Alex Miller

Problem Set 1

**Linear Algebra**

1.

(a)

Within the mata environment, this same operation was performed with the following commands:

**: educ=(2\3\2\1\1);**

**: x=J(5,4,.);**

**: for (i=1; i<=rows(x); i++) {**

**x[i,1]=1;**

**x[i,2]=(educ[i]==1?1:0);**

**x[i,3]=(educ[i]==2?1:0);**

**x[i,4]=(educ[i]==3?1:0);**

**};**

After entering our initial vector and adding the column of 1s, a series of conditional statements in a loop are used to determine dummy variable values, which give the following matrix:

**1 2 3 4**

**+-----------------+**

**1 | 1 0 1 0 |**

**2 | 1 0 0 1 |**

**3 | 1 0 1 0 |**

**4 | 1 1 0 0 |**

**5 | 1 1 0 0 |**

**+-----------------+**

The equation for our hypothetical model in this problem is as follows:

The first column of our X matrix consists solely of the value 1 in order to estimate the model intercept (a constant value – β0). The second column of the matrix is a dummy variable (1 for true and 0 for false) representing observations with less than high school education. The information from this column will be used to determine an estimate for β1. The third column of the matrix is a dummy variable representing observations with at least a high school education, but no degree. This column will be used to determine an estimate for β2. The forth column of the matrix is a dummy variable representing observations with a degree or higher education. This column will be used to determine an estimate for β3.

The rank of the matrix X was determined with the command:

**: rank(x);**

The result of this command returned 3. Since the matrix X has 4 columns, X does not have full rank. This is because there are not independent signals coming from each column. The inclusion of all three dummy variables in the model introduces perfect collinearity; Because there are only three categorical options, if you are not either of the first two, you must belong to the last category.

2.

(a)

The same operation was performed in the mata environment with the following commands:

**: z=(2,6\9,2);**

**: m=(4\8);**

**: luinv(z);**

**: z\*luinv(z);**

First, the data was entered into the two matrices. Second, the inverse of Z is calculated. Lastly, Z is multiplied by it’s inverse to ensure the inverse was correctly calculated. The result was as such:

**1 2**

**+---------------+**

**1 | -.04 .12 |**

**2 | .18 -.04 |**

**+---------------+**

(b)

The same results were achieved in mata with the following commands:

**: z'z;**

**: rows(z'z)==cols(z'z);**

The Z’ was multiplied with Z, and then the resulting matrix was checked for being square. These commands returned:

**[symmetric]**

**1 2**

**+-----------+**

**1 | 85 |**

**2 | 30 40 |**

**+-----------+**

and a **1** indicated that the rows (2) did indeed match the columns (2).

(c)

The above operations were done within mata with the following command:

**: (z\*m)'==m'z';**

The above command returned a **1** indicating that those two matrix products were equivalent.

**OLS**

3.

5.

Only two assumptions are necessary for b to be an unbiased estimate of β (E(b)= β). Those two assumptions are: E[ϵ]=0 and E[ϵ|x]=0. Heterskedasticity is not a determining factor for the bias of the estimator; it only factors into the bias for the variance of our estimators.

6.

(a)

The dataset contains the full range of data utilized by Mroz’s analysis. Mroz’s paper notes 753 observations and there are 753 rows in the dataset. Mroz had 428 working women noted in his study, and 428 rows in the dataset have a 1 for the working dummy variable. These two numbers were counted with the following two commands in STATA:

**. sum \***

**. sum \* if lfp==1**

(b)

The model used to estimate a labor supply curve for working women for this question is as follows:

*(Working hours)= β0 + β1\*(wage) + β2\*(family income) + β3\*(children under 6) + β4\*(children from 6-18) + β5\*(unemployment rate) + β6\*(city dummy) + β7\*(mother’s education) + β8\*(father’s education)*

This model was estimated using the following commands:

**. drop if lfp==0**

**. reg whrs ww faminc kl6 k618 un cit wmed wfed, robust**

First, observations where women did not work were excluded from the model. This was done mainly because the issue being studied is not incentives to enter or exit the labor market, but rather the choice of hours of labor once one is already in the labor market. Then, the model was estimated using robust standard errors. The following table reflects the results of that analysis:

**Linear regression Number of obs = 428**

**F( 8, 419) = 7.82**

**Prob > F = 0.0000**

**R-squared = 0.1101**

**Root MSE = 739.25**

**----------------------------------------------------------------**

**| Robust**

**whrs | Coef. Std. Err. t P>|t|**

**-------------+--------------------------------------------------**

**ww | -36.46969 9.866642 -3.70 0.000**

**faminc | .0137707 .0040066 3.44 0.001**

**kl6 | -252.7355 115.9428 -2.18 0.030**

**k618 | -104.4524 25.83098 -4.04 0.000**

**un | -19.27316 11.33525 -1.70 0.090**

**cit | -74.34996 80.14519 -0.93 0.354**

**wmed | 6.310049 12.8807 0.49 0.624**

**wfed | -17.17906 13.22798 -1.30 0.195**

**\_cons | 1606.146 171.5121 9.36 0.000**

**----------------------------------------------------------------**

β0 was estimated as 1606.15, so at a hypothetical wage of zero, the model estimates that about 1606 hours of work will be performed in a year. β1 was estimated as -36.47; for every $1 increase in wage, we can expect about a 36 hour reduction in work per annum. This sort of behavior is expected towards the top of a normal backwards-bending labor supply curve. β4 estimates the effect of overall family income on hours worked, but the positive relationship might only be indicative of the fact that the wife’s income is also part of the family’s. Both β3 and β4 estimate the cost to work of having children at being very high and quite significant.

