Uganda National Panel Survey:

Regression Analysis

Wave 1: 2010 - 2011
Wave 2: 2011 - 2012

COURSE:
GHDP 7048 Advanced Econometrics I.

PRESENTED BY:
Amber Ni

Summary Statistics Tables

WAVE 1

Descriptive Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
hhsize	2850	7.539	3.877	1	33
hhMeanAge	2848	23.579	12.081	7.5	92
hhMeanEduYrs	2183	4.591	2.719	0	14
hhhSex	2817	.692	.462	0	1
hhhMarital	2817	.936	.674	0	2
hhhAge	2817	47.071	15.206	17	101
hhhHighestEdu	2244	1.795	1.016	0	5
hhhEduYrs	2244	7.337	4.099	0	17
depRatio	2850	.561	.219	0	1
region	2850	2.389	1.122	1	4
urban	2850	.205	.404	0	1
elec	2833	.132	.338	0	1
room count	2834	2.975	1.683	0	12
handwash toilet	2832	.154	.459	0	2
logged water	2819	4.094	.708	0	8.7
meals day	2831	2.475	.734	0	14
ypce	2820	667002.17	11285193	11927.621	5.782e+08
ypceLn	2820	12.572	.782	9.387	20.175
poor	2850	.247	.432	0	1

WAVE 2

Descriptive Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
hhsize	2714	7.067	3.585	1	29
hhMeanAge	2713	22.948	11.791	8.4	91
hhMeanEduYrs	2415	7.177	3.9	0	17
hhhSex	2706	.695	.46	0	1
hhhMarital	2705	.928	.678	0	2
hhhAge	2706	46.956	15.235	15	100
hhhHighestEdu	2151	1.83	1.042	0	5
hhhEduYrs	2150	7.38	4.048	0	17
depRatio	2714	.539	.221	0	1
region	2716	2.272	1.218	0	4
urban	2716	.225	.418	0	1
elec	2636	.121	.326	0	1
room count	2637	2.949	1.767	0	19
handwash toilet	2632	.111	.387	0	2
logged water	2617	4.055	.762	0	9.21
meals day	2632	2.498	.664	0	5
ypce	2655	432182.27	606645.09	13516.509	16815060
ypceLn	2655	12.605	.802	9.512	16.638
poor	2714	.244	.43	0	1

Pair-wise Correlation

Pairwise correlations

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(1) hhsize	1.000						
(2) hhMeanAge	-0.316	1.000					
	(0.000)						
(3) hhMeanEduYrs	0.012	0.058	1.000				
	(0.399)	(0.000)					
(4) depRatio	0.249	-0.065	-0.320	1.000			
	(0.000)	(0.000)	(0.000)				
(5) ypce	0.003	0.000	0.009	-0.029	1.000		
	(0.848)	(0.991)	(0.532)	(0.033)			
(6) ypceLn	-0.199	0.071	0.444	-0.415	0.205	1.000	
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)		
(7) poor	0.119	-0.016	-0.241	0.243	-0.031	-0.679	1.000
	(0.000)	(0.221)	(0.000)	(0.000)	(0.023)	(0.000)	

Pairwise correlations

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
(1) region	1.000									
(2) urban	-0.198	1.000								
	(0.000)									
(3) elec	-0.313	0.424	1.000							
	(0.000)	(0.000)								
(4) room_count	0.039	-0.013	0.076	1.000						
	(0.004)	(0.348)	(0.000)							
(5) handwash_toilet	-0.136	0.163	0.262	0.196	1.000					
	(0.000)	(0.000)	(0.000)	(0.000)						
(6) logged_water	-0.039	0.083	0.070	0.289	0.136	1.000				
	(0.004)	(0.000)	(0.000)	(0.000)	(0.000)					
(7) meals_day	-0.061	0.168	0.197	0.149	0.126	0.174	1.000			
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)				
(8) ypce	-0.023	0.011	0.023	-0.002	0.012	-0.012	0.019	1.000		
	(0.091)	(0.428)	(0.092)	(0.904)	(0.388)	(0.386)	(0.160)			
(9) ypceLn	-0.260	0.340	0.479	0.152	0.257	0.064	0.315	0.205	1.000	
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)		
(10) poor	0.132	-0.170	-0.209	-0.126	-0.128	-0.087	-0.258	-0.031	-0.679	1.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.023)	(0.000)	

