Revision Questions

Table of contents

1 Basics

1. Consider two random k-dimension vectors $\{X, Y\}$ and non-random $k \times k$ matrices $\{A, B\}$. Show,

   1.1. $Var(AX) = AVar(X)A'$

   1.2. $Var(AX + b) = AVar(X)A'$ for non-random k-dimension vector b

   1.3. $Cov(AX, BY) = ACov(X, Y)B'$

2. Suppose $X \sim N(\mu, \Sigma)$, with $X \in \mathbb{R}^k$. Find the distribution, $Y = AX + b$, for non-random $k \times k$ matrix A and non-random k-dimension vector b.

2 CLRM

1. Which of the CLRM assumptions is required for identification of β? Demonstrate this claim.
2. Provide an example of a model that is non-linear in regressors, but linear in parameters. Similarly, provide an example of a model that is linear in regressors, but non-linear in parameters.
3. Suppose, the true data generating process was given by,

$$Y = X\beta + \epsilon$$

   where $E[\epsilon|X] = \alpha$ and X included a constant term with parameter β_1. Is β_1 identified?
4. Given the result $\beta = E[X_iX_i']^{-1}E[X_iY_i]$. With two regressors, $X_i = [1\ X_{2i}]'$, show

   4.1. $\beta_2 = \frac{Cov(X_{2i}, Y_i)}{Var(X_{2i})}$

   4.2. $\beta_1 = E[Y_i] - \beta_2 E[X_i]$

3 OLS

1. Consider the projection matrices $P_X = X(X'X)^{-1}X'$ and $M_X = I_n - P_X$. Show,

   1.1. $P_X X = X$

   1.2. $M_X X = 0$

   1.3. $P_X M_X = 0$

   1.4. $X'P_X = X'$
2. Show that $X'X$, where X is a $n \times k$ random matrix, can be expressed as $\sum_{i=1}^{n} X_iX_i'$, where X_i is a $k \times 1$ vector.
3. Consider the partitioned regression model,

$$Y = X_1\beta_1 + X_2\beta_2 + u$$

   Show,

   3.1. $E[\hat{\beta}_1|X] = \beta_1$ where $\hat{\beta}_1 = (X_1'M_2X_1)^{-1}X_1'M_2Y$.

   3.2. Write down the conditional variance of the OLS estimator for β_1, assuming homoskedasticity: $E[uu'|X] = \sigma^2 I_n$.

   3.3. Write down the conditional variance of the OLS estimator for β_1, assuming heteroskedasticity: $E[uu'|X] = \Omega$.

4. Demonstrate the BLUE result: the OLS estimator ($\hat{\beta}$) is the Best Linear Unbiased Estimator. Consider the alternative unbiased, linear estimator $b = AY$; such that,

$$E[b|X] = \beta$$

4.1. Show that since b is unbiased, it must be that $AX = I_k$.

4.2. Using the above result, show that under CLRM 1-6 (i.e., including homoskedasticity),

$$Cov(\hat{\beta}, b|X) = Var(\hat{\beta}|X)$$

4.2. Show that $Var(b|X) - Var(\hat{\beta}|X) \geq 0$ (i.e. a positive semi-definite matrix), by solving for

$$Var(\hat{\beta} - b|X)$$

5. Consider the GLS estimator ($\tilde{\beta}$), which solves the problem

$$\min_b \, (Y - Xb)'\Omega^{-1}(Y - Xb)$$

where $E[\varepsilon\varepsilon'|X] = \Omega$.

5.1. Show that $Var(\hat{\beta}|X) - Var(\tilde{\beta}|X) \geq 0$, where $\hat{\beta}$ is the OLS estimator.

5.2. Under which assumption are the two estimators equivalent?

4 Linear Tests

1. Under CLRM 1-6, solve for the finite sample distribution of $Rb\hat{e}ta$, where R is non-random $k \times k$ matrix.

2. Consider the multiple, linear hypotheses:

$$\begin{aligned} H_0 : \beta_2 - 4\beta_4 &= 1 \\ \beta_3 &= 3 \\ \beta_5 &= \beta_6 \end{aligned}$$

2.1. Write the 4 hypotheses in the form $R\beta = r$.

2.2. Write down the F-statistic for the test (assuming homoskedasticity) as well as it's finite sample distribution.

2.3. What is the asymptotic distribution of this test statistic.

2.4. For a linear model with $k = 6$, write the restricted model corresponding to the above hypotheses.

3. Consider the model of household food expenditure,

$$foodexp_i = \beta_1 + \beta_2 inc_i + \beta_3 hhsize_i + \beta_4 hhsize_i^2 + \varepsilon_i$$

3.1. Suppose you wish to test the hypothesis of increasing returns to food consumption in the household: each additional household member is marginally cheaper to feed. Which is a more powerful test:

$$H_0 : \beta_4 = 0 \qquad \text{against} \qquad H_1 : \beta_4 \neq 0$$

or,

$$H_0 : \beta_4 \geq 0 \qquad \text{against} \qquad H_1 : \beta_4 < 0$$

3.2. Transform the model to test the hypothesis $H_0 : -\beta_3/\beta_4 = 5$, using just the coefficient on the variable $hhsize$.

