Using Stata with Multiple Regression & Matrices

1. Matrix calculations with Stata. Stata has several built-in functions that make it work as a matrix calculator. These functions are probably primarily helpful to programmers who want to write their own routines.

To get the M matrix, you can use the `mat accum` command. The `mat accum` command adds $X_0$ to the list of variables (where $X_0 = 1$ for all cases) and then computes all cross-products.

```stata
. use http://www.nd.edu/~rwilliam/stats1/statafiles/reg01.dta, clear
. mat accum m = income educ jobexp
(obs=20)
. mat list m
```

```
symmetric m[4,4]
    income    educ    jobexp    _cons
income  13742.27
    educ  6588.3  3285
jobexp  6448.9  2999  3767
    _cons  488.3  241  253  20
```

To get the XP matrix of cross-product deviations from the means, we add the `dev` and `noconstant` parameters. The `dev` parameter subtracts the mean of the variable from each case while `noconstant` keeps $X_0$ from being added to the data.

```stata
. mat accum xp = income educ jobexp, dev noconstant
(obs=20)
. mat list xp
```

```
symmetric xp[3,3]
    income    educ    jobexp
income  1820.4255
    educ  704.28499  380.95
jobexp  271.90499  -49.65  566.55
```

The covariance matrix can now be computed from the xp matrix. The sample size used when computing the xp matrix is stored by Stata in a scalar called `r(N)`.

```stata
. mat s = xp/(r(N)-1)
. mat list s
```

```
symmetric s[3,3]
    income    educ    jobexp
income   95.811867
    educ  37.067631  20.05
jobexp  14.310789  -2.6131579  29.818421
```

The `corr` function can be used to compute the correlations of the variables. The correlations can be computed from either the xp or covariance matrix.
. mat r = corr(s)

. mat list r

symmetric r[3,3]
   income   educ   jobexp
income   1
   educ  .84572271   1
   jobexp .26773898 -10687254   1

It wouldn’t be as much fun, but you can just use the `corr` program to get the covariances and correlations. To get the correlations,

```
corr income educ jobexp
(obs=20)
```

```
|       income     educ   jobexp
|---------------------------
income |  1.0000
educ   |  0.8457   1.0000
jobexp |  0.2677 -10.69  1.0000
```

To get the covariances instead, use the `cov` parameter.

```
corr income educ jobexp, cov
(obs=20)
```

```
|       income     educ   jobexp
|---------------------------
income |   95.8119
educ   |   37.0676   20.05
jobexp |  14.3108 -2.61316  29.8184
```

2. Do it yourself regression. Want to double-check Stata’s regression estimates? You can do it with Stata’s matrix commands. Recall that \( b = (X'X)^{-1}X'Y \). In words, we say \( b \) equals \( X \) prime \( X \) inverse \( X \) prime \( Y \). \( X'X \) is the cross-product matrix of the \( X \)’s with each other, including \( X_0 \). To compute it in Stata,

```
mat accum xprimex = educ jobexp
(obs=20)
```

```
mat list xprimex
```

```
symmetric xprimex[3,3]
   educ   jobexp   _cons
   educ   3285
   jobexp 2999   3767
   _cons  241   253   20
```

\( X'Y \) is the cross-product of \( Y \) with each of the \( X \)’s. `mat vecaccum` will compute \( Y'X \) for us. It computes the cross-product of the first variable listed with all the other variables listed.
. mat vecaccum yprimex = income educ jobexp

. mat list yprimex

yprimex[1,3]
        educ   jobexp    _cons
income  6588.3  6448.9   488.3

Note that this is a row vector. To get $X'Y$, which is a column vector, we simply transpose $Y'X$.

. mat xprimey = yprimex'

. mat list xprimey

xprimey[3,1]
     income
        educ  6588.3
     jobexp  6448.9
    _cons   488.3

Now we are ready for the final calculation!

. mat b = inv(xprimex)*xprimey

. mat list b

b[3,1]
      income
        educ  1.9333928
     jobexp   .64936536
    _cons  -7.0968549

3. [Optional] Proof that $b = (X'X)^{-1}X'Y$. Let $X$ be an $N \times K$ matrix (i.e. $N$ cases, each of which has $K$ X variables, including $X_0$.) $Y$ is an $N \times 1$ matrix. $e$ is an $N \times 1$ matrix. Then, if the assumptions of OLS regression are met,

<table>
<thead>
<tr>
<th>Equation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y = Xb + e$</td>
<td></td>
</tr>
<tr>
<td>$Y - e = Xb$</td>
<td>Subtract $e$ from both sides</td>
</tr>
<tr>
<td>$X'(Y - e) = X'Xb$</td>
<td>Premultiply both sides by $X'$</td>
</tr>
<tr>
<td>$X'Y = X'Xb$</td>
<td>If the assumptions of OLS regression are met, $X'e = 0$ because the Xs are uncorrelated with the residuals of $Y$</td>
</tr>
<tr>
<td>$(X'X)^{-1}X'Y = (X'X)^{-1}X'Xb$</td>
<td>Premultiply both sides by $(X'X)^{-1}$</td>
</tr>
<tr>
<td>$(X'X)^{-1}X'Y = b$</td>
<td>$(X'X)^{-1}X'X = I$ and $Ib = b$</td>
</tr>
</tbody>
</table>