Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
Exam 2025-07-07 - Part B - 2024/2025

Problem 2: Modeling viewing time across languages

A tech company is analyzing how much time users spend watching their newly released *Functional Training* workout series, which is available in 10 different language versions on their streaming platform.

Data has been collected from 1000 user accounts, with 100 users for each language. The goal is to predict the number of hours a user will spend watching the workout series, denoted by the variable `Views`, using the following predictors:

- `Premium_account` $\in \{0, 1\}$: indicates whether the user has a premium (ad-free) subscription.
- `Laptop_time`: average number of hours per day the user spends on a laptop.
- `Phone_time`: average number of hours per day the user spends on a phone.
- `Social_connections`: number of friends the user has on the platform.
- `Fitness_level` $\in \{0, 1\}$: user's self-declared fitness level (0 = beginner, 1 = advanced).

a) Fit a linear model, referred to as **M0**, assuming independent and identically distributed observations and no interaction terms. In this model, `Views` is predicted using the variables `Premium_account`, a linear combination of device usage defined as (`Laptop_time` + $\frac{1}{2}$`Phone_time`), `Social_connections`, and `Fitness_level`.

Report the estimated coefficients for (`Laptop_time` + $\frac{1}{2}$`Phone_time`) and `Social_connections`.

Then, test whether there is sufficient evidence at the 1% significance level to claim that `Social_connections` has a negative effect on `Views`.

b) Based on model **M0**, compute a 95% prediction interval *[lower, upper]* for `Views` for a beginner user (fitness level = 0) watching the English version of the series, who has a premium account, spends on average 5 hours daily on a laptop and 2 hours on a phone, and has 10 friends on the platform.

c) Extend model **M0** to a new model **M1**, accounting for the language in which the content is viewed. Specifically, introduce a random effect modelling a direct influence on the number of hours spent watching the series.

Report the total number of parameters to be estimated in **M1**.

d) Report the residual standard error from model **M1**. Also, identify the language group associated with the highest expected `Views`, after adjusting for all other covariates.

e) Propose a further extension, model **M2**, that builds upon **M1** by accounting for potential heteroskedasticity: namely, the possibility that the variance of `Views` differs by `Fitness_level`. Formulate the model accordingly to test whether variability in viewing time is higher for beginners or advanced users.

Compare models **M1** and **M2** using appropriate quantitative criteria. Which model offers the best fit to the data?

Upload your results here: `https://forms.office.com/e/eswucY0KpZ`