3.3. Transform the model to test the same hypothesis, using just the coefficient on the variable $hhsize^2$.

5 Panel Data

1. Show that $\tilde{Y}_i = Y_i - \bar{Y}_i$, where $\bar{Y}_i = \frac{1}{T}\sum_{t=1}^{T} Y_{it}$, can be written as $Y = M_\ell Y_i$, where $M_\ell = \ell(\ell'\ell)^{-1}\ell'$ and ℓ is a $T \times 1$ vector of 1's.
2. Show that $X'X$, where X is a $nT \times k$ random matrix, can be expressed as $\sum_{i=1}^{n} X_i' X_i$, where X_i is a $T \times k$ matrix.
3. Show that $I_n \otimes M_\ell$ is an idempotent matrix and $\sum_{i=1}^{n} \tilde{X}_i' \tilde{X}_i = X'(I_n \otimes M_\ell)X$.
4. Show that for T=3, $D'(DD')^{-1}D = M_\ell$.
5. Demonstrate that the OLS estimator of the 'within-unit' transformed model is unbiased. Is it efficient?
6. Conditional variance of random effects model,

6.1. Solve for $Var(\hat{\beta}^{OLS}|X)$, the variance of the ('pooled') OLS estimator for the random effects model.

6.2. Solve for $Var(\hat{\beta}^{GLS}|X)$, the variance of the GLS estimator for the random effects model.

6.3. Verify that $Var(\hat{\beta}^{OLS}|X) - Var(\hat{\beta}^{GLS}|X) \geq 0$, is a positive-semidefinite matrix.

7. Conditional variance of first-differenced transformation

   7.1. Solve for $Var(\hat{\beta}^{FD-OLS}|X)$, the variance of the OLS estimator for the FD transformation.

   7.2. Solve for $Var(\hat{\beta}^{FD-GLS}|X)$, the variance of the GLS estimator for the FD transformation.

   7.3. Verify that $Var(\hat{\beta}^{FD-OLS}|X) - Var(\hat{\beta}^{FD-GLS}|X) \geq 0$, is a positive-semidefinite matrix.

8. Why is it important that the assumed model underlying the 'pooled' OLS estimator and 'within' OLS estimator are the same in the Hausman test? That is, why cannot we not compare the LSDV estimator of the fixed-effects model with the 'pooled' OLS estimator?

9. What is the purpose of the Mundlack correction?

6 Binary Outcome Models

1. Consider the Logit model.

   1.1. Write down the likelihood function.

   1.2. State the maximization problem that solves for ML estimator.

   1.3. Solve for the F.O.C.s of the ML problem.

   1.4. Solve for the asymptotic variance-covariance matrix of $\hat{\beta}^{ML}$.

2. Consider the Probit model.

   2.1. Write down the likelihood function.

   2.2. State the maximization problem that solves for ML estimator.

   2.3. Solve for the F.O.C.s of the ML problem.

   2.4. Solve for the asymptotic variance-covariance matrix of $\hat{\beta}^{ML}$.

3. From the proof of the asymptotic normality, show

$$E\left[\frac{\partial^2 f(Y_i|X_i;\beta_0)/\partial\beta\partial\beta'}{f(Y_i|X_i;\beta_0)}\right] = 0$$

<IPython.core.display.HTML object>

```
Iteration 0:  Log likelihood = -209.35624
Iteration 1:  Log likelihood = -205.66439
Iteration 2:  Log likelihood = -205.27888
Iteration 3:  Log likelihood = -205.27756
Iteration 4:  Log likelihood = -205.27756

Logistic regression                                     Number of obs = 165,038
                                                        LR chi2(7)    =    8.16
                                                        Prob > chi2   =  0.3189
Log likelihood = -205.27756                             Pseudo R2     =  0.0195

------------------------------------------------------------------------------
    employed | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
      eduyrs |  -.2139028   .0990457    -2.16   0.031    -.4080287   -.0197768
       exper |   .0408304   .0861167     0.47   0.635    -.1279552    .2096161
      exper2 |  -.0012522   .0024549    -0.51   0.610    -.0060637    .0035593
   1.married |  -.1047635   .5283537    -0.20   0.843    -1.140318    .9307907
    1.female |   .0208576   .5035107     0.04   0.967    -.9660052    1.007721
     1.child |   .4950251   .7490084     0.66   0.509    -.9730045    1.963055
             |
female#child |
        1 1  |   .6419965   1.046185     0.61   0.539    -1.408489    2.692482
             |
       _cons |   11.64254   1.476069     7.89   0.000     8.749502    14.53559
------------------------------------------------------------------------------
```

4. Consider the logit model output above, estimated using the dataset from Problem Set 4. The sample and variables are defined as in PS 4.