Pairwise correlations

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(1) hhhSex	1.000							
(2) hhhMarital	0.387	1.000						
	(0.000)							
(3) hhhAge	-0.153	-0.145	1.000					
	(0.000)	(0.000)						
(4) hhhHighestEdu	0.120	0.022	-0.077	1.000				
	(0.000)	(0.140)	(0.000)					
(5) hhhEduYrs	0.149	0.041	-0.159	0.843	1.000			
	(0.000)	(0.006)	(0.000)	(0.000)				
(6) ypce	0.013	-0.008	-0.003	0.011	0.013	1.000		
	(0.327)	(0.546)	(0.834)	(0.462)	(0.374)			
(7) ypceLn	0.075	-0.037	-0.137	0.409	0.388	0.205	1.000	
	(0.000)	(0.006)	(0.000)	(0.000)	(0.000)	(0.000)		
(8) poor	-0.084	-0.010	0.093	-0.220	-0.224	-0.031	-0.679	1.000
	(0.000)	(0.478)	(0.000)	(0.000)	(0.000)	(0.023)	(0.000)	

- Variables that are highly correlated with consumption expenditure (Logged annual consumption per capita):
  - hhsize, hhMeanEduYrs,depRatio, hhhAge, hhhHighest~u, hhhEduYrs, region, urban, elec, room_count, handwash_t~t, meals_day
- Independent variables that are highly correlated with each other:
  - hhhEduYrs and hhhHighest~u
- Independent variables that aare moderately related with each other:
  - hhsize and hhMeanAge
  - hhhMarital and hhhSex
  - elec and region

Machine Learning Model Selection

Stepwise

	(1) Full_model	(2) SW_B_ar2	(3) SW_B_aic	(4) SW_B_ar2	(5) SW_F_aic
Household size	-0.0459***	-0.0459***	-0.0457***	-0.0457***	-0.0457***
Mean age of HH mem~s	0.00529***	0.00529***	0.00525***	0.00525***	0.00525***
Mean years of educ~H	0.0185***	0.0185***	0.0186***	0.0186***	0.0186***
Dependency ratio	-0.770***	-0.770***	-0.766***	-0.766***	-0.766***
Sex of HH head	0.0385	0.0385	0.0470*	0.0470*	0.0470*
What is the presen~t	0.0161	0.0161			
Age of HH head	-0.00394***	-0.00394***	-0.00398***	-0.00398***	-0.00398***
Highest education ~H	0.0523**	0.0523**	0.0520**	0.0520**	0.0520**
Years of education~d	0.00905*	0.00905*	0.00907*	0.00907*	0.00907*
Region of Residen~10	-0.0835***	-0.0835***	-0.0829***	-0.0829***	-0.0829***
Urban/Rural Identi~r	0.173***	0.173***	0.171***	0.171***	0.171***
Does this house ha~y	0.470***	0.470***	0.470***	0.470***	0.470***
How many rooms doe~o	0.0762***	0.0762***	0.0763***	0.0763***	0.0763***
Do you have a hand~l	0.148***	0.148***	0.148***	0.148***	0.148***
How many litres of~r	0.0904***	0.0904***	0.0920***	0.0920***	0.0920***
How many meals, in~f	0.209***	0.209***	0.209***	0.209***	0.209***
Constant	12.08***	12.08***	12.08***	12.08***	12.08***
R-squared	0.502	0.502	0.502	0.502	0.502
Adjusted R-squared	0.500	0.500	0.500	0.500	0.500
Number of observat~s	3824	3824	3824	3824	3824
Number of covariates	16	16	15	15	15

