PHP2510

ASSIGNMENT #3 WEEKS 12-14

PURPOSE

Regression is the statistical tool we use to model, predict, and understand the complex relationships in the world around us. The purpose of this assignment is to provide practical, hands-on experience in the complete workflow of linear regression, moving from a raw dataset and a research question to a fully diagnosed, refined, and interpreted model. This assignment will also help you master the art and science of model diagnostics, learn to justify complex modeling decisions, and apply your R coding skills to a realistic data challenge. The group challenge will also encourage collaboration and cooperation with your peers and learning from each other.

Topics Covered: Multiple Linear Regression, Hypothesis Testing (in the context of Regression)

INSTRUCTIONS

I have generated 4 complex public health datasets. Two teams have been randomly formed and assigned to each dataset, creating a head-to-head competition.

Your team must explore the data and investigate:

- Transformations: Is the outcome skewed? Are some predictors skewed? Consider log(), sqrt(), etc.
- Non-linearity: Does a predictor have a non-linear relationship with the outcome? Consider quadratic terms.
- Interactions: Does the effect of one predictor depend on the level of another?
- Variable Coding: How should you handle categorical variables? Ordinal variables?Model Diagnostics: Once you have a "candidate" model, you must inspect the residuals. Are they normally distributed? Is there homoscedasticity (constant variance)? Are there high-leverage points? Your model is not valid if it violates key assumptions.
- Hypothesis Testings: How can I confirm my model adjustments are meaningful?

These synthetic datasets were built using a "true" underlying model. Your job is to get as close to that true model as possible. This true model will be statistically sound, parsimonious, and tell an interpretable story about the data.

Good luck, and may the best model win!

GRADING & SUBMISSION

Each team captain will submit **one** PDF via Canvas that includes:

- Introduction & Team:
    - List your team name, dataset, and team members
- The Process (10 pts): Explain your process of model building:
    - How did you explore the data? What did you find?
    - What candidate models did you try? Show your thought process
    - What hypothesis tests did you perform? How did this impact your process?
- The Final Model (10 pts): Present your final chosen model:
    - Show the summary() output

    - Write out the final model equation
    - Explain why this is your final model
- Interpretation (10 pts):
    - Generally explain the functional form of your model and it's use of complex terms or transformations
    - Interpret the meaning of at least three coefficients from your final model
- Diagnostics (10 pts):
    - Discuss the diagnostics for your final model
    - Include and describe the standard diagnostic plots
    - State your conclusions about model fit
- What We Learned (10 pts):
    - A brief paragraph on what your team learned from this assignment

Additionally, each individual should use this form to state: "I attest that all team members contributed meaningfully to this assignment." If this is not the case, please respond accordingly in the form and I will determine appropriate grading for the team.

A winning team will be determined based on which model is "closest" to the true model (known from my data generation process); ties will be broken based on the quality of the overall submission. Each member of each winning team will receive 10 bonus points on the assignment (thereby making a grade of 60/50 = 120% possible).

DATA SETS & TEAMS

All datasets are available in Canvas Files (Data -> Assignment #3)

Dataset	Teams	
Dataset 1: Air Quality & Respiratory Health Outcome: fev1 (Forced Expiratory Volume) What factors predict respiratory function in this population?	**The Residuals** **Shravya Sunkugari (TC)** Emily Y. Jin Soyu Hong Hailey Barrell	**No Outliers Here** **Erin K. Finn (TC)** Lauren E. Lee Ruth M. Moreira Ulloa Noah L. Gomes
Dataset 2: Hospital Readmission Risk Outcome: readmission_cost What patient and hospital factors predict 30-day readmission cost?	**The Skew Slayers** **Barron Clancy (TC)** Madilyn H. Matsunaga Laura Wu Audrey Sieng	**Beta Crew** **Alyssa R. Sherry (TC)** Matthew T. Liu Bianca L. Farro AJ Wu
Dataset 3: Workplace Wellness Program Outcome: stress_score_change (Post - Pre) How effective is a new wellness program on employee stress?	**Log-ical Thinkers** **Katherine Dunham (TC)** Joshua Dantus Anh Vu Preston W. Rossi Cailyn E. Clemons	**The Leverage Points** **Melissa R. Ponce (TC)** Grace H. Minano Lopez Sara M. Brinton Julci L. Areza Sophia L. Yang

Dataset 4: Food Security & Child Nutrition Outcome: haz What household and child-level factors predict child nutritional status (haz)?	**The Role Models** **Julia E. Shrier (TC)** Huyen N. Nguyen Eurie L. Seo Sasha Gordon Phoebe Koehler	**The Regressionals** **Ruviha A. Homma (TC)** Ishan D. Shah Jamiley Y. Avila Shuyue Xu Kenneth Kalu