CPS2

1. Random Utility Model: Conditional Logit

(1) Import data from file `RevealedPreferenceData.mat`, which contains panel scanner data on purchases of margarine. This data set was originally analyzed in Allenby and Rossi (Quality Perceptions and Asymmetric Switching Between Brands, *Marketing Science*, 1991), and contains purchases on ten brands of margarine by 561 panelists, making a total of 4470 purchases. We also have information on various demographic characteristics of each household, including household income, family size, education, and if the head of the household is retired. The data has 44700 rows and 8 columns:
   - Column 1: Person = Individual identifier.
   - Column 2: Number of occasions.
   - Column 3: Choice = Choice indicator, 0/1; for each individual and each choice occasion, it is is a multinomial indicator of one of the 10 brands, namely PPk_{Stk},PBB_{St}, PFl_{Stk}, $PHse_{Stk}$, $PGen_{Stk}$, $PImp_{Stk}$, PSS_{Tub}, PPk_{Tub}, PFl_{Tub}, $PHse_{Tub}$, where *Pk* is Parkay; *BB* is BlueBonnett, *Fl* is Fleischmanns, *Hse* is house, *Gen* is generic, *Imp* is Imperial, *SS* is Shed Spread. *Stk* indicates stick, and *Tub* indicates Tub form.
   - Column 4: Income
   - Column 5: Family size
   - Column 6: College educated
   - Column 7: Retired
   - Column 8: Price in dollar

(2) Write the Log-Likelihood function and the Gradient. *Hint*: you have a sample of N individuals who make repeated choices. Assuming independence, the individual likelihood is a product of the densities for each choice; assuming i.i.d. samples, the overall likelihood to be maximized is the product of the individual ones. The formula for the individual likelihood and the gradient (score equations) are given in the lecture notes. To write the Log-Likelihood you need to define the utility of option j for individual i in occasion t as a function of the covariates and the parameters. The structure of the LL function should be as follows: as inputs you should have the parameters to be estimated, the set of regressors (covariates) and the set of individual choices; as outputs you should have the value of the LL and the value of the gradient.

The Log-likelihood function should look like something as `[LL,G] = ... loglik($x_0$,covariates,choices)`, so you have to set not only LL as output but also the gradient G as described in the slides. Refer to the Lecture Notes 6, example with Stated Preference data for inspiration...

(3) Consider two cases:
   (a) *CASE 1*: The utility of subject i for option $j = 1, \ldots, 10$ in choice task t is a function of price and a set of Alternative Specific Constants β_j

$$U_{ijt} = \alpha Price_{jt} + \beta_j + \epsilon_{ijt}$$

   where ϵ_{ijt} are i.i.t. Type 1 Extreme Value (Gumbel) distributed. (Hint: To create the ASC's, you might want to appropriately use the Matlab function `eye`).
   (b) *CASE 2*: The utility of subject i for option $j = 1, \ldots, 10$ in choice task t is a function of price which depends on te set of observed individual demographic characteristics (Income, FamilySize, CollegeEducated, Retired), and again on the set of Alternative Specific Constants β_j

$$U_{ijt} = \alpha_0 Price_{jt} + \alpha_I Price_{jt} Income_i + \alpha_F Price_{jt} FamilySize_i + ...$$
$$\alpha_C Price_{jt} CollegeEducated_i + \alpha_R Price_{jt} Retired_i + \beta_j + \epsilon_{ijt}$$

   where ϵ_{ijt} are again i.i.t. Type 1 Extreme Value (Gumbel) distributed.

(4) Maximize the LL function using the MATLAB command `fmincon`, and find the point estimates of the parameters, their standard errors and 95% confidence intervals in both cases (Hint: to find the standard errors, use the estimated Hessian, an output of the fmincon function). Report and comment your estimates.

(5) Using the estimated parameter calculate the averaged own-price and cross-price elasticities for the 10 goods (a 10×10 matrix) in the two cases, and comment on your results.

2. MSE

(1) Generate $n = 1000$ independent samples of size $m = 10$, where each observation $Y_{ij} \sim \mathcal{N}(\theta, \sigma^2)$, with known true mean $\theta = 1$ and standard deviation $\sigma = 1$. For each sample (i.e., each row of your simulated matrix), compute the sample mean estimator $\hat{\theta}_i$
(2) Implement two alternative estimators:
   - Shrinkage estimator: $\tilde{\theta}_i = \lambda \hat{\theta}_i$ with $\lambda = 0.8$
   - Constant estimator: $\theta_i^{\text{const}} = 0$ for all i
(3) For all three estimators (sample mean, shrinkage, and constant): Compute the bias, variance, and mean squared error (MSE)
(4) Present the results in a table with three rows and three columns: Bias^2, Variance, MSE
(5) Plot the empirical distribution (histogram or density) of the three estimators, labeling the estimators in the plot
(6) Which estimator has the smallest MSE? Repeat the analysis with $m = 5$ and $m = 50$. What do you observe about the bias-variance tradeoff as sample size increases?
(7) Be sure to comment your code and your results

3. Bayes

(1) Consider the following data from the Pfizer COVID-19 vaccine trial:
   - Vaccinated group: 8 out of 18,198 participants were infected
   - Placebo group: 162 out of 18,325 participants were infected

   You may treat these outcomes as binomial. The vaccine efficacy (VE) is defined as: $VE = 1 - \frac{\theta_v}{\theta_p}$
(2) Consider an uninformative prior: Beta$(1, 1)$ for the binomial success probability θ:
   - Compute the posterior distributions of θ_v and θ_p analytically and plot them
   - Compute the posterior means, variances, and 95% credible intervals for these two posterior distributions
   - Simulate from the posterior distributions of θ_v and θ_p, compute the derived distribution of VE, and plot its histogram.
(3) Redo the analysis with an informative prior: Beta$(0.5, 99.5)$
(4) Interpret the results:
   - What is the estimated vaccine efficacy, and what is its 95% credible interval?
   - How does the choice of prior affect the inference?
(5) Repeat the analysis assuming a smaller sample (10% of the original size)
(6) Be sure to comment your code and your results