APSTA-2122

Regression Project, due Friday, April 21, 5:00 PM
Submit via NYU Classes

Introduction

The Berkeley Guidance Study, under the direction of Jean Macfarlane, started with a sample of infants who were born in Berkeley, California in 1928-1929. Most of the children were Caucasian and Protestant, and two-thirds came from middle-class families. The basic cohort includes 136 of these children who participated in the study through the 1930s and up to the end of World War II. Annual data collection ended in 1946. In this project, you are asked to prepare a short data analysis using these data. The dataset contains a short list of variables pertaining to the child at three time points: age 2, age 9 and age 18. The variables collected in this study include: Sex, Height (cm) and Weight (kg) at ages 2, 9 and 18, leg circumference (cm) and strength (kg) at ages 9 and 18, and Somatotype (a 1 to 7 scale of body type).

Grading Criteria

Instructions for each section are described below. This assignment will be graded both on correctness of answers provided, as well as the presentation of results and the accompanying code. All code should be clearly organized and well-commented. If you consult outside resources, please cite them.

Part 1: Exploring Various Models (40 points)

This section may be written like a homework assignment - written either in a text editor (e.g. Word), or as a R markdown file. That is, explain your conclusions to a statistical audience and include all relevant code as an accompanying .R file or embedded within an R markdown file. **If you only submit a .R file, we will not grade it.**

1. Model height growth from age 2 to age 9 by answering the following questions:

   (a) Create a scatter plot of heights at age 9 on heights at age 2, with the points colored based on the gender of the child. Does there appear to be a different pattern for boys than for girls?

   (b) Fit a simple linear regression of heights at age 9 on heights at age 2.

      - Report and interpret the estimated regression coefficients.
      - Test the hypothesis of $H_0 : \beta_1 = 0$ against the two-sided alternative.

- Show numerically that the value of the T-statistic for the above hypothesis test is equal to the square root of the F-statistic from the ANOVA at the bottom of the regression output.
- Check the normality and homoscedasticity assumptions on the residuals. Include any plots you consult.

(c) Considering a model that allows for separate intercepts for boys and girls, is this model significantly better than the simple linear regression fit above?

(d) Considering a model that allows for both the separate slope and separate intercepts for boys and for girls, is this model significantly better than the simple linear regression fit above?

2. Model height growth from age 9 to age 18 by answering the following questions:

   (a) Create a scatter plot of heights at age 18 on heights at age 9, with the points colored based on the gender of the child. Does there appear to be a different pattern for boys than for girls?

   (b) Fit a simple linear regression of heights at age 18 on heights at age 9. Report the estimated regression coefficients.

   (c) Considering a model that allows for separate intercepts for boys and girls, is this model significantly better than the simple linear regression fit above?

   (d) Considering a model that allows for both the separate slope and separate intercepts for boys and for girls, is this model significantly better than the simple linear regression fit above?

   (e) Choose which of the above 3 models you think best describes the data and interpret the parameter estimates for this model.

3. Create a new dataset that includes only the boys in the sample. Use this new dataset to model the change in weight from age 9 to age 18.

   (a) Fit two linear regression models: (M1) Weight at age 18 on weight at age 9 and (M2) Weight at age 18 on weight at age 9 and leg circumference at age 9. Explain why weight at age 9 is significant in one model but not the other. Justify your answer by calculating the appropriate correlation coefficient.

   (b) The *hat matrix* can be calculated as $H = X(X^TX)^{-1}X^T$, where X is the design matrix. The diagonal values of the hat matrix determine the leverage that each point has in the fit of the regression model.

      - Explain why this matrix is known as the hat matrix. (You may need to do some research to answer this question).
      - Calculate this matrix in `R`. Show that the leverage of one of the points is much higher than any of the other points.

- Fit two simple linear regression models, both regressing weight at age 18 on weight at age 9. One model should use all of the boys in the dataset, and the other should remove the high-leverage point. Compare the coefficients for weight at age 9 obtained from both models.
- Create a scatter plot of weight at age 18 on weight at age 9. Plot both regression lines fit in the previous part on the plot in different colors.
- Based on the above parts, which regression line you think better fits the data? Report and interpret the estimated regression parameters for the model you choose.

4. Create a new dataset that includes only the girls in the sample. Use this new dataset to model Somatotype in the following ways.

   (a) Plot somatotype against weight at each of the three time points. Comment on how the relationship between weight and somatotype changes over time.

   (b) Create new variables:

$$\begin{aligned} DW9 &= WT9 - WT2 \\ DW18 &= WT18 - WT9 \\ AVE &= \frac{1}{3}(WT2 + WT9 + WT18) \\ LIN &= WT18 - WT2 \\ QUAD &= WT2 - 2 \cdot WT9 + WT18 \end{aligned}$$

   DW9 and DW18 measure the change in weight between consecutive timepoints. AVE, LIN, and QUAD measure the average, linear and quadratic trends over time (since the timepoints are roughly evenly spaced).

