Homework 5

BSTA 512/612

2024-03-07

> **!** Important
>
> **THIS PAGE IS UNDER CONSTRUCTION!!**

Directions

- Please upload your homework to Sakai. **Upload both your .Rmd code file and the knitted .html file**.
- For each question, make sure to include all code and resulting output in the html file to support your answers.
- Show the work of your calculations using R code within a code chunk. Make sure that both your code and output are visible in the knitted html file.
- Write all answers in complete sentences as if communicating the results to a collaborator.
  - Points (usually 0.5-1) will be deducted for not including a sentence summarizing results in the context of the research study.
  - Questions not requiring a sentence are: *none - include a summary for all questions*

  Tip: It is a good idea to try kitting your document from time to time as you go along! Note that knitting automatically saves your Rmd file and knitting frequently helps you catch your errors more quickly.

HW 5 specific directions

- For all hypothesis tests, include
  - the null and alternative hypotheses (and if applicable the regression model(s) being tested),
  - R code to run the test, and
  - a conclusion in the context of the problem.
- You do not need to run the hypothesis tests using the formula unless directed otherwise.

Question 1: Association model building

Use the data from Chapter 16 Problem 5 to answer the questions below. These are NOT questions from the book.

a)

Create a figure with pairwise scatter plots of the variables. Note that depression index (DEP) is the outcome measure (or response variable, Y).

- **Do you see any signs of the linearity assumption not being met? If so, for which variables?**
- **Do you see any strong correlations between independent variables that could potentially cause collinearity problems? Confirm your observations by calculating pairwise correlations among the predictors.**

b)

Check if the DEP data are normally distributed. If needed, what would be an appropriate transformation for this variable?

c)

Test whether the association between DEP and WP is significant using alpha = .05.

d)

Suppose researchers would like to use a log transformation for the MC variable, based on what has been done in other studies. Do you agree with this choice? Why or why not? Whether or not you agree, test whether the association between DEP and log(MC) is significant, using alpha = .05, and use log(MC) instead of MC for the remainder of the assignment.

e)

Test whether SEX is an effect modifier that changes the association between DEP and WP. Use alpha = .10.

f)

Test whether SEX is an effect modifier that changes the association between DEP and log(MC). Use alpha = .10.

g)

From the results obtained in parts (e) and (f), should we further check whether SEX is a confounder? Why or why not? If yes, determine whether SEX is a confounder for the associations between DEP and WP and between DEP and log(MC)

h)

Determine whether AGE is a confounder for the associations between DEP and WP and between DEP and log(MC).

i)

What is your final association model based on the results from the previous questions?

j)

Perform model diagnostics for your final association model. Use the steps outlined below.

j1)

Determine whether the independence assumption has been met.

j2)

Determine whether the linearity assumption has been met.

j3)

Determine whether the homoscedasticity assumption has been met.

j4)

Determine whether the normality assumption has been met.

j5)

Determine whether there any outliers (e.g., high leverage, lack of fit) or influential points (e.g., dffits, Cook's distance, dfbetas)?

j6)

Is there evidence of collinearity in the model? If there is collinearity, make changes to your model to reduce the collinearity and justify your method.

k)

Using the final association model obtained from part (j5), interpret the (adjusted) association between DEP and MC.

l)

What is the R-squared value for your final association model? Explain it in the context of study.

m)

For your final association model, run a hypothesis test and report the 95% CI for the slope of the WP variable. Interpret the results (both test and CI).

Question 2: Prediction model building

Continue to use the data from Chapter 16 Problem 5 to answer the questions below. These are NOT questions from the book.

Suppose that in addition to the independent variables, the interactions `log(MC)*SEX`, `WP*SEX`, `WP*MC_log`, `WP*AGE`, `MC_log*AGE`, and `SEX*AGE` are of interest in building a prediction model.

a)

Run the four automatic selection procedures discussed in class (forward selection, backward elimination, forward stepwise, and backward stepwise) using all the independent variables and the interactions listed above. What are the parsimonious models from the different procedures? How are they similar or different? If the results are different, provide some reasons as to why they are different.

Forward Selection

Backwards Selection

Forward Stepwise

Backwards Stepwise

Comparison of models

b)

Restricting to just the variables that appeared in any of the prediction models obtained in the previous part, use Mallow's C_p and the models' adjusted R-squared values to decide on a parsimonious final prediction model. Include an explanation on how you chose your final prediction model.