ECON 532: Homework 1

Shabab Ahmed

19 January, 2022

1 OLS

In this problem, we were asked to minimize the sum of squared errors using a built-in optimizer rather than through matrix algebra. The regression model is:

$$\text{Delay}_i = \beta_0 + \beta_1 \text{Distance}_i + \beta_2 \text{Departure Delay}_i + \beta_{3-8} \text{Day of week fixed effects}_i + \epsilon_i$$

We created dummy variables for each of the categories of the categorical variable 'Day of week fixed effects'. We will have to remove one of the dummy variables to run the regression if we want to include the constant term. This is to avoid perfect collinearity among the explanatory (independent) variables. That is, we need to select one category (dummy variable) as the baseline category (the category against which we compare all the other categories). We remove the indicator variable for 'Day 1'. Therefore, our β contains the coefficients in the following order:

$$\begin{bmatrix} \text{Constant} \\ \text{Distance} \\ \text{Departure Delay} \\ \text{Day 2} \\ \text{Day 3} \\ \text{Day 4} \\ \text{Day 5} \\ \text{Day 6} \\ \text{Day 7} \end{bmatrix}$$

We use MATLAB's *fminsearch* function along with an initial guess of $\beta = \mathbf{0}$ to find the minimizer of the sum of squared error:

$$\widehat{\beta} = \arg\min_{\beta} \sum_i (Y_i - X_i\beta)^2$$

We also calculated the analytical solution using $\widehat{\beta}_{ols} = (XX)^1 XY$. The following table contains both the numeric and analytic solution:

	Numerical	Analytical
$\widehat{\beta}_0$	-0.72714	-0.42353
$\widehat{\beta}_1$	-0.003412	-0.0031331
$\widehat{\beta}_2$	1.0163	1.0166
$\widehat{\beta}_3$	-0.33608	-0.91821
$\widehat{\beta}_4$	0.968	0.16934
$\widehat{\beta}_5$	-0.026091	-0.33446
$\widehat{\beta}_6$	-0.27277	-1.7165
$\widehat{\beta}_7$	0.009296	-1.5408
$\widehat{\beta}_8$	-0.9494	-0.84353

We can see that the solutions are different between the two methods. However, if we calculate the objective function at the estimated parameters with the two different methods, we find that the sum of squared residuals (objective function evaluated at the estimated values) only differ by 0.2% between the two methods. This gives us an indication that the numerical solution is not doing too poorly. We can also notice that the coefficients on 'Distance' and 'Departure Delay' from the numerical solution is much closer to the analytical solution. However, the differences are much greater between the numerical and analytical solution on the coefficients of the dummy variables and the constant term. This maybe due to the fact that 'Distance' and 'Departure Delay' variables are continuous in the sense that they affect the objective function for all of the data points. The dummy variables only affect the objective function for a portion of the data, namely, when it is that particular day of the week.

2 Maximum Likelihood

In this problem, we estimate a logit model using maximum likelihood. We know that the probability of a flight arriving more than 15 minutes late is:

$$P(Y_i > 15|X_i;\beta) = \frac{\exp(\beta X_i)}{1+\exp(\beta X_i)}$$

Then:

$$P(Y_i \leq 15|X_i;\beta) = 1 - \frac{\exp(\beta X_i)}{1+\exp(\beta X_i)} = \frac{1}{1+\exp(\beta X_i)}$$

We can easily calculate the log of these probabilities as the following:

$$\ln P(Y_i > 15|X_i;\beta) = \beta X_i - \ln(1+\exp(\beta X_i))$$

$$\ln P(Y_i \leq 15|X_i;\beta) = \ln(1) - \ln(1+\exp(\beta X_i)) = -\ln(1+\exp(\beta X_i))$$