   4.1. Compute the marginal effect of an additional year of education on the probability of being employed for a childless, unmarried, female with 15 years of education and 5 years of (potential) experience.

   4.2. Compute the marginal effect of being married for a male with children, 12 years of education and 10 years of experience.

7 Endogenous Selection Models

1. Consider the endogenous the right-censored Tobit model. The observed distribution is,

$$f(y) = \begin{cases} f^*(y) & \text{for} \quad y < 0 \\ F^*(0) & \text{for} \quad y = 0 \\ 0 & \text{for} \quad y > 0 \end{cases}$$

Suppose, $Y^* = X_i'\beta + \varepsilon_i$, where the error term has a normally distributed (conditional on X).

1.1. Solve for $E[Y_i|Y_i < 0]$

1.2. Write down the likelihood function of the observed data.

1.3. Solve for $\frac{\partial E[Y_i|X_i, Y_i<0]}{\partial X_i}$

1.4. Solve for $\frac{\partial E[Y_i|X_i]}{\partial X_i}$

1.5. How do the answers compare to the case where Y is left-censored at 0?

```
Iteration 0:  Log likelihood = -209.35624
Iteration 1:  Log likelihood = -205.60572
Iteration 2:  Log likelihood = -205.33806
Iteration 3:  Log likelihood = -205.33756
Iteration 4:  Log likelihood = -205.33756

Probit regression                                       Number of obs = 165,038
                                                        LR chi2(7)    =    8.04
                                                        Prob > chi2   =  0.3293
Log likelihood = -205.33756                             Pseudo R2     =  0.0192

------------------------------------------------------------------------------
    employed | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
      eduyrs |  -.0544498   .0255841    -2.13   0.033    -.1045937   -.0043059
       exper |   .0095296   .0223186     0.43   0.669    -.0342139    .0532732
      exper2 |  -.0002998   .0006319    -0.47   0.635    -.0015382    .0009387
   1.married |  -.0259376   .1384233    -0.19   0.851    -.2972423    .2453672
    1.female |   .0049027   .1321027     0.04   0.970    -.2540139    .2638192
     1.child |   .1297358   .1937444     0.67   0.503    -.2499962    .5094679
             |
female#child |
        1 1  |   .1540261   .2634162     0.58   0.559    -.3622602    .6703123
             |
       _cons |   4.341448   .3821216    11.36   0.000     3.592503    5.090392
------------------------------------------------------------------------------
```

2. Consider an endogenous threshold model, for (log of) employee earnings. Included in the model is a linear term in years of education, quadratic terms in years of (potential) experience, a married dummy-variable, and a female dummy-variable.

   2.1. Consider the output above from the selection equation, which includes an additional interaction between gender and presence of children (under 18). Compute the Inverse Mills Ratio for a married women, with children, 15 years of education and 8 years of experience.

   2.2. Write down the likelihood function of the observed outcomes: $[ln(wage_i), Employed_i]$. Where the latter is a dummy variable indicating employment (positive earnings).

   2.3. The Heckit output is given below. What can you conclude regarding selection into employment. And, under what assumptions.

   2.4. Does the interaction bewteen gender and parenthood belong in the main equation?

```
note: two-step estimate of rho = 19.806689 is being truncated to 1

Heckman selection model -- two-step estimates   Number of obs     =    115,352
(regression model with sample selection)              Selected    =    115,331
                                                      Nonselected =         21

                                                Wald chi2(5)      =       4.66
                                                Prob > chi2       =     0.4590

------------------------------------------------------------------------------
             | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
lnwage       |
      eduyrs |   .1289768   .0827439     1.56   0.119    -.0331983     .291152
       exper |   .0559093   .0878259     0.64   0.524    -.1162263    .2280449
      exper2 |  -.0010644   .0023925    -0.44   0.656    -.0057535    .0036248
   1.married |   .1015441   .4325466     0.23   0.814    -.7462316    .9493198
    1.female |  -.3315884   .3953545    -0.84   0.402    -1.106469    .4432921
       _cons |   4.387386   1.504299     2.92   0.004     1.439014    7.335759
-------------+----------------------------------------------------------------
employed     |
       exper |   .0079269   .0219371     0.36   0.718    -.0350691    .0509228
      exper2 |  -.0001872   .0006005    -0.31   0.755     -.001364    .0009897
   1.married |  -.0687001   .1375189    -0.50   0.617    -.3382323    .2008321
    1.female |   -.023606   .1337126    -0.18   0.860    -.2856778    .2384659
     1.child |   .1552287   .1938012     0.80   0.423    -.2246146    .5350721
             |
```

```
female#child |
        1 1  |   .1308613   .2683362     0.49   0.626    -.3950681    .6567907
             |
       _cons |   3.482178   .1598692    21.78   0.000      3.16884    3.795516
-------------+----------------------------------------------------------------
/mills       |
      lambda |   66.28671    929.642     0.07   0.943    -1755.778    1888.352
-------------+----------------------------------------------------------------
         rho |    1.00000
       sigma |  66.286706
------------------------------------------------------------------------------
```