The Analytics Edge

Test your knowledge of Discrete Choice and Model Selection in R

1. This problem set uses data on the choice of the heating system in California houses. The dataset in the file **Heating.csv** consists of observations for 900 single-family houses in California that were newly built and had central air-conditioning. The choice is among heating systems. Five types of systems are considered to have been possible:

   - gas central (gc)
   - gas room (gr)
   - electric central (ec)
   - electric room (er)
   - heat pump (hp)

   There are 900 observations where the variables are:

   - **idcase**: observation number (1-900)
   - **depvar**: identifies the chosen alternative (gc, gr, ec, er, hp)
   - **ic.alt**: installation cost for the 5 alternatives (alt = gc, gr, ec, er, hp)
   - **oc.alt**: annual operating cost for the 5 alternatives (alt = gc, gr, ec, er, hp)
   - **income**: annual income of the household (in tens of thousands of dollars)
   - **agehed**: age of the household head
   - **rooms**: number of rooms in the house
   - **region**: a factor with levels ncostl (northern coastal region), scostl (southern coastal region), mountn (mountain region), valley (central valley region)

   Note that the attributes of the alternatives, namely, installation cost and operating cost, take a different value for each alternative. Therefore, there are 5 installation costs (one for each of the 5 systems) and 5 operating costs. To estimate the logit model, the researcher needs data on the attributes of all the alternatives, not just the attributes for the chosen alternative. For example, it is not sufficient for the researcher to determine how much was paid for the system that was actually installed (i.e., the bill for the installation). The researcher needs to determine how much it would have cost to install each of the systems if they had been installed. The importance of costs in the choice process (i.e., the coefficients of installation and operating costs) is determined through comparison of the costs of the chosen system with the costs of the non-chosen systems.

For these data, the costs were calculated as the amount the system would cost if it were installed in the house, given the characteristics of the house (such as size), the price of gas and electricity in the house location, and the weather conditions in the area (which determine the necessary capacity of the system and the amount it will be run.) These cost are conditional on the house having central air-conditioning. (That's why the installation cost of gas central is lower than that for gas room: the central system can use the air-conditioning ducts that have been installed.)

You'll see that the first household chose alternative 1 (gas central), has an income of $70,000, the head of household is 25 years old, the house has 6 rooms, and is located in the north coastal area.

(a) Run a logit model with installation cost and operating cost as the only explanatory variables, without intercepts.

   i. Do the estimated coefficients have the expected signs?

   ii. Are both coefficients significantly different from zero?

   iii. Use the average of the probabilities to compute the predicted share. Compute the actual shares of houses with each system. How closely do the predicted shares match the actual shares of houses with each heating system?

   iv. The ratio of coefficients usually provides economically meaningful information in discrete choice models. The willingness to pay (*wtp*) through higher installation cost for a one-dollar reduction in operating costs is the ratio of the operating cost coefficient to the installation cost coefficient. What is the estimated *wtp* from this model? Note that the annual operating cost recurs every year while the installation cost is a one-time payment. Does the result make sense?

(b) Add alternative-specific constants to the model in (a). With K alternatives, at most $K-1$ alternative specific constants can be estimated. The coefficient of $K-1$ constants are interpreted as relative to Kth alternative. Normalize the constant for the alternative hp to 0.

   i. How well do the estimated probabilities match the shares of customers choosing each alternative in this case?

   ii. Calculate the *wtp* that is implied by the estimate. Is this reasonable?

   iii. Suppose you had included constants for alternatives ec, er, gc, hp with the constant for alternative gr normalized to zero. What would be the estimated coefficient of the constant for alternative gc? Can you figure this out logically rather than actually estimating the model?

(c) Now try some models with sociodemographic variables entering.

   i. Enter installation cost divided by income, instead of installation cost. With this specification, the magnitude of the installation cost coefficient is inversely related to income, such that high income households are less concerned with installation costs than lower income households. Does dividing installation cost by income seem to make the model better or worse than the model in (b)?

ii. Instead of dividing installation cost by income, enter alternative-specific income effects. You can do this by using the | argument in the mlogit formula. What do the estimates imply about the impact of income on the choice of central systems versus room system? Do these income terms enter significantly?

2. A sample of residential electricity customers were asked a series of choice experiments. The data is provided in the file **Electricity.csv**. In each experiment, four hypothetical electricity suppliers were described. The person was asked which of the four suppliers he/she would choose. As many as 12 experiments were presented to each person. Some people stopped before answering all 12. There are 361 people in the sample, and a total of 4308 experiments. In the experiments, the characteristics of each supplier were stated.