* p<0.05, ** p<0.01, *** p<0.001

Stepwise backward using AIC, forward using adjusted R-squared and forward using AIC **all suggest to remove "hhhMarital"**

Lasso

```
. eststo lasso: reg ypceLn $lassoVars
```

Source	SS	df	MS		
Model	1086.20615	16	67.8878842	Number of obs	= 3,824
Residual	1077.87105	3,807	.283128722	F(16, 3807)	= 239.78
				Prob > F	= 0.0000
				R-squared	= 0.5019
				Adj R-squared	= 0.4998
Total	2164.07719	3,823	.566067798	Root MSE	= .5321

ypceLn	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
hhsize	-.0459025	.0032387	-14.17	0.000	-.0522524	-.0395527
hhMeanAge	.0052909	.0013879	3.81	0.000	.0025698	.0080121
hhMeanEduYrs	.0185431	.0031515	5.88	0.000	.0123644	.0247218
depRatio	-.7695984	.0506705	-15.19	0.000	-.8689423	-.6702546
hhhSex	.0385485	.0222023	1.74	0.083	-.0049811	.082078
hhhMarital	.0160997	.0150532	1.07	0.285	-.0134133	.0456127
hhhAge	-.0039421	.0009492	-4.15	0.000	-.005803	-.0020812
hhhHighestEdu	.0522631	.0163288	3.20	0.001	.020249	.0842772
hhhEduYrs	.0090527	.004108	2.20	0.028	.0009986	.0171067
region	-.0834768	.0079997	-10.43	0.000	-.099161	-.0677926
urban	.1725154	.0240199	7.18	0.000	.1254223	.2196084
elec	.4699676	.0315398	14.90	0.000	.4081312	.5318041
room_count	.0762384	.0058861	12.95	0.000	.0646981	.0877786
handwash_toilet	.1482957	.0204881	7.24	0.000	.108127	.1884645
logged_water	.0903533	.0141767	6.37	0.000	.0625587	.118148
meals_day	.2085508	.013983	14.91	0.000	.181136	.2359657
_cons	12.0755	.0801663	150.63	0.000	11.91832	12.23267

Lasso result suggested that **nothing needs to be removed** from the model

VIF for Multicollinearity

. vif, uncentered // variables

Variable	VIF	1/VIF
_cons	86.80	0.011521
logged_water	48.77	0.020505
hhhAge	27.56	0.036285
meals_day	18.27	0.054720
hhhEduYrs	16.01	0.062444
hhhHighest~u	15.65	0.063911
hhMeanAge	14.20	0.070416
depRatio	11.93	0.083852
hhsize	10.62	0.094202
hhMeanEduYrs	6.74	0.148458
room_count	6.10	0.164033
region	5.88	0.170069
hhhSex	5.01	0.199710
hhhMarital	4.27	0.234424
elec	1.75	0.571528
urban	1.73	0.578691
handwash_t~t	1.26	0.795396
Mean VIF	16.62	

>10

. vif, uncentered // variables

Variable	VIF	1/VIF
_cons	86.80	0.011521
~~logged_water~~	48.77	0.020505
hhhAge	27.56	0.036285
meals_day	18.27	0.054720
~~hhhEduYrs~~	16.01	0.062444
hhhHighest~u	15.65	0.063911
~~hhMeanAge~~	14.20	0.070416
depRatio	11.93	0.083852
hhsize	10.62	0.094202
hhMeanEduYrs	6.74	0.148458
room_count	6.10	0.164033
region	5.88	0.170069
hhhSex	5.01	0.199710
hhhMarital	4.27	0.234424
elec	1.75	0.571528
urban	1.73	0.578691
handwash_t~t	1.26	0.795396
Mean VIF	16.62	

Finalized Variables

HH characteristics

hhsize hhMeanEduYrs depRatio region urban elec room_count handwash_t~t meals_day

HHH characteristics

hhhSex hhhAge hhhHighest~u

Quadratic terms

hhsize2 = hhsize^2

hhMeanEduYrs2 = hhMeanEduYrs^2

hhhAge2 = hhhAge^2

Eliminated

hhhEduYrs logged_water hhMeanAge hhhMarital

Regression Analysis

Pooled OLS

- The coefficients on most of our independent variables **are statistically significant** at a **95%** confidence level.

