42.510 Statistical Learning for Data Science FALL 2025

(Practice) Assignment 2.5

No need to submit solutions. These are practice questions on discrete choice models, model selection and model assessment. Answers are provided in the .Rmd (and .html) files.

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 coeffiicients 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 will 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 coefficients 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) The present value of the future operating costs is the discounted sum of operating costs over the life of the system: $PV = \sum_{t=1}^{L}[OC/(1+r)^t]$ where r is the discount rate and L is the life of the system. As L rises, the PV approaches OC/r. Therefore, for a system with a sufficiently long life (which we will assume these systems have), a one-dollar reduction in OC reduces the present value of future operating costs by $(1/r)$. This means that if the person choosing the system were incurring the installation costs and the operating costs over the life of the system, and rationally traded-off the two at a discount rate of r, the decision-maker's *wtp* for operating cost reductions would be $(1/r)$. Define a new variable `lcc` (lifecycle cost) that is defined as the sum of the installation cost and the (operating cost)/r. Run a logit model with the lifecycle cost as the only explanatory variable. Estimate the model for $r = 0.12$. Comment on the value of log-likelihood of the models obtained in (a) as compared to (b).

(c) 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 coeffiient of the constant for alternative `gc`? Can you figure this out logically rather than actually estimating the model?

(d) 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 (c)?

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?

(e) We now are going to consider the use of the logit model for prediction. Estimate a model with installation costs, operating costs, and alternative specific constants. Calculate the probabilities for each house explicitly.

i. The California Energy Commission (CEC) is considering whether to offer rebates on heat pumps. The CEC wants to predict the effect of the rebates on the heating system choices of customers in California. The rebates will be set at 10% of the installation cost. Using the estimated coeffiients from the model, calculate predicted shares under this new installation cost instead of original value. How much do the rebates raise the share of houses with heat pumps?

ii. Suppose a new technology is developed that provides more efficient central heating. The new technology costs $200 more than the electric central heating system. However it saves 25% of the electricity such that its operating costs are 75% of the operating costs of `ec`. We want to predict the original market penetration of this technology. Note that there are now 6 alternatives instead of 5. Calculate the probability and predict the market share (average probability) for all 6 alternatives using the model that is estimated on the 5 alternatives. Use the original installation costs for the heat pumps rather than the reduced costs from the previous question. What is the predicted market share for the new technology? From which of the original five systems does the new technology draw the most customers?

2. 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 $0, 1, \ldots, p$ 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) Are the following statements **True** or **False**:

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

3. 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
   - `Top25perc`: New students from top 25
   - `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^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.