Generative AI for Data Analysis

Writing code, method descriptions, and important considerations

Greta Linse, Sara Mannheimer, Ava Yazdian, Sally Slipher

1

Montana State University Library

Guidance on responsible use of generative AI in academic work

2

Using Generative AI:

Writing code and more for data analysis

3

Example use cases

- Translating code from one language to another
- Writing code given a set of instructions/steps/pseudo code
- Create code for a desired plot by describing it in words or by giving it an example

4

Translating code from one language to another

- PROMPT: Given the following code translate to R.
- Demonstration:
  - Translate NHANES Sample Code for Logistic Regression that is only available in SAS-callable SUDAAN version 11
- FOLLOW UP PROMPT 1: What will the output look like?
- FOLLOW UP PROMPT 2: Write a results paragraph using these results.

5

SAS code

```
*****************************************************************;
    *Linear trends analysis by age group using Logistic Regression*;
*****************************************************************;

proc rlogist data = Rx FILETYPE = SAS DESIGN = WR ;
* Nest statement: PSUs nested within Strata accounts for the design effects*;
 NEST SDMVSTRA SDMVPSU / MISSUNIT;
 * Weight statement: specify appropriate weight, accounts for the unequal probability of sampling and non-response.*;
 WEIGHT wtint2yr;
 * Class statement: specify categorical variable(s). Reflevel statement can be included to choose reference category
 for the categorical variables. By default SUDAAN uses the highest category*;
 class one /nofreq;
* Subpopx statement: specify the subpopulation of interest (the inclusion criteria)*;
 subpopx inAnalysis;
 * Model statement: specifies dependent variable and independent variable(s) *;
 MODEL  anyOpioid = ageCat;
 * Output statement: outputs the results to a file*;
 output / betas=default filename=fig3_logistReg_betas_all replace;
   * Test statement: produces statistics and P values for the Satterthwaite adjusted CHI square (satadjchi),
 the Satterthwaite adjusted F (satadjf), and Satterthwaite adjusted degrees of freedom (printed by default).
 If this statement is omitted, the nominal degrees of freedom,
 the WALDF and the p-value corresponding to the WALDF and WALDP will be produced.*;
 test waldf satadjf satadjchi;
run;
```

6

Resulting R code

```
library(survey)

# Assuming Rx is your data frame and has the following columns:
# - SDMVSTRA: Stratum
# - SDMVPSU: Primary Sampling Unit (PSU)
# - wtint2yr: Survey weight
# - one: Categorical variable
# - ageCat: Age category
# - anyOpioid: Dependent variable (binary)
# - inAnalysis: Subset indicator

# Create survey design object
survey_design <- svydesign(
  ids = ~SDMVPSU,
  strata = ~SDMVSTRA,
  weights = ~wtint2yr,
  data = subset(Rx, inAnalysis == 1)
)

# Fit logistic regression
model <- svyglm(
  anyOpioid ~ ageCat,
  design = survey_design,
  family = quasibinomial()
)

# Summary of the model
summary(model)

# Optional: Extract and save coefficients
betas <- coef(model)
write.csv(betas, "fig3_logistReg_betas_all.csv", row.names = TRUE)
```


The output will look similar to this:

```
Call:
svyglm(formula = anyOpioid ~ ageCat, design = survey_design, family = quasibin

Survey design:
n = number of observations
n.clusters = number of PSUs
population sizes: w1, w2, ...

Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)    X.XXX       X.XXX      X.XX     X.XX
ageCat1        X.XXX       X.XXX      X.XX     X.XX
ageCat2        X.XXX       X.XXX      X.XX     X.XX

(Dispersion parameter for quasibinomial family taken to be X.XXX)

Quasi-likelihood R-squared: X.XXX
```

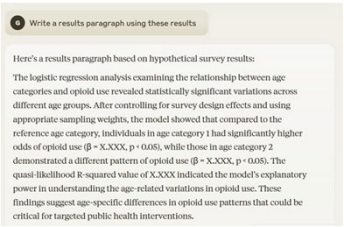
Key elements include:

- Coefficients for each age category
- Standard errors
- t-values
- p-values

Response to: What will the output look like?
Asked of Claud.AI on November 21, 2024

7

Writing up the results


Write a results paragraph using these results

Here's a results paragraph based on hypothetical survey results:

The logistic regression analysis examining the relationship between age categories and opioid use revealed statistically significant variations across different age groups. After controlling for survey design effects and using appropriate sampling weights, the model showed that compared to the reference age category, individuals in age category 1 had significantly higher odds of opioid use (β = X.XXX, p < 0.05), while those in age category 2 demonstrated a different pattern of opioid use (β = X.XXX, p < 0.05). The quasi-likelihood R-squared value of X.XXX indicated the model's explanatory power in understanding the age-related variations in opioid use. These findings suggest age-specific differences in opioid use patterns that could be critical for targeted public health interventions.

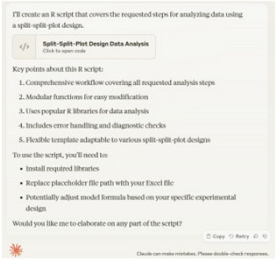
Response to: Write a results paragraph using these results.
Asked of Claud.AI on November 21, 2024

8

Writing code given a set of instructions

- PROMPT: Write R code that does the following steps:

1. Import data from an Excel file
2. Check for missing values
3. Perform exploratory data analysis
4. Generate a statistical model for a split-split-plot design
5. Check model assumptions
6. Summarize the results


I'll create an R script that covers the requested steps for analyzing data using a split-split-plot design.

Split-Split-Plot Design Data Analysis
Click to open code

Key points about this R script:

1. Comprehensive workflow covering all requested analysis steps
2. Modular functions for easy modification
3. Uses popular R libraries for data analysis
4. Includes error handling and diagnostic checks
5. Flexible template adaptable to various split-split-plot designs

To use the script, you'll need to:

- Install required libraries
- Replace placeholder file path with your Excel file
- Potentially adjust model formula based on your specific experimental design

Would you like me to elaborate on any part of the script?

Copy Retry

Claude can make mistakes. Please double-check responses.

9

Resulting R Code

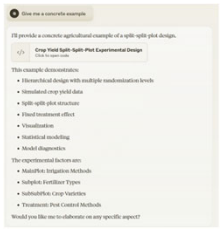
10

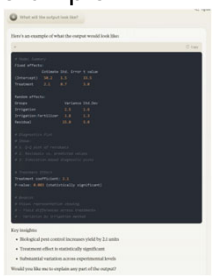
Knowledge required

- Need to understand how to install packages
  - What is lme4 and DHARMa? Are they actually popular libraries?
- The suggested code uses functions with messages and errors. Seems pretty complicated!
- Have to read the code carefully because there are some optional code chunks that are included.
- Realized that the treatment was included as a random effect instead of a fixed effect.
- Follow up Prompt 1: What is MainPlot?
- Follow up Prompt 2: Add a fixed treatment effect
- Follow up Prompt 3: Give me a concrete example
- Follow up Prompt 4: What will the output look like?

11

Output from second example





12