Problem Set 4 Solutions

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

$$\ln\left[\text{wage}_i\right] = \beta_0 + \beta_1 \text{abil}_i + \beta_2 \text{educ}_i + \beta_3 \text{exper}_i + e_i \tag{1}$$

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

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

  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

a. Once model2 has been reduced to the lowest BIC version, what is the difference between models 1 and 2?

$$\begin{aligned}\ln\left[\text{wage}_i\right] &= \beta_0 + \beta_1 \text{abil}_i + \beta_2 \text{educ}_i + \beta_3 \text{exper}_i + \beta_9\left(\text{educ}_i \times \text{exper}_i\right) + e_i \\ \ln\left[\text{wage}_i\right] &= \beta_0 + \beta_2 \text{educ}_i + \beta_7\left(\text{abil}_i \times \text{educ}_i\right) + \beta_9\left(\text{educ}_i \times \text{exper}_i\right) + e_i\end{aligned}$$

b. What is the variable $(\text{educ}_i \times \text{exper}_i)$ doing to the model? We call this variable an interaction variable.

$$\begin{aligned}\frac{d\ln\left[\text{wage}\right]}{d\text{educ}} &= \beta_2 + \beta_9 \text{exper} = 0.049 + 0.006\left(\text{exper}\right) \\ \frac{d\ln\left[\text{wage}\right]}{d\text{educ}} &= \beta_2 + \beta_7 \text{abil} + \beta_9 \text{exper} = 0.069 + 0.004\left(\text{abil}\right) + 0.003\left(\text{exper}\right)\end{aligned}$$

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

$$g\left(\text{approve}\right) = \beta_0 + \beta_1\text{white} \tag{3}$$

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

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

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

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

- Compute the White heteroskedasticity robust standard errors for model5

Interpretations

a. In model3, we can't directly interpret the coefficient on white like we did in the linear model because the logit link function $g\left(\cdot\right)$ confuses things a bit. That said, the sign and significance of β_1 still have the same effect. What is this coefficient indicating roughly?

Without controlling for other factors, it seems that being white increases the approval rate for loans.

b. After adding 14 more variables in model4, how does β_1 change? Is it still significant?

After controlling for many other factors, there is still a significant increase in the approval rate for white customers.

c. After adding the interaction between white and obrat in model5, how has β_1 changed now?

β_1 has become completely insignificant.

d. What is $\left(\text{white} \times \text{obrat}\right)$, and why do you think it has affected the model so greatly?

$$\begin{aligned} E\left[g\left(\text{approve}\right)|\text{white} = 0\right] &= \beta_0 + \beta_2\text{hrat} + \beta_3\text{obrat} + \cdots + \beta_{15}\text{vr} \\ E\left[g\left(\text{approve}\right)|\text{white} = 1\right] &= \left(\beta_0 + \beta_1\right) + \beta_2\text{hrat} + \left(\beta_3 + \beta_{16}\right)\text{obrat} + \cdots + \beta_{15}\text{vr} \end{aligned}$$

So in this model, there is no statistical level-difference between approval for people who white or non-white. But there is a significant difference between how white and non-white people are approved based on their other obligations.

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

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

- Compute the White heteroskedasticity robust standard errors for model6.

Interpretations

a. Interpret the coefficient on educ in model6 (hint: the link function for Poisson is $g\left(y\right) = \ln\left[y\right]$)?

Every 1 year increase in eductation is associated with a 5.95% decrease in the number of cigarettes smoked.

b. Find the rate of change for $\ln\left[\text{cigs}\right]$ with respect to age. Find this number when a person is 20 and when a person is 60.

$$\frac{d\ln\left[\text{cigs}\right]}{d\text{age}} = \beta_2 + 2\beta_3\text{age} = 0.1139858 - 2\left(0.0013679\right)\text{age}$$

$$\begin{aligned} \left.\frac{d\ln\left[\text{cigs}\right]}{d\text{age}}\right|_{\text{age}=20} &= \frac{5.927\%}{\text{year}} \\ \left.\frac{d\ln\left[\text{cigs}\right]}{d\text{age}}\right|_{\text{age}=60} &= \frac{-5.0162\%}{\text{year}} \end{aligned}$$

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

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

- 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: context45
- Using model8, find the predicted classifications from for the context45 data using the 'predict' function. Store these as a new variable 'hdisease_pred' inside of context45.

Interpretations

a. Comparing model7 to model8, which model might be over-fitting the data? Which model might be under-fitting the data?

model7 under-fits. model8 over-fits.

b. Why don't we include dset in these models?

Because we don't have identifying information on dset for context5.

Question 5

Data

The WAGE1.csv data is the data from your first problem set.

Analysis

- Read the data WAGE1.csv into a new variable: context5
- Consider using $k-$means to segment the WAGE1.csv data. Plot the within-group sum of squares for $k = 1, 2, \ldots, 10$.
- Set the seed to be 2 as we did in class and use $k-$means with 10 initial starting positions to estimate $k = 3$ means (which may or may not be the correct number) using context5. Store the $k-$means result in: model9 (Hint: be careful about your seed, or you might get a different cluster order than the grading system does.)
- Find the estimated means from model9.
- Using model9, segment the data into three groups and run the following linear model for clusters 1, 2, and 3. Store the results in model10, model11, and model12 respectively.

$$\text{wage}_i = \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{exper}_i + \beta_3 \text{tenure}_i + e_i$$

Interpretations

a. Using the elbow test on the within- sum of squares plot, find the optimal number of clusters for this data set.

It looks like 2 is the right number to me, but I would accept anything from 2 to 4.

b. Looking at the means from model1, describe the different clusters. [Hint: Look at the education, experience, and tenure variables in particular.]

Cluster 1 : Older blue collar workers (low educ, high exp/ten)
Cluster 2 : Younger white collar workers (high educ, low exp/ten)
Cluster 3 : Older white collar workers (high educ, high exp/ten)

c. Discuss the differences between models 2, 3, and 4.

	Cluster 1	Cluster 2	Cluster 3
(Intercept)	3.78799	-3.21053	-4.8574
educ	0.40735	0.52524	0.73631
exper	-0.10172	0.17286	0.05905
tenure	0.13653	0.29156	0.21796

For blue collar workers, education, experience, and tenure have relatively low value (everyone earns about the same). For young white collar workers, experience and tenure are immensely valuable whereas education has a relatively low value. For older white collar workers, education is immensely valuable.

Question 6

Data

The data in murder.csv contains murder rates (murders per 100,000 persons) for the 50 US states plus D.C. from 1965 to 2014

Analysis

- Read the data murder.csv into a new variable: context6
- Run principal components analysis on the 51 series and store the result in: model13. [Ignore stationarity for this question, although in general, you should make data stationary before operating on it]
- Generate the scree plot for model13.
- Store the first principal component inside of context6 as: factor
- Show a time series plot of the factor

Interpretations

a. Based on the scree plot, how many principal components should we use for this data?

1 principal component looks reasonable to me.

b. Describe the overall movement of the factor over time. When is the factor greatest?

The factor is greatest in 1990-1992. Crime was rising up to that point. It has been falling ever since.