   - The price of the supplier was either one of these options:
     - a fixed price at a stated cents per kWh, with the price varying over suppliers and experiments (**pf1, pf2, pf3, pf4**)
     - a time-of-day (tod) rate under which the price is 11 cents per kWh from 8am to 8pm and 5 cents per kWh from 8pm to 8am. These tod prices did not vary over suppliers or experiments: whenever the supplier was said to offer tod, the prices were stated as above (**tod1, tod2, tod3, tod4**)
     - a seasonal rate under which the price is 10 cents per kWh in the summer, 8 cents per kWh in the winter, and 6 cents per kWh in the spring and fall. Like tod rates, these prices did not vary. Note that the price is for the electricity only, not transmission and distribution, which is supplied by the local regulated utility (**seas1, seas2, seas3, seas4**).
   - The length of contract that the supplier offered was also stated, in years (such as 1 year or 5 years.) During this contract period, the supplier guaranteed the prices and the buyer would have to pay a penalty if he/she switched to another supplier. The supplier could offer no contract in which case either side could stop the agreement at any time. This is recorded as a contract length of 0 (**cl1, cl2, cl3, cl4)**.
   - Some suppliers were also described as being a local company or a "well-known" company. If the supplier was not local or well-known, then nothing was said about them in this regard (**loc1, loc2, loc3, loc4, wk1, wk2, wk3, wk4**).

   The actual choices made are captured in **choice** with **id** capturing the customer identity.

   (a) Run a mixed logit model without intercepts and a normal distribution for the 6 parameters of the model and taking into account the panel data structure.

      i. Using the estimated mean coefficients, determine the amount that a customer with average coefficients for price and length is willing to pay for an extra year of contract length.

ii. Determine the share of the population who are estimated to dislike long term contracts (i.e. have a negative coefficient for the length.)

(b) The price coefficient is assumed to be normally distributed in these runs. This assumption means that some people are assumed to have positive price coefficients, since the normal distribution has support on both sides of zero. Using your estimates from before, determine the share of customers with positive price coefficients (Hint: Use the pnorm function to calculate this share). As you can see, this is pretty small share and can probably be ignored. However, in some situations, a normal distribution for the price coefficient will give a fairly large share with the wrong sign. Revise the model to make the price coefficient fixed rather than random. A fixed price coefficient also makes it easier to calculate the distribution of willingness to pay (*wtp*) for each non-price attribute. If the price coefficient is fixed, the distribution of *wtp* for an attribute has the same distribution as the attribute's coefficient, simply scaled by the price coefficient. However, when the price coefficient is random, the distribution of *wtp* is the ratio of two distributions, which is harder to work with. What is the estimated value of the price coefficient/ Compare the log likelihood of the new model with the old model.

(c) You think that everyone must like using a known company rather than an unknown one, and yet the normal distribution implies that some people dislike using a known company. Revise the model to give the coefficient of wk a uniform distribution (do this with the price coefficient fixed). What is the estimated distribution for the coefficient of wk and the estimated price coefficient?

3. Suppose we perform best subset, forward stepwise, and backward stepwise selection on a single data set. For each approach, we obtain p+1 models, containing 0, 1, 2, ..., p predictors. Provide your answers for the following questions:

(a) Which of the three models with k predictors has the smallest training sum of squared errors?

(b) Which of the three models with k predictors has the smallest test sum of squared errors?

(c) True or False:

i. The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the (k + 1)-variable model identified by forward stepwise selection.

ii. The predictors in the k-variable model identified by backward stepwise are a subset of the predictors in the (k+ 1)- variable model identified by backward stepwise selection.

iii. The predictors in the k-variable model identified by backward stepwise are a subset of the predictors in the (k + 1)- variable model identified by forward stepwise selection.

iv. The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the (k + 1)-variable model identified by backward stepwise selection.

v. The predictors in the k-variable model identified by best subset are a subset of the predictors in the (k+1)-variable model identified by best subset selection

4. In this question, we will use the data in **College.csv** to investigate how well we can predict the number of applications received for universities and colleges in the US. The dataset has the following fields:

   - Private: Private/public indicator
   - Apps: Number of applications received
   - Accept: Number of applicants accepted
   - Enroll: Number of new students enrolled
   - Top10perc: New students from top 10% of high school class
   - Top25perc: New students from top 25% of high school class
   - F.Undergrad: Number of full-time undergraduate students
   - P.Undergrad: Number of part-time undergraduate students
   - Outstate: Out of state tuition
   - Room.Board: Room and board costs
   - Books: Estimated book costs
   - Personal: Estimated personal spending
   - PhD: Percent of faculty with PhDs
   - Terminal: Percent of faculty with a terminal degree
   - S.F.Ratio: Student/faculty ratio
   - perc.alumni: Percent of alumni who donate
   - Expend: Instructional expenditure per student
   - Grad.Rate: Graduation rate

   (a) Split the data set into a training set and a test set using the seed 1 and the sample() function with 80% in the training set and 20% in the test set. How many observations are there in the training and test sets?

   (b) Fit a linear model using least squares on the training set. What is the average sum of squared error of the model on the training set? Report on the average sum of squared error on the test set obtained from the model.

   (c) Use the backward stepwise selection method to select the variables for the regression model on the training set. Which is the first variable dropped from the set?

   (d) Plot the adjusted R^2 for all these models. If we choose the model based on the best adjusted R^2 value, which variables should be included in the model?

   (e) Use the model identified in part (d) to estimate the average sum of squared test error. Does this improve on the model in part (b) in the prediction accuracy?

(f) Fit a LASSO model on the training set. Use the command to define the grid for λ:
grid <- 10$\wedge$ seq(10,-2, length=100)
Plot the behavior of the coefficients as λ changes.

(g) Set the seed to 1 before running the cross-validation with LASSO to choose the best λ. Use 10-fold cross validation. Report the test error obtained, along with the number of non-zero coefficient estimates.