Interpretation of important variables:

- **Household size:**
  - [hhsize] A 1-unit increase in household size is, on average, associated with a 10.8% decrease in per capita expenditure ($p < 0.01$), suggesting that **larger households might have lower per capita consumption**
  - [hhsize2] The negative impact of household size diminishes at higher values -> **the marginal impact of adding another household member becomes less negative for larger households**
- **Household Head's Education:**
  - [hhhHighestEdu] Compared to the baseline category (No formal education), **higher levels of education are significantly associated with higher per capita expenditure**
  - **Completed University shows the largest effect** (0.589, $p < 0.01$), meaning households with university-educated heads have 58.9% higher per capita expenditure compared to the uneducated group
- **Region:**
  - [region] Household consumption is significantly lower in Eastern (-43.3%), Northern (-37.1%), and Western (-37.4%) regions ($p < 0.01$), while the Central region shows a slight, non-significant decline (-0.099, $p = 0.053$), indicating **Kampala has the highest per capita expenditure**

```
. eststo ols: reg ypceLn $chosenVars, robust

Linear regression                               Number of obs     =      3,840
                                                F(24, 3815)       =     163.94
                                                Prob > F          =     0.0000
                                                R-squared         =     0.5263
                                                Root MSE          =       .522
```

ypceLn	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
hhsize	-.1082314	.0084852	-12.76	0.000	-.1248674	-.0915955
hhsize2	.0032185	.0003681	8.74	0.000	.0024969	.0039401
hhMeanEduYrs	.0153727	.0089862	1.71	0.087	-.0022456	.032991
hhMeanEduYrs2	.0008407	.0005602	1.50	0.133	-.0002575	.001939
depRatio	-.6915157	.0534802	-12.93	0.000	-.7963682	-.5866631
hhhSex	.0910173	.0203993	4.46	0.000	.0510227	.1310118
hhhAge	.0031387	.0040996	0.77	0.444	-.004899	.0111764
hhhAge2	-.0000429	.0000402	-1.07	0.287	-.0001217	.000036
hhhHighestEdu						
Less than primary	.1908588	.0800231	2.39	0.017	.0339666	.347751
Completed primary	.2588802	.0821224	3.15	0.002	.0978723	.4198882
Completed O-level	.2981048	.0889548	3.35	0.001	.1237012	.4725083
Completed A-level	.3689073	.0895147	4.12	0.000	.1934059	.5444086
Completed University	.5894282	.1039	5.67	0.000	.3857233	.7931332
region						
Central	-.099664	.0515011	-1.94	0.053	-.2006362	.0013083
Eastern	-.4325959	.0533992	-8.10	0.000	-.5372897	-.3279022
Northern	-.37068	.0527941	-7.02	0.000	-.4741874	-.2671725
Western	-.3739826	.0538503	-6.94	0.000	-.4795608	-.2684045
urban	.1602458	.0254155	6.31	0.000	.1104164	.2100751
elec	.3925232	.0318967	12.31	0.000	.3299869	.4550595
room_count	.0700362	.0066653	10.51	0.000	.0569684	.083104
handwash_toilet						
Yes with water only	.1278687	.0320778	3.99	0.000	.0649775	.1907599
Yes with water and soup	.3039058	.0479915	6.33	0.000	.2098144	.3979973
meals_day	.2273649	.0243191	9.35	0.000	.1796853	.2750446
wave	.0892028	.0197424	4.52	0.000	.0504962	.1279094
_cons	12.38246	.1454655	85.12	0.000	12.09727	12.66766

