Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
January 20, 2025

Problem n.2

A tech company is analyzing the users' viewing time of their latest *Functional Training* workout series, streamed on their platform in 10 different languages. The company collects data for 1000 accounts (100 accounts per language). The objective is to predict `Views`, the number of hours a user will watch *Functional Training*, based on the following factors: premium account status (`Premium_account` $\in \{0, 1\}$, where 1 indicates a premium account, with no advertisement), average daily time on the laptop (`Laptop_time`, in hours) and on the phone (`Phone_time`, in hours) declared by the user, number of friends on the platform (`Social_connections`), and fitness level (`Fitness_level` $\in \{0, 1\}$, categorized as 0 for beginner and 1 for advanced).

a) Fit a model **M0** without interactions, assuming the data are independent and identically distributed. In this model, `Views` is predicted based on the variables `Premium_account`, (`Laptop_time` + $\frac{1}{2}$`Phone_time`), `Social_connections`, and `Fitness_level`.

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

Test whether we can affirm with 99% confidence that `Social_connections` has a negative effect on `Views`.

b) Compute and report the 95% prediction interval *[lower, upper]* for `Views` through **M0** for a beginner English user with a premium account, who stays 5 hours on average on the laptop and 2 on the phone, and has 10 friends on the platform.

c) Update now **M0** into a new model **M1**, that you deem the most appropriate, able to account for the grouping of users induced by the different languages.

How do **M0** and **M1** differ (e.g., in terms of formulation, parameters to be estimated,...)?

d) Report the residual standard error of the model **M1**, and the language according to which the `Views` are higher, all the other factors being the same.

e) Propose a new model **M2**, modifying **M1**, able to account the fact that the variability in the outcome changes with `Fitness_level`. Specifically, you want to test whether views are more variable for participants who start at a different fitness level.

Compare **M1** and **M2**, quantitatively supporting your answer. What is the best model?

Upload your solution https://forms.office.com/e/XKTgEMKqhr