PSYCH 10 Final Project · Fall 2026

Week 7 · M6 · The model

Full name (in CAPITAL LETTERS) _____ **SUNet ID** (the part before @stanford.edu) _____ **Section** (e.g., TUES 9:30) _____

Please write legibly so the teaching team can log your attendance: if we can't read your name, we can't give you credit.

Please hand this sheet in to your TA at the end of section (each student should hand in their own copy). A copy is on the course website (psych10.github.io).

Where this milestone fits

In M4, your group posed two questions; in M5, you cleaned the dataset and plotted each variable and the question 1 relationship. In M6, you will fit a model that answers question 1 with numbers: how strong is the relationship? Everything you produce in M4, M5, and M6 will go into your final report, so please write each answer as a sentence you could paste into the report.

Deliverables

Each group should submit two files on Gradescope (the link is on Canvas): the group's `M6-the-model.Rmd`, from the `milestones/` folder, and the knit PDF. One member of the group should upload both files and add the other members to the submission. By Sunday, your group's shared folder should also hold the M6 `.Rmd`.

Steps

Your group will do most of these steps in section and finish the rest together before Sunday. Every function on this sheet appears in CourseKata, up to Chapter 9, so if one looks unfamiliar, please look it up in CourseKata. Step 7, the shuffles, can take several minutes to run on a large dataset. The `.Rmd` has a place for every step on this sheet.

1. **Get the latest version of your folder.** Download your group's shared folder from Google Drive, so that your local project folder has everyone's latest changes: the earlier milestone `.Rmd` files, the raw data, `R/prepare-data.R`, and the processed data file in `data/processed/`. (If you download single files instead of the whole folder, please put each one in the same place it has in the shared folder.) Then download `M6-the-model.Rmd` from the course website into the `milestones/` folder. Decide who will upload the changed files at the end of section.
2. **Load your processed data and restate your questions.** In the first chunk of `M6-the-model.Rmd`, read `processed-data.csv` from `data/processed/`, the file your script wrote in M5, not the raw file in `data/raw/`. Then copy in your two questions from M4 and your M4 prediction for question 1: the direction of the relationship (which group is higher, or whether the outcome rises or falls with the explanatory variable) and its rough size in the units of the outcome.
3. **Fit the empty model.** Fit `lm(outcome ~ NULL, data = processed)`, the model that predicts the same value for every case. Report its one estimate, `b0`, in the units of your outcome, and say in a sentence what that number is (the mean of the outcome, the empty model's prediction for every case). The empty model is the baseline that the question 1 model has to improve on in step 6.
4. **Fit the model for question 1.** Fit `lm(outcome ~ explanatory, data = processed)`, with the outcome and explanatory variable from your question 1, and plot the model on the data: `gf_model()` added to a jitter or box plot for a categorical explanatory variable, or `gf_lm()` added to a scatter plot for a quantitative explanatory variable.
5. **Say what the estimates mean, in the units of your data.** For a group model (categorical explanatory variable): `b0` is the mean of the reference group, so name that group, and `b1` is the difference between the second group's mean and the reference group's mean (a model with three or more groups has one more `b` for each additional group). For a regression model (quantitative explanatory variable): `b1` is the predicted change in the outcome for each one-unit increase in the explanatory variable, and `b0` is the predicted outcome when the explanatory variable equals 0, so say whether 0 lies inside the range of your explanatory variable. Write each estimate as a sentence with its number and units, for example *"students at Stanford sleep an average of 0.5 hours per night more than students at UCSF."*
6. **Measure the strength of the relationship.** Run `supernova()` on your question 1 model and report PRE and F from the table. PRE is the proportion by which your model reduces the empty model's error, so say in a sentence what that proportion means for your question, for example *"hours of sleep account for 12% of the variation in exam scores."* If your explanatory variable has exactly two groups, also report Cohen's d from `cohensD()`.

7. **Could chance alone produce a b1 of this size?** Save your real `b1` with `sample_b1 <- b1(outcome ~ explanatory, data = processed)`. Then shuffle the outcome a thousand times, which breaks any relationship between the two variables, and save the `b1` from each shuffle: `shuffled_b1s <- do(1000) * b1(shuffle(outcome) ~ explanatory, data = processed)`. (If `do(1000)` takes more than a few minutes on your dataset, use `do(100)` in section and rerun with 1000 at home.) Make a histogram of the shuffled `b1`s with your real `b1` marked on the x-axis, as in the `.Rmd`. Then compute the proportion of shuffled `b1`s at least as far from 0 as your real `b1`, in either direction, and say in a sentence what that proportion means: if the outcome had no relationship with the explanatory variable, how often would a shuffle alone produce a `b1` at least as far from 0 as your real `b1`?
8. **Compare the result with your M4 prediction.** Did the direction match the direction you predicted in M4? Was the size close to the size you predicted, larger, or smaller? Report the comparison in a sentence with both numbers, the predicted size and the fitted `b1`. If the result differs from the prediction, that difference is one of your findings; please do not revise the prediction to match the result.
9. **Report what you have found so far for question 2.** Your model has only one explanatory variable, because we will not cover models with two explanatory variables until the last week of classes. If the other variable in your question 2 is in your dataset, look at the question 1 relationship within each group of that variable (if that variable is quantitative, first split it into two or three groups at cut points you state in the `.Rmd`), for example by fitting the same model separately to each group or by faceting your step 4 plot with `gf_facet_grid()`, and say whether `b1` is about the same in every group or differs between groups in its direction or size. If that variable is not in your dataset, say what pattern you would expect to see across its groups if you had the variable you described in M5.
10. **Say which of your M4 DGPs could still be at work.** Go back to the data-generating processes (DGPs) you wrote in M4. For each one, say whether your results so far rule it out, are consistent with it, or cannot distinguish it from the other DGPs, and give the reason. Use all of your evidence: the model and the shuffles from this milestone; your question 2 plot or table from M5 (step 10) and from step 9 here, which bear most directly on the DGP behind your second question; and the variables you listed in M4 as the ones that could distinguish your DGPs. If the shuffles rarely produce a `b1` as far from 0 as your real `b1`, chance alone is an unlikely explanation for the relationship, but a rare shuffle result does not rule out a third variable or the reverse direction.
11. **Knit to PDF in a fresh session** (Session → Restart R, then Knit → Knit to PDF). One member should submit the `.Rmd` and the PDF on Gradescope and add the other members to the submission.
12. **Upload the M6 `.Rmd` to the shared folder.**

Using AI tools on this milestone

The course encourages you to use AI tools on the project. This week, you may use an AI tool to help you write and debug the code on this sheet: the two models, the plots, the shuffles, and the proportion in step 7. To get code you can use, give the tool the `str()` output for your processed data and a plain-language description of what you want, one step at a time. Every member of the group should be able to explain the purpose of each line of code and the meaning of each number in the output, because the same functions and tables will be on the practice exam on Monday, Nov 9. Please write the sentences that interpret the estimates, the PRE, the shuffle proportion, and the comparison with your prediction in your own words.

How it's graded (2 points)

	Points
The empty model and the question 1 model fitted from the processed data, with the model plotted on the data	0.3
`b0` and `b1` interpreted in sentences, with numbers and the outcome's units	0.4
PRE and F from `supernova()` (Cohen's d for two groups), and a sentence on the meaning of the PRE	0.3
The shuffled `b1`s with the real `b1` marked, the proportion at least as far from 0, and a sentence on its meaning	0.3
The comparison with the M4 prediction: direction and size, with both numbers	0.3
Question 2 so far, and each M4 DGP marked ruled out, consistent, or not yet distinguishable from the others	0.4