Business 41903

Instructor: Christian Hansen

Problem Set 4

1. Use the data in TFAdata.txt to answer this question. The data consist of 9915 observations from the 1990 SIPP. Before answering the question, remove 1915 observations (at random) which will be used for an out-of-sample comparison in part e and part j. The variables in the data are

- *net_tfa* = net total financial assets (in dollars)
- *age* = age of household head (in years)
- *inc* = household income (in dollars)
- *fsize* = family size
- *educ* = years of completed education of household head
- *db* = indicator for having a defined benefit pension
- *marr* = indicator for being married
- *male* = indicator for male household head
- *twoearn* = indicator for being a two-earner household
- *e401* = indicator for 401(k) eligibility
- *p401* = indicator for participating in a 401(k)
- *pira* = indicator for participating in an IRA
- *hown* = indicator for owning a home

a. Estimate $\mathrm{E}[net\_tfa|X = \{age, inc, fsize, educ, marr, twoearn\}]$ using a kernel estimator with product kernel and bandwidth(s) chosen by cross-validation. Carefully explain how you handle coarsely discrete variables (i.e. *marr* and *twoearn*). What are the bandwidths? How do they compare to the range of the individual conditioning variables?

b. Estimate $\mathrm{E}[net\_tfa|X = \{age, inc, fsize, educ, marr, twoearn\}]$ using k-nn with the number of neighbors chosen by cross-validation. Carefully explain how you handle coarsely discrete variables (i.e. *marr* and *twoearn*). Comment on how many neighbors you chose and how this relates to the sample size.

c. Estimate E[$net\_tfa|X = \{age, inc, fsize, educ, marr, twoearn\}$] using series with basis elements chosen by cross-validation. Carefully explain how you handle coarsely discrete variables (i.e. *marr* and *twoearn*). Carefully document the basis you are using and how you chose to add elements to the expansion along the path considered for cross-validation. Comment on which terms you end up selecting.

d. Estimate E[$net\_tfa|X = \{age, inc, fsize, educ, marr, twoearn\}$] using a linear model via OLS.

e. Use the 1915 observations you held out to compare the estimates obtained in parts a.-d. Specifically, let $\widehat{g}_j(x)$ for $j \in \{a, b, c, d\}$ be the estimator of the conditional expectation obtained in part a.-d. respectively. Calculate the mean square forecast error as $\frac{1}{1915}\sum_{i \in hold-out}(\widehat{g}_j(x_i) - y_i)^2$ and the mean absolute forecast error as $\frac{1}{1915}\sum_{i \in hold-out}|\widehat{g}_j(x_i) - y_i|$. Which procedure performs best according to each metric? Do the performance discrepancies seem large? [Note: Assuming independent sampling, you can compute a standard error for the mean square forecast error and for the mean absolute forecast error conditioning on the estimated model.]

f. Estimate E[$e401|X = \{age, inc, fsize, educ, marr, twoearn\}$] using a kernel estimator with product kernel and bandwidth(s) chosen by cross-validation. Carefully explain how you handle coarsely discrete variables (i.e. *marr* and *twoearn*). What are the bandwidths? How do they compare to the range of the individual conditioning variables?

g. Estimate E[$e401|X = \{age, inc, fsize, educ, marr, twoearn\}$] using k-nn with the number of neighbors chosen by cross-validation. Carefully explain how you handle coarsely discrete variables (i.e. *marr* and *twoearn*). Comment on how many neighbors you chose and how this relates to the sample size.

h. Estimate E[$e401|X = \{age, inc, fsize, educ, marr, twoearn\}$] using series with basis elements chosen by cross-validation. Carefully explain how you handle coarsely discrete variables (i.e. *marr* and *twoearn*). Carefully document the basis you are using and how you chose to add elements to the expansion along the path considered for cross-validation. Comment on which terms you end up selecting. How many predicted values are smaller than 0 or larger than 1?

i. Estimate E$[net_tfa|X = \{age, inc, fsize, educ, marr, twoearn\}]$ using a logit model with a linear index in $\{age, inc, fsize, educ, marr, twoearn\}$.

j. Use the 1915 observations you held out to compare the estimates obtained in parts f.-i. Specifically, let $\widehat{g}_j(x)$ for $j \in \{f, g, h, i\}$ be the estimator of the conditional expectation obtained in part f.-i. respectively. Calculate the mean square forecast error as $\frac{1}{1915}\sum_{i \in hold-out}(\widehat{g}_j(x_i) - y_i)^2$ and the misclassification rate $\frac{1}{1915}\sum_{i \in hold-out}\mathbf{1}(\widehat{y}_{j,i} \neq y_i)$ where $\widehat{y}_{j,i}$ is the Bayes-classifier based on model j - i.e. $\widehat{y}_{j,i} = \mathbf{1}(\widehat{g}_j(x_i) \geq .5)$. Which procedure performs best according to each metric? Do the performance discrepancies seem large? [Note: Assuming independent sampling, you can compute a standard error for the mean square forecast error and for the misclassification rate conditioning on the estimated model.]