   (c) Fit the following three models:

$$M1 : Somatotype \sim WT2 + WT9 + WT18$$

$$M2 : Somatotype \sim WT2 + DW9 + DW18$$

$$M3 : Somatotype \sim AVE + LIN + QUAD$$

   Compare and contrast these models by answering the following questions:

   - What attributes of the models are the same across all three models? What attributes of the models are different?
   - Why does the coefficient for DW18 in model 2 equal the coefficient for WT18 in model 1, but the coefficient for DW9 in model 2 does not equal the coefficient for WT9 in model 1?
   - Show algebraically why M1 and M3 are equivalent by showing how the coefficients in M3 can be obtained by algebraically manipulating the coefficients in M1.

(d) Fit the following model:

$$M4 : Somatotype \sim WT2 + WT9 + WT18 + DW9$$

Explain why some parameters are not estimated.

Part 2: Explaining Somatotype (60 points)

This section should read as stand-alone descriptive scientific report (that is, it should read independently of the previous part). You should write this report in a text editor, such as Microsoft Word or LaTeX. All figures and tables must be properly formatted - that is, do not copy and paste your tables from `R`. This section must not exceed **15 pages** double-spaced including any necessary figures and tables. Please use appropriate section headers to increase the readability of your report. Include your code in a well-commented .R file. Your report should be formatted as follows:

Introduction Section ($\approx 1 - 2$ pages)

1. Start with an introduction to the problem. Why might people be interested in understanding growth data?
2. Introduce the study behind this dataset. You may choose to do a quick internet search to learn more about this study. Make sure to include a summary of the study design, study population, and the variables collected.
3. Introduce the idea of somatotype. What is somatotype? What does a large value or a small value represent? Are there any other measures that could have been used instead?
4. Pose research questions that your analysis will answer.

Descriptive Section ($\approx 3 - 4$ pages)

1. Describe the patterns in this dataset by providing numerical and graphical summaries of the variables of interest. This section should include:
   - Table(s) containing descriptive statistics of the people in the sample. For continuous variables, present the following five measures: sample mean, sample standard deviation, sample median, min and max. For categorical variables, present the number of individuals in each category as well as the corresponding sample proportion. You may choose to create separate descriptive statistic tables for each gender separately, as well as for all of the children together.
   - Figures pertaining to exploratory data analysis that shed light about some of the relationships in the data. Only include plots that are relevant to the questions of interest. Some examples of graphs are

    - Histogram of the somatotype variable.
    - Scatterplots between height and weight variables at different time points.
    - Boxplots of the continuous variables against sex.
    - Graphs showing the trend in weight gain or height increase over time.

- All plots need appropriate captions and should be explained in the text so that a general scientific audience may understand what is being presented (i.e. you may assume that you reader knows what a mean or median is and how to read a box plot, but not how to interpret patterns seen over different box plots or histograms). We will grade you on presentation of your figures as well as what is in the text, so make sure you spend time on making them look nice. You may want to consider using the `ggplot2` package in `R`.

Explanatory Model (≈ 4 pages)

1. Build the best explanatory model for the factors that may be associated with somatotype. Consider somatotype as a continuous response variable and any other variable (including those you derived above) as the predictor variables.

    (a) You must use at least two model selection methods discussed in class to choose an optimal explanatory model among candidate models that include additive relationships among any of the variables, as well as any two-way interaction term with gender. . These may include (but are not limited to) F statistics, R^2, adjusted R^2, AIC and BIC stepwise methods, penalized regression methods, cross-validation.

    (b) Do your two model selection methods agree? If not, which do you believe is the better explanatory model. Justify your conclusion with empirical evidence. In addition, you should also use the substantive knowledge you have, and any results from Part (1), to aid you with this decision. Articulate your model selection strategies clearly and briefly explain why you are making any decisions.

    (c) Include a table summarizing your final model, that includes the estimates of the regression coefficients and their standard errors, the relevant test statistics and p-values (or 95% confidence intervals of the estimates). Interpret the coefficients in the table in the text of the report. These interpretations should be correct, and understandable by scientists interested in understanding Somatotype (not just statisticians).

    (d) Conduct residual diagnostics for your choice of model and report your findings. Are the assumptions of multiple linear regression satisfied for your final model.

Conclusion Section ($\approx 1 - 2$ pages)

1. Recap the problem of interest and the research questions you pose.

2. How do your analyses answer the research questions? Make sure to recap the methods used and the results you found.
3. What are the limitations to your study? Make sure to address limitations in the study design, as well as the choice of analysis. If you could collect more (or different) data, would you? If so, what would you collect?
4. End with a conclusion paragraph that summarizes your report.