Regression Analysis

Fixed Effect

- A lot of coefficients **lost their significance** from Pooled OLS to FE + some variables are **omitted** + **R-squared decreased**
  - FE models only use **within-household variation**. If a variable's effect is largely driven by differences between households, rather than changes within a household over time, it will lose significance in the FE model (e.g., hhsize, hhhSex, hhhAge, etc.)
  - Many **time-invariant factors** (e.g., regional dummies) are absorbed by the fixed effects
- FE model is **better for causal inference** because they remove **unobserved** household-level cofounders + OLS may **overestimate significance** by not accounting for unobserved heterogeneity

Interpretation of important variables:

- **HH mean education years**:
  - [hhMeanEduYrs] The negative but non-significant coefficient on hhMeanEduYrs suggests that additional years of education may initially be associated with lower earnings.
  - [hhMeanEduYrs2] The positive and significant coefficient suggests that **as education levels increase, the benefits of education on per capita expenditure become stronger**
- **HH Dependency ratio**:
  - [depRatio] an increase in the dependency ratio is associated with a strong decline in log per capita expenditure
- **Food security**:
  - [meals_day] an increases in the number of meals per day within a household is strongly associated with higher per capita expenditure

```
. eststo fe: xtreg ypceLn $chosenVars, robust fe i(HHID)
note: 2.region omitted because of collinearity.
note: 3.region omitted because of collinearity.
note: 4.region omitted because of collinearity.

Fixed-effects (within) regression               Number of obs      =         3,840
Group variable: HHID                            Number of groups   =         2,408

R-squared:                                      Obs per group:
     Within  = 0.0794                                         min =             1
     Between = 0.3678                                         avg =           1.6
     Overall = 0.3271                                         max =             2

                                                F(21, 2407)       =          6.61
corr(u_i, Xb) = 0.2663                          Prob > F          =        0.0000

                                    (Std. err. adjusted for 2,408 clusters in HHID)
------------------------------------------------------------------------------
             |               Robust
      ypceLn | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       hhsize |  -.0292955   .0430877    -0.68   0.497    -.1137884    .0551973
      hhsize2 |  -.0000879   .0011594    -0.08   0.940    -.0023613    .0021856
   hhMeanEduYrs |  -.0157712    .010241    -1.54   0.124    -.0358533    .0043108
  hhMeanEduYrs2 |   .0017627   .0005954     2.96   0.003     .0005952    .0029302
       depRatio |  -.6746691   .1389538    -4.86   0.000    -.9471506   -.4021876
         hhhSex |   .0168221   .0892467     0.19   0.851    -.1581863    .1918304
         hhhAge |   .0305718   .0276746     1.10   0.269    -.0236967    .0848403
        hhhAge2 |  -.0002914   .0002596    -1.12   0.262    -.0008004    .0002176
                |
      hhhHighestEdu |
  Less than primary |   .2488849   .1178713     2.11   0.035     .0177451    .4800247
  Completed primary |   .2446682   .1251978     1.95   0.051    -.0008383    .4901748
  Completed O-level |   .3185641   .1414425     2.25   0.024     .0412024    .5959258
  Completed A-level |    .337227   .1379581     2.44   0.015     .0666981    .6077559
Completed University |   .1731626   .1720242     1.01   0.314    -.1641683    .5104934
                |
             region |
            Central |  -.0835402   .0732033    -1.14   0.254    -.2270883    .0600078
            Eastern |          0  (omitted)
           Northern |          0  (omitted)
            Western |          0  (omitted)
                |
              urban |   .1349049   .1248544     1.08   0.280    -.1099282    .3797381
               elec |   .1168535   .0861647     1.36   0.175    -.0521112    .2858181
         room_count |   .0276695   .0124364     2.22   0.026     .0032824    .0520566
                |
      handwash_toilet |
 Yes with water only |   .0393565   .0408939     0.96   0.336    -.0408345    .1195474
Yes with water and soap |    .028148   .0885993     0.32   0.751    -.1455907    .2018867
                |
         meals_day |   .0907619   .0217813     4.17   0.000     .0480498    .1334739
              wave |   .0130858   .0221549     0.59   0.555    -.0303588    .0565303
             _cons |   11.86946   .6830274    17.38   0.000     10.53008    13.20884
-------------+----------------------------------------------------------------
           sigma_u |  .61098438
           sigma_e |  .4127288
               rho |  .68666235   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Regression Analysis

Random Effect

- **Increased R-squared** compared to FE + coefficients' magnitudes very **similar to Pooled OLS** + **no omitted** variables
  - FE model only explains variation within households, while the RE model considers **both within-household and between-household variations**.
  - RE does not eliminate time-invariant variables (e.g., region) because it assumes **household-specific effects are uncorrelated** with explanatory variables
- RE's **assumption may be violated** because it is almost impossible for household-specific effects to be uncorrelated with explanatory variables. (e.g., a strong preference for saving - often unobserved, may be both related to consumption and explanatory variables like dependency ratio)

Hausman Test

P-value = 0.0000 (< 0.05) → Reject the null hypothesis → **FE is preferred** because there is correlation between the individual effects and regressors

