Supplemental Notes: Multicollinearity

Discussion multi-collinearity

- LSA4: The regressors are perfectly multi-collinear if one of the regressors is a perfect linear function of the others

- Example: X_{1i}=(=1 if observation i is male)
  X_{2i}=(=1 if observation i is female)
  So: $X_{1i} + X_{2i} = 1$, perfectly collinear with intercept

- LSA4 is "testable". If two (or more) regressors are perfectly collinear, Stata will throw one out of the regression

- It simply means that you cannot separately identify the effect of the multi-collinear regressors on Y

"Imperfect" Multi-Collinearity

- Consider the model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$

- Suppose $X_{2i} = \gamma_0 + \gamma_1 X_{1i} + v_i$, where $\gamma_1 \neq 0$
- $\Rightarrow$The smaller the variance of v_i, the more collinear X_2 and X_1 become

- The more multi-collinear X_2 and X_1 are, the more "unstable" the OLS estimates of β_1 and β_2 become, and also the larger their standard errors become

- STATA simulation example to demonstrate this result

STATA simulation example

- Fix population parameter values: $\beta_0=5$, $\beta_1=1$, and $\beta_2=-2$
- Generate X_{1i}~Uniform[0,20]
- Generate $X_{2i} = -5 + 3*X_{1i} + v_i$ $v_i \sim N(0,\delta^2)$
- Generate random error terms $u_i \sim N(0,\sigma^2)$
- By construction $E[u_i|X_{1i},X_{2i}]=0$

- Construct $Y_i = 5 + 1*X_{1i} + -2*X_{2i} + u_i$, $i=1,\ldots,10{,}000$

- Consider cases with different values for $Var(v_i)$

- The smaller the value of $Var(v_i)$, the more multi-collinear X_1 and X_2 become

Case 1: $Var(v_i) = 1$

Variable	Obs	Mean	Std. Dev.	Min	Max
b0	50	5	0	5	5
b1	50	1	0	1	1
b2	50	-2	0	-2	-2
b0_estim	50	5.101321	.207221	4.58962	5.475749
b1_estim	50	.9436798	.1127736	.6963274	1.222902
b2_estim	50	-1.981726	.0379931	-2.073606	-1.898584
s_b1_estim	50	.1512652	.000971	.1498006	.1538676
s_b2_estim	50	.0503188	.000323	.0498316	.0511845
R2	50	.9705463	.000374	.9695491	.9710928
b21_estim	50	3.001116	0	3.001116	3.001116
r_12	50	.9983299	0	.9983299	.9983299
meanX2	50	24.96389	0	24.96389	24.96389
meanX1	50	9.988091	0	9.988091	9.988091
meanY	50	-34.9447	.058045	-35.08013	-34.82299
nobs	50	10000	0	10000	10000
sample	50	25.5	14.57738	1	50

Case 2: $Var(v_i)=0.25$

Variable	Obs	Mean	Std. Dev.	Min	Max
b0	50	5	0	5	5
b1	50	1	0	1	1
b2	50	-2	0	-2	-2
b0_estim	50	5.375429	.7487377	3.485523	6.9843
b1_estim	50	.7792146	.4543434	-.2164099	1.885361
b2_estim	50	-1.926904	.1519723	-2.294426	-1.594339
s_b1_estim	50	.6039451	.0038768	.5980974	.6143356
s_b2_estim	50	.2012752	.001292	.1993264	.2047381
R2	50	.9703999	.0003765	.969398	.9709502
b21_estim	50	3.000279	0	3.000279	3.000279
r_12	50	.9998953	0	.9998953	.9998953
meanX2	50	24.96418	0	24.96418	24.96418
meanX1	50	9.988091	0	9.988091	9.988091
meanY	50	-34.94528	.058045	-35.08071	-34.82357
nobs	50	10000	0	10000	10000
sample	50	25.5	14.57738	1	50

Case 3: $Var(v_i)=0.005$

Variable	Obs	Mean	Std. Dev.	Min	Max
b0	50	5	0	5	5
b1	50	1	0	1	1
b2	50	-2	0	-2	-2
b0_estim	50	23.28391	37.97198	-68.64539	106.3698
b1_estim	50	-9.965876	22.79383	-59.84771	45.16391
b2_estim	50	1.654792	7.598506	-16.72061	18.28276
s_b1_estim	50	30.19135	.1938002	29.89902	30.71077
s_b2_estim	50	10.06376	.0646	9.966321	10.2369
R2	50	.9703856	.0003769	.9693832	.9709364
b21_estim	50	3.000005	0	3.000005	3.000005
r_12	50	.9999999	0	.9999999	.9999999
meanX2	50	24.96427	0	24.96427	24.96427
meanX1	50	9.988091	0	9.988091	9.988091
meanY	50	-34.94547	.0580453	-35.0809	-34.82376
nobs	50	10000	0	10000	10000
sample	50	25.5	14.57738	1	50

Implications

- Strong near multi-collinearity leads to:

- Unreliable estimates of the regression coefficients

- Very large standard errors