Thus, the log likelihood of observing the flight arrival delays in the data is going to be given by:

$$\begin{aligned}
L(\beta) &= \sum \ln P(Y_i|X_i;\beta) \\
&= \sum \mathbb{1}_{Y_i>15}(Y_i)\left(\beta X_i - \ln(1+\exp(\beta X_i))\right) + \mathbb{1}_{Y_i\leq 15}(Y_i)\left(-\ln(1+\exp(\beta X_i))\right) \\
&= \sum \mathbb{1}_{Y_i>15}(Y_i)\left(\beta X_i - \ln(1+\exp(\beta X_i))\right) + \left(1-\mathbb{1}_{Y_i>15}(Y_i)\right)\left(-\ln(1+\exp(\beta X_i))\right) \\
&= \sum \mathbb{1}_{Y_i>15}(Y_i)\left(\beta X_i\right) + -\ln(1+\exp(\beta X_i))
\end{aligned}$$

where $\mathbb{1}_A(Y_i)$ is the indicator function which is equal to 1 if event A is true and 0 otherwise. In our code, we can make use of the binary variable for a flight arriving more than 15 minutes late as a proxy for this indicator function. We also recognize that maximizing the log-likelihood is equivalent to minimizing the negative of the log-likelihood. We use MATLAB's *fminsearch* function along with an initial guess of $\beta = \mathbf{0}$. The explanatory variables are arranged in the following order:

$$X_i' = \begin{bmatrix}\text{Constant} & \text{Distance} & \text{Departure Delay}\end{bmatrix}$$

We obtain the following solution:

$$\widehat{\beta}_{MLE} = \begin{bmatrix}\widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2\end{bmatrix} = \begin{bmatrix}-2.613 \\ -0.00013675 \\ 0.1295\end{bmatrix}$$

3 GMM

In the data, X contains a vector of 3 observed variables and Z contains 4 instruments. Our model is:

$$Y_i = \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \epsilon_i$$

with 4 instruments $Z_1, \ldots, Z_4$. The moment conditions are $\mathbb{E}[\epsilon|Z] = 0$ which implies the unconditional moment restriction $\mathbb{E}[Z(Y - X\beta)] = 0$.

Since we have more instruments than β's we will need a weighting matrix W to minimize our objective function:

$$Q_n(\beta; W) = g_n(\beta)' W g_n(\beta)$$

where

$$g_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} g_i(\beta).$$

We will use a two-step estimator. The first step estimator uses $W = I_4$. We use MATLAB's *fminsearch* along with intial guess of $\beta = \mathbf{0}$ to get the solution. We calculate the standard errors using the formula given in the problem set. The following table contains the results:

1st Stage	**Estimate**	**Standard error**
$\widehat{\beta}_0$	1.9187	0.18444
$\widehat{\beta}_1$	1.1036	0.18665
$\widehat{\beta}_2$	3.6526	0.16614

We also provide the estimated covariance matrix:

$$\widehat{V}_1 = \begin{bmatrix} 34.0171601251727 & -2.23654910713410 & 2.08889597684759 \\ -2.23654910713410 & 34.8385536607512 & -9.29029189314368 \\ 2.08889597684759 & -9.29029189314369 & 27.6019444842134 \end{bmatrix}$$

For the second stage estimator, we use the weight matrix $\widehat{W} = \Sigma^{-1}(\widehat{\beta})$ for

$$\Sigma(\widehat{\beta}) = \sum_{i=1}^{n} \widehat{\epsilon}_i^2 z_i z_i'.$$

We use the MATLAB's *fminsearch* again along with an initial guess of $\beta = \mathbf{0}$. The results are provided below:

2nd Stage	**Estimate**	**Standard error**
$\widehat{\beta}_0$	1.921	0.18444
$\widehat{\beta}_1$	1.094	0.18642
$\widehat{\beta}_2$	3.6614	0.16589

The estimated covariance matrix is:

$$\widehat{V}_2 = \begin{bmatrix} 34.0180565431051 & -2.22307242797331 & 2.07841083777408 \\ -2.22307242797331 & 34.7519436873242 & -9.21539765249504 \\ 2.07841083777408 & -9.21539765249505 & 27.5203787876216 \end{bmatrix}$$

The point estimates of the second stage estimator is closer to the true value than the first stage estimator. However, they are not equal to the true value and this makes sense because we have omitted variable bias. The standard errors of the second stage estimators are also lower than the first stage estimators. This also makes sense because the weight matrix used in the second stage is supposed to be optimal in the sense that the estimator is efficient. This means that use of this weight matrix should give the lowest standard errors. However, we only use an estimate of the optimal weight matrix and we could probably get lower standard errors by doing iterated GMM.