

which has no error. All of the points lie exactly on the plane and both β and γ are unknown. The planes were arranged so that $\beta = 0.5$ for every run.

Values of x and y were recorded by teams for three different experimental protocols or plans. The third value z was not. In practice, the values of z are never known – there is always some variate that is not measured, perhaps not even thought of, that might be part of relating y to x. These are called **lurking variables** and they will always exist.

The purpose of this question is to investigate and compare the different experimental plans. Of particular interest is whether or not x causes y, that is testing $H_0 : \beta = 0$. And if so, to estimate the value of β defining the causal relationship.

2 Plan

Three experimental plans were considered:

- `"observational"` where six (x, y) pairs were observed in a particular configuration by each team.
- `"randomized"` where three tower-markers were randomly allocated to each of only two different x values, from which the ys were determined. Each team produced two replicates here.
- `"randomizedBlock"` where tower markers were sorted into pairs by height and one marker of each pair were randomly assigned the lower of the two x values and the other to the higher x value. Again, y was determined after allocation for all six markers.

In this question, you will be working only with the data collected using the **observational** plan. See the file background.pdf for more information.

3 Data

Set up the following:

```
## Set this up for your own directory
imageDirectory <- "MyAssignmentDirectory/img"  # e.g. in current "./img"
dataDirectory <- "MyAssignmentDirectory/data"  # e.g. in current "./data"
path_concat <- function(path1, ..., sep="/") paste(path1, ..., sep = sep)
```

The full data set is then read in as:

```
labData <- read.csv(file = path_concat(dataDirectory, "labData.csv"))
```

The data can be subsetted according to the three different experimental plans.

a. *(1 mark)* Select that subset of the data corresponding to the observational plan. Assign it to the variable `observational`. Show your code.

4 Analysis

b. *(4 marks)* Plot the (x, y) pairs from all of the observational data

Use `xlim = c(0, 30), ylim = c(0,30), pch = 19, col = adjustcolor("black", 0.3)` in the call to `plot()`.

Label the plot meaningfully.

Fit a straight line model of y on x and add this fitted line to the plot. Save the fit object. Report the value of the slope estimate.

Show your code.

c. **The measuring system.** Recall the model fitted in part (b).

i. *(3 marks)* Plot the residuals (on the vertical axis) against the x values (on the horizontal axis).

Use `xlim = c(0, 30)`.

Make sure the plot is meaningfully labelled. Add a horizontal line at 0.

Show your code.

ii. *(2 marks)* Based on what you see in the above residual plot, what would you say about the measuring system for y? Justify your answer.

d. **Learning from repetition** Each team executed the same plan. To gain a better appreciation of the qualities of that plan, we investigate the individual team estimates of β.

i. *(4 marks)* First fit a separate line for each team's data. Capture the slope estimate of each fit and collect these into a single vector. Report the average of the estimated slopes.

Show your code.

ii. *(3 marks)* Recall that in fitting a straight line model of y on x, the least-squares estimate of the slope coefficient for x, using n points is

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

You will have already observed that the average of the slopes for the individual team data is identical to the slope estimate of the observational data combined (i.e. ignoring the teams).

Must this be so? A little more notation may help.

Let the teams be indexed $j = 1, \ldots, J$, and each team has exactly m pairs of observations which will be indexed by i. Then, the slope estimate for the jth team's data pairs (x_{ij}, y_{ij}), for $i = 1, \ldots, m$ is just

$$\hat{\beta}_j = \frac{\sum_{i=1}^{m}(x_{ij} - \overline{x}_j)y_{ij}}{\sum_{i=1}^{m}(x_{ij} - \overline{x}_j)^2}$$

where $\overline{x}_j = (\sum_{i=1}^{m} x_{ij})/m$ is the average of the xs from team j.

Explain mathematically, or otherwise prove, why in this study we have

$$\frac{\sum_{j=1}^{J} \hat{\beta}_j}{J} = \hat{\beta} \quad \text{from above with} \quad n = m \times J.$$

iii. *(3 marks)* Draw a meaningfully labelled histogram of the individual slope coefficient estimates for all teams.

Show your code.

Use `xlim = c(-1, 1), col = "lightgrey"` in `hist()` and an appropriate `main` title and `xlab`.

Add a vertical red dashed line at the average of the slope estimates.

Add a vertical blue dashed line at the true value of β.

5 Conclusion

e. *(3 marks)* What do you conclude about the quality of team slope estimates from the observational study?

f. *(2 marks)* What effect, if any, has been produced by a lurking variable? Explain.