```
   chi2(20) = (b-B)'[(V_b-V_B)^(-1)](b-B)
            = 155.06
Prob > chi2 = 0.0000
```

```
. ******************************************************************
. * Random-Effects Model
. ******************************************************************
.         eststo re: xtreg ypceLn $chosenVars, robust re i(HHID)

Random-effects GLS regression                   Number of obs     =      3,840
Group variable: HHID                            Number of groups  =      2,408

R-squared:                                      Obs per group:
     Within  = 0.0593                                         min =          1
     Between = 0.5931                                         avg =        1.6
     Overall = 0.5250                                         max =          2

                                                Wald chi2(24)     =    3217.02
corr(u_i, X) = 0 (assumed)                      Prob > chi2       =    0.0000

                                    (Std. err. adjusted for 2,408 clusters in HHID)
```

ypceLn	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
hhsize	-.1022406	.008867	-11.53	0.000	-.1196196	-.0848615
hhsize2	.0029804	.0003821	7.80	0.000	.0022314	.0037293
hhMeanEduYrs	.005812	.0083529	0.70	0.487	-.0105594	.0221835
hhMeanEduYrs2	.001133	.0005133	2.21	0.027	.0001271	.002139
depRatio	-.7254262	.0558121	-13.00	0.000	-.8348159	-.6160366
hhhSex	.081138	.0219325	3.70	0.000	.0381511	.1241249
hhhAge	.0037484	.0044246	0.85	0.397	-.0049237	.0124206
hhhAge2	-.0000474	.0000436	-1.09	0.277	-.0001329	.0000381
hhhHighestEdu						
Less than primary	.2412013	.0735642	3.28	0.001	.0970182	.3853844
Completed primary	.3186143	.0759823	4.19	0.000	.1696918	.4675369
Completed O-level	.3703792	.0834948	4.44	0.000	.2067324	.534026
Completed A-level	.4532421	.0834521	5.43	0.000	.289679	.6168052
Completed University	.646509	.1017386	6.35	0.000	.4471049	.845913
region						
Central	-.1082478	.0491132	-2.20	0.028	-.2045078	-.0119878
Eastern	-.4478605	.0524324	-8.54	0.000	-.5506262	-.3450948
Northern	-.3949083	.0520351	-7.59	0.000	-.4968952	-.2929213
Western	-.3906785	.0532468	-7.34	0.000	-.4950404	-.2863166
urban	.1690546	.0274605	6.16	0.000	.1152329	.2228763
elec	.3930107	.0338448	11.61	0.000	.3266761	.4593454
room_count	.0663017	.0068359	9.70	0.000	.0529035	.0796998
handwash_toilet						
Yes with water only	.1116018	.0311417	3.58	0.000	.0505652	.1726384
Yes with water and soup	.2544563	.0492462	5.17	0.000	.1579356	.3509771
meals_day	.1978755	.0224851	8.80	0.000	.1538055	.2419454
wave	.0714909	.0173958	4.11	0.000	.0373958	.105586
_cons	12.47654	.1452293	85.91	0.000	12.1919	12.76118
sigma_u	.31648115					
sigma_e	.4127288					
rho	.37027108	(fraction of variance due to u_i)				

OLS Assumptions

Linear in Parameters

- **Def:** OLS assumes the relationships between the dependent variable and the predictors are linear (but in my case, some explanatory variables and the dependent variable are not in a linear relationship).
- **Solutions:**
  - log transformation (mitigates problems with outliers)
  - quadratic function (deals with non-linear relationships)

Random Sampling

- **Def:** Random sampling is necessary for the analysis to have external validity
- **Assessment:**
  - two waves for analysis were carried out on a nationally representative sample of households -> random sampling
  - my analysis cares more about internal validity, instead of the generalization of this study to other population
  - the sample doesn't have missing data that will cause endogenous sample selection; I've removed outliers from mistakes