The additional “background” variables, as Mroz calls them, serve to minimize the possibility of omitted variable bias within the model. The city dummy should help capture otherwise exogenous effects on hours worked between cities and the countrysides. Similarly, the unemployment rate will capture exogenous variables about regional economic performance while parental education should capture some measure of social standing or upbringing.

(c)

1.

In order to replicate the “canned” regression package STATA offers, the following commands were run in mata:

**: y=st\_data(.,("whrs"))**

**: x=st\_data(.,("ww", "faminc", "kl6", "k618", "un", "cit", "wmed", "wfed"))**

**: x=(J(rows(x),1,1),x);**

**: b = invsym(x'x)\*(x'y);**

**: b**

**: yhat = x\*b;**

**: ehat = y-yhat;**

**: s2 = (ehat'ehat)/(rows(x)-rows(b))**

**: rmse=sqrt(s2)**

**: varcov = invsym(x'x)\*s2**

**: se=sqrt(diagonal(varcov))**

**: tstat = diagonal(diag(b) \* invsym(diag(se)))**

First, the relevant data were imported from STATA into mata. A row of 1s was appended to our independent variables to allow the calculation of our constant term. The OLS estimator was calculated, and both fitted values and error terms were collected. Using those error terms, we’re able to compute the mean squared error and the root mean squared error. A variance-covariance matrix was calculated from x and the sum of the squared variation in the error term, and that number was used to determine standard errors and t-statistics. These commands produced the following results:

**b =**

**+----------------+**

**1 | 1606.145543 |**

**2 | -36.46968706 |**

**3 | .0137707056 |**

**4 | -252.7354932 |**

**5 | -104.4524396 |**

**6 | -19.27316192 |**

**7 | -74.34995671 |**

**8 | 6.310049378 |**

**9 | -17.17906122 |**

**+----------------+**

**s2 = 546488.6759**

**rmse = 739.248724**

**se =**

**+---------------+**

**1 | 167.3847437 |**

**2 | 11.40827477 |**

**3 | .0033505123 |**

**4 | 92.39424636 |**

**5 | 27.57785316 |**

**6 | 12.01233581 |**

**7 | 79.59858625 |**

**8 | 13.10748882 |**

**9 | 12.55597533 |**

**+---------------+**

**t-stat =**

**+----------------+**

**1 | 9.595531276 |**

**2 | -3.196774953 |**

**3 | 4.110029822 |**

**4 | -2.735402941 |**

**5 | -3.787547894 |**

**6 | -1.60444748 |**

**7 | -.9340612719 |**

**8 | .4814079544 |**

**9 | -1.368198071 |**

**+----------------+**

For a typical numbers of degrees of freedom, most t-statistics above 2 show that the results are significant to a 0.05 alpha. In this case, our t-statistic shows that the constant, the wage, family income, number of children under 6, and the number of children from 6-18 are all significant determinates of the number of hours these women worked. Unfortunately, as we’ll see later, there is a relatively good chance that there is heteroskedasticity in this data, so the standard errors used to compute these t-statistics should not be used for inference even though the betas are unbiased.

2.

In order to calculate the model R2, the following commands were used in mata:

**: ybar = mean(y)**

**: r2 = 1-((ehat'ehat)/((y:-ybar)'(y:-ybar)))**

To calculate the R2 here, I used 1 minus the ratio of the sum of the squared errors to the total variation in our dependent variable from its mean. This leaves us with the percent of the variation in our dependent variable that is explainable by the model.

**r2 = 0.1101090218**

In other words, only about 11% of the total variation in women’s work hours is explainable given our current model.

3.

The following commands were used in mata in order to estimate robust standard errors:

**: e2 = ehat\*ehat'**

**: vhat=diag(e2)**

**: rvarcov = invsym(x'x)\*x'\*vhat\*x\*invsym(x'x)**

**: robustse = sqrt(diagonal(rvarcov))**

The output from these commands is:

**robustse =**

**+---------------+**

**1 | 169.699215 |**

**2 | 9.762352529 |**

**3 | .0039642329 |**

**4 | 114.7173364 |**

**5 | 25.55794746 |**

**6 | 11.21544078 |**

**7 | 79.29806214 |**

**8 | 12.74455115 |**

**9 | 13.08816119 |**

**+---------------+**

As opposed to normal standard errors, these standard errors use the weight of each error term in order to correct what would otherwise be biased standard errors. The observations that have large error variance are given less weight in calculating these robust standard errors than for observations that have a very small error variance.

(d)

The STATA packaged heteroskedasticity test was performed by first running the regression again and then with the command:

**. estat hettest**

This test gave a **Chi2 value of 3.74** with an associated **p-value of 0.053.** Although 0.053 is not technically small enough to reject the null hypothesis of homoskedasticity, it is close enough to warrant using robust standard errors. The “manual” method is performed as follows:

**. predict e, resid**

**. predict yhat, xb**

**. gen r = (e^2)/(e(rss)/e(N))**

**. reg r yhat**

**. display e(mss)/2**

The **Chi2 value for this test came out as 3.7431639.** Intuitively, this test works because it is looking directly at the relationship between the first model’s residuals (squared errors) and our fitted values (xb). After obtaining both the squared errors and the fitted values from the first model, the regression between the two tells us to what extent the variation in the error term can be explained by variation in x. If the model sum of squares for that regression is large, then the error term varies greatly with x and there is a large problem with heteroskedasticity.