Problem Set 8

ECON 6343: Econometrics III
Prof. Tyler Ransom
University of Oklahoma

Due: October 29, 9:00 AM

Directions: Answer all questions. Each student must turn in their own copy, but you may work in groups. You are encouraged to use any and all Artificial Intelligence resources available to you to complete this problem set. Clearly label all answers. Show all of your code. Turn in jl-file(s), output files and writeup via GitHub. Your writeup may simply consist of comments in jl-file(s). If applicable, put the names of all group members at the top of your writeup or jl-file.

You may need to install and load the following package:

```
MultivariateStats
```

You will need to load the following previously installed packages:

```
Optim

HTTP

GLM

LinearAlgebra

Random

Statistics

DataFrames

DataFramesMeta

CSV

MultivariateStats
```

In this problem set, we will practice estimating models that require dimension reduction of the covariates. These include Principal Components Analysis (PCA) and factor analysis.

1. Load the dataset `nlsy.csv` and estimate the following linear regression model:

   $$\log(wage) = \beta_0 + \beta_1 black + \beta_2 hispanic + \beta_3 female + \beta_4 school + \beta_5 gradHS + \beta_6 grad4yr + \varepsilon$$

2. Compute the correlation among the six `asvab` variables.

3. Estimate the same regression model as above, but now add the six `asvab` variables contained in the CSV file. Given your answer from question #2, do you think it will be problematic to directly include these in the regression?

4. Rather than including a large set of correlated variables, let's instead include the first principle component of this set as one additional regressor in the model from question #1.

   - Use the package `MultivariateStats`
   - In this package is a function called `fit()`
   - `M = fit(PCA, asvabMat; maxoutdim=1)` will give the first principle component, but note that `asvabMat` needs to be a $J \times N$ matrix, **not** a $N \times J$ matrix as you would usually use. See the examples at https://multivariatestatsjl.readthedocs.io/en/stable/pca.html for more details.
   - To get the first principle component returned as data, you will need to use `asvabPCA = MultivariateStats.transform(M, asvabMat)`.
     Again, `asvabPCA` is a $1 \times N$ array. You will need to reshape this array to add it as a covariate to your regression model.

5. Repeat question 4, but use `FactorAnalysis` instead of `PCA` (the syntax should be exactly the same)

6. Now estimate the full measurement system using either maximum likelihood or simulated method of moments. (You can take your pick, but I recommend using MLE.)

   The measurement system and log wage equation are specified as follows:

   $$asvab_j = \alpha_{0j} + \alpha_{1j} black + \alpha_{2j} hispanic + \alpha_{3j} female + \gamma_j \xi + \varepsilon_j \quad (1)$$

   $$\log(wage) = \beta_0 + \beta_1 black + \beta_2 hispanic + \beta_3 female + \beta_4 school + \beta_5 gradHS + \beta_6 grad4yr + \delta \xi + \varepsilon \quad (2)$$

   where ξ is a person-specific random factor that is assumed to be drawn from a standard normal distribution. If we could observe ξ, it would be an N-length vector (same as $\log(wage)$ and each of the $asvab_j$'s).

   The likelihood function for each observation of this system of equations is given by

   $$\mathscr{L}_i = \left\{ \prod_j \frac{1}{\sigma_j} \phi\left(\frac{M_{ij} - X_i^m \alpha_j - \gamma_j \xi_i}{\sigma_j} \right) \right\} \frac{1}{\sigma_w} \phi\left(\frac{y_i - X_i \beta - \delta \xi_i}{\sigma_w} \right) \quad (3)$$

where $\phi(\cdot)$ is the standard normal pdf. M_j is an $N \times 1$ vector of ASVAB scores for ASVAB$_j$. X^m is a matrix with N rows and the following columns: an intercept, and dummies for black, hispanic and female. y is the log wage and X is the $N \times K$ covariate matrix in question #1. The σ's are variance parameters to be estimated.

As mentioned above, ξ_i is a random factor that is distributed standard normal. Because it is unobserved, we will need to integrate it out of the likelihood function. Thus, the log likelihood function for the model is

$$\ell = \sum_i \log\left(\int \mathscr{L}_i \mathrm{d}F(\xi)\right) \tag{4}$$

I recommend using Gauss-Legendre quadrature (see Problem Set 4) to approximate this integral, but you could also use Monte Carlo integration.

7. Have an AI write unit tests for each of the functions you've created (or components of each) and run them to verify that they work as expected.