

Problem Set 4

BUAN 6356

Due: Tuesday, 2017-11-21-11:59pm

Deliverable:

an R source-code file named ps4.r

Question 1

Data

The htv.csv data set includes information on wages, education, parents' education, and several other variables for 1,230 working men in 1991.

Analysis

- Read the data htv.csv into a new variable: context1
- Run the following linear model using the 'lm' function. Store the result in: model1

$$\ln[\text{wage}_i] = \beta_0 + \beta_1 \text{abil}_i + \beta_2 \text{educ}_i + \beta_3 \text{exper}_i + e_i \quad (1)$$

- Find the AIC and BIC of model1.
- Run the following quadratic model using the 'lm' function. Store the result in: model2

$$\begin{aligned} \ln[\text{wage}_i] = & \beta_0 + \beta_1 \text{abil}_i + \beta_2 \text{educ}_i + \beta_3 \text{exper}_i + \beta_4 \text{abil}_i^2 + \beta_5 \text{educ}_i^2 + \beta_6 \text{exper}_i^2 + \\ & \beta_7 (\text{abil}_i \times \text{educ}_i) + \beta_8 (\text{abil}_i \times \text{exper}_i) + \beta_9 (\text{educ}_i \times \text{exper}_i) + e_i \end{aligned} \quad (2)$$

Store the result in model2.

- Find the AIC and BIC of model2. You should have found that model2 has a higher BIC than model1, making it less optimal.
- Remove as many variables as you can from model2 to find the key subset of variables that makes model2 have the lowest BIC that is possible. Make sure this minimum BIC model is your final (turn-in) version of model2.

Interpretations

- Once model2 has been reduced to the lowest BIC version, what is the difference between models 1 and 2?
- What is the variable $(\text{educ}_i \times \text{exper}_i)$ doing to the model? We call this variable an interaction variable.

Question 2

Data

The `loanapp.csv` data contains information on individuals and whether or not they were approved for a loan to buy a house. The binary variable to be explained is approve_i , which is equal to one if a mortgage loan to an individual was approved. The key explanatory variable is white_i , a dummy variable equal to one if the applicant was white. The other applicants in the data set are black and Hispanic.

Analysis

- Read the data `loanapp.csv` into a new variable: `context2`
- Run the following logistic model using the ‘`glm`’ function. Store the result in: `model3`

$$g(\text{approve}) = \beta_0 + \beta_1 \text{white} \quad (3)$$

- Compute the White heteroskedasticity robust standard errors for `model3`.
- Run the following logistic model using the ‘`glm`’ function. Store the result in: `model4`

$$\begin{aligned} g(\text{approve}) = & \beta_0 + \beta_1 \text{white} + \beta_2 \text{hrrat} + \beta_3 \text{obrat} + \beta_4 \text{loanprc} + \beta_5 \text{unem} + \beta_6 \text{male} \\ & + \beta_7 \text{married} + \beta_8 \text{dep} + \beta_9 \text{sch} + \beta_{10} \text{cosign} + \beta_{11} \text{chist} + \beta_{12} \text{pubrec} \\ & + \beta_{13} \text{mortlat1} + \beta_{14} \text{mortlat2} + \beta_{15} \text{vr} + e_i \end{aligned} \quad (4)$$

- Compute the White heteroskedasticity robust standard errors for `model4`.
- Run the following logistic model using the ‘`glm`’ function. Store the result in: `model5`

$$\begin{aligned} g(\text{approve}) = & \beta_0 + \beta_1 \text{white} + \beta_2 \text{hrrat} + \beta_3 \text{obrat} + \beta_4 \text{loanprc} + \beta_5 \text{unem} + \beta_6 \text{male} \\ & + \beta_7 \text{married} + \beta_8 \text{dep} + \beta_9 \text{sch} + \beta_{10} \text{cosign} + \beta_{11} \text{chist} + \beta_{12} \text{pubrec} \\ & + \beta_{13} \text{mortlat1} + \beta_{14} \text{mortlat2} + \beta_{15} \text{vr} + \beta_{16} (\text{white} \times \text{obrat}) + e_i \end{aligned} \quad (5)$$

- Compute the White heteroskedasticity robust standard errors for `model5`

Interpretations

- In `model3`, we can’t directly interpret the coefficient on `white` like we did in the linear model because the logit link function $g(\cdot)$ confuses things a bit. That said, the sign and significance of β_1 still have the same effect. What is this coefficient indicating roughly?
- After adding 14 more variables in `model4`, how does β_1 change? Is it still significant?
- After adding the interaction between `white` and `obrat` in `model5`, how has β_1 changed now?
- What is $(\text{white} \times \text{obrat})$, and why do you think it has affected the model so greatly?

Question 3

Data

We use the data in `smoke.csv` to estimate a demand function for daily cigarette consumption. Since most people do not smoke, the dependent variable, `cigs`, is zero for most observations. A linear model is not ideal because it can result in negative predicted values.

Analysis

- Read the data `smoke.csv` into a new variable: `context3`
- Run the following Poisson model using the ‘`glm`’ function. Store the result in: `model6`

$$g(\text{cigs}) = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{age} + \beta_3 \text{age}^2 + \beta_4 \log[\text{income}] + \beta_5 \text{restaurn} \quad (6)$$

- Compute the White heteroskedasticity robust standard errors for `model6`.

Interpretations

- Interpret the coefficient on `educ` in `model6` (hint: the link function for Poisson is $g(y) = \ln[y]$)?
- Find the rate of change for $\ln[\text{cigs}]$ with respect to `age`. Find this number when a person is 20 and when a person is 60.

Question 4

Data

The data in `hdisease.csv` contains patient information for patients admitted to four hospitals for different studies. The patients were admitted for chest pain and were diagnosed with or without heart disease. When they were diagnosed with heart disease, the disease was ranked in stages 1 to 4.

Analysis

- Read the data `hdisease.csv` into a new variable: `context4`
- Run the following recursive partitioning model using the `'evtree'` function from the `'evtree'` package. Store the result in: `model7`

$$E[\text{hdisease}] = f(\text{age}, \text{cp}, \text{trestbps}, \text{thalach}, \text{exang}) \quad (7)$$

- Run the same recursive partitioning model using the `'ctree'` function from the `'party'` package. Store the result in `model8`.
- Read the data `hdisease-new.csv` into a new variable: `context5`
- Using `model8`, find the predicted classifications from for the `context5` data using the `'predict'` function. Store these as a new variable `'hdisease_pred'` inside of `context5`.

Interpretations

- a. Comparing `model7` to `model8`, which model might be over-fitting the data? Which model might be under-fitting the data?
- b. Why don't we include `dset` in these models?