No perfect collinearity

- **Def:** No independent variable in the model is an exact linear combination of other independent variables
- **Assessment:**
  - I set a baseline for each categorical variable to avoid perfect multicollinearity
  - FE automatically dropped variables due to perfect multicollinearity
  - I used pair-wise correlation to check the correlation between independent variables and removed highly correlated ones
  - I checked VIF and removed some with VIF >10

Zero Conditional Mean

- **Def:** The value of the explanatory variables must contain no information about the mean of the unobserved factors (exogeneity)
- **Solutions:**
  - quadratic function and log transformation to ensure we capture true relationships
  - I used FE to account for any observed and unobserved time-invariant variables
- **Limitation:**
  - there can still be other variables that are omitted from the regression and correlated with both explanatory and dependent variables

Homoskedasticity

- **Def:** The variation of the error term,$\sigma\varepsilon 2$, is the same for all Xi.
- **Solution:**
  - compute robust standard errors

Normality of error terms

- **Def:**
  - Unobserved factors are normally distributed around the population regression function
  - The form and the variance of the distribution does not depend on any of the explanatory variables.
- **Solution:**
  - I used "kdensity" and "qnorm" to check the normality of the residuals.
  - generally not a concern in larger sample sizes

Conclusion

Main findings

FE makes more sense among all model I tried -> **findings based on FE result**:

- **Dependency ratio, food security** (measured by meals per day), and **housing space** (measured by room counts) are **strong predictors** of consumption.
  - A higher dependency ratio is associated with lower consumption
  - Better food security and larger housing space correlate with higher consumption.
- **Mean years of education** of a household are correlated with consumption.
  - Early education attainment may initially lead to lower consumption due to delayed entry into the job market
  - **Higher education levels yield greater long-term returns**.
- **Kampala** is the most economically advantaged region.

Policy Implication

- **Support households with high dependency ratios:**
  - Social assistance, targeted cash transfers, or child care subsidies
- **Improving food security:**
  - Investment in agricultural productivity, food assistance programs, and price stabilization policies
- **Affordable housing initiatives:**
  - Affordable housing, financial support for homeownership, and rental assistance
- **Education policy initiatives:**
  - Scholarships program, vocational training, and financial aid programs
- **Regional economic development:**
  - Regional investment strategies, infrastructure development, and business incentives to promote economic growth in less developed areas.

Limitation

- **Omitted variable bias:**
  - There can still be other variables that are omitted from the FE regression, correlated with both explanatory and dependent variables and time-variant
  - I didn't controlled for time-fixed effects that can take into account of those time-variant variables but fixed across households. (Including variable wave only assumed a linear time trend)
- **Elimination of time-invariant characteristics:**
  - Variable that does not change over time within a entity cannot be estimated
- **Small within-group variation:**
  - If within-group variation is small, the estimates become less precise
- **Low external validity**