11-1-2008

How Does the Persons per Household Variable Affect Population Estimation and How to Measure It

Xiaomin Ruan

How does the Persons per Household Affect Population Estimation and how to Measure it

Xiaomin Ruan

THE UNIVERSITY of NEW MEXICO

University of New Mexico
BBER Population Estimation & Projection Program
For the Data User Conference, Nov. 2008

Outlines

- Concepts related to persons per household (PPH)
- Effects of PPH in housing unit method
- Components that affect PPH
- PPH estimation model selection and validation
- Exploration of forecasting PPH

Concepts Related to Persons per Household

- **Persons per Household (PPH)**

  — the number of persons in one household. Also named as Household Size by Census Bureau.

  PPH = Total Population / # Households (or # Housing Units)

- **Household**

  — A household includes all persons who **occupy** a housing unit.

- **Housing Unit (HU)**

  — A housing unit is a house, an apartment, a mobile home, a group of rooms, or a single room that is **occupied** as a separate living quarters.

PPH in Population Estimation

Last Year Housing Unit Stock + New Housing Units - Demolished Housing units = Current Housing Units

Current Housing Units X Occupancy Rate X Average Persons per Occupied Housing Units = Persons in Housing Units

HU OR PPH

+

Estimated Current Group Quarter Persons Persons in Facilities

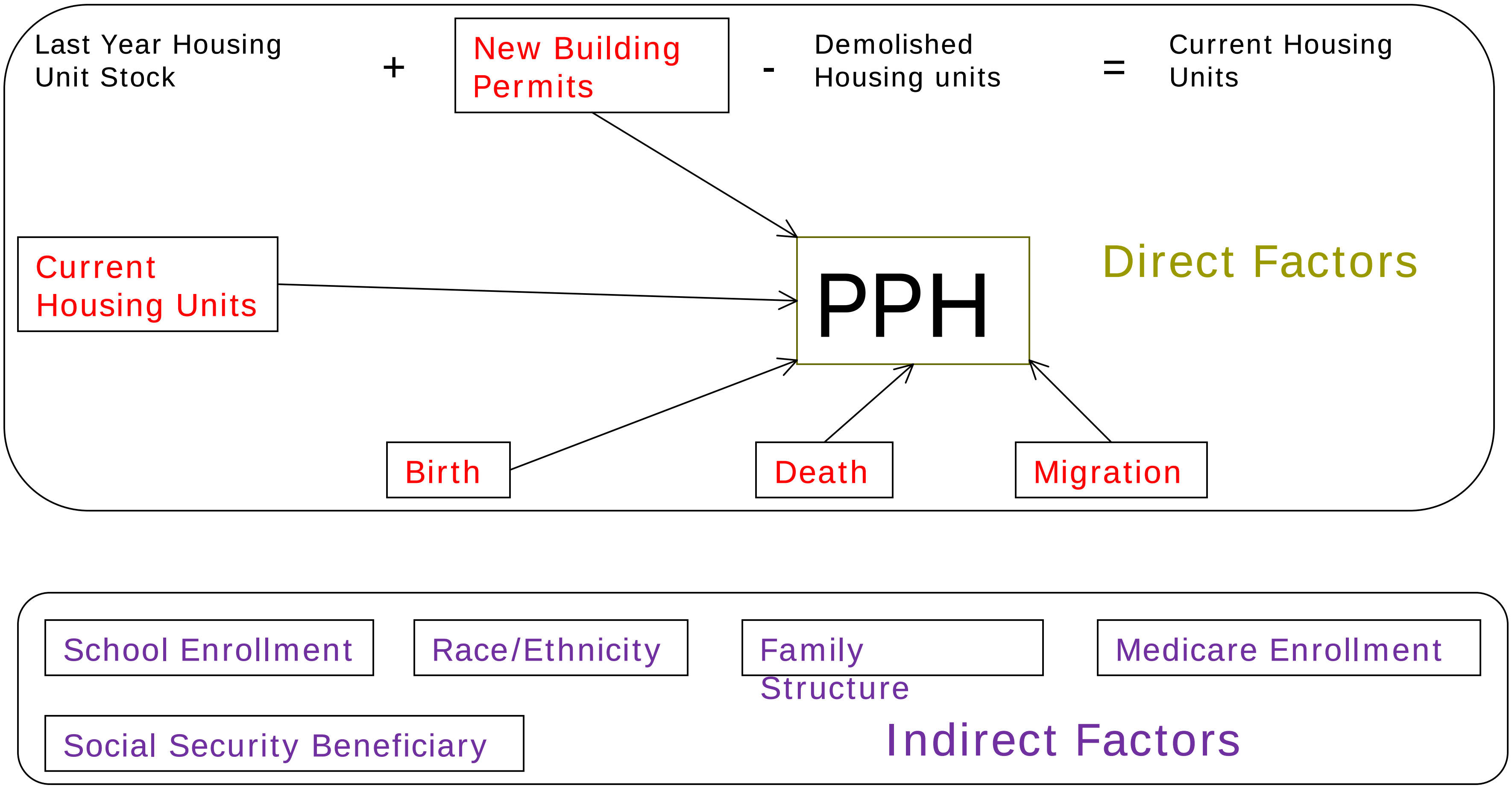
=

Total Persons

Sensitivity of Population Estimation to PPH

Geographic Level		Census 2000	PPH Variation			
			-1%	-2%	-5%	-10%
Tract	PPH	2.34	2.32	2.27	2.16	1.94
	Occupied HU	1,183	-	-	-	-
	Population	2,771	2,743	2,688	2,554	2,299
	Absolute Diff.	-	-28	**-83**	-217	-472
County	PPH	2.52	2.49	2.47	2.39	2.27
	Occupied HU	220,936	-	-	-	-
	Population	556,678	551,111	545,544	528,844	501,010
	Absolute Diff.	-	-5,567	**-11,134**	-27,834	-55,668
State	PPH	2.68	2.66	2.63	2.55	2.41
	Occupied HU	677,971	-	-	-	-
	Population	1,819,046	1,800,856	1,782,665	1,728,094	1,637,141
	Absolute Diff.	-	-18,190	**-36,381**	-90,952	-181,905

Components that affect PPH



Available Source Data

Data Source	Time Span	Note
ACS	2002-2007	Only MSA available
OASDI	1984-2006	Cibola missed 84-87 data, while Catron and Harding have imcomplete 86 and 87 data.
Medicare	1998-2006	Complete
Birth	1990-2006	Complete
Death	1990-2006	Complete
IRS Migration	1981-2007	Missed year 82, 83, 90, 91, 92 data
School Enrollment	1986-2006	Complete grade 1 to 12
Building Permit	2000-2006	Complete
Employment	1995-2006	Complete
Race from Birth	1990-2005	Complete
Race from Death	1990-2005	Complete
Race from Sch_Enroll	1989-2004	Complete

Regression Method

- Regression method uses the variation in independent variables to explain the variation in PPH

- Once the regression model is validated, the independent variable values can be plugged in inter decennial years to estimate the missing PPH and construct a PPH time series to support population estimation.

- Structural test (Chow Test) is critical to the regression model. Once the test shows the variance of residuals of Year 2000's PPH is indifferent from that of 1990, the estimated coefficients can then be applied to other years. Combining with validating process, the structural test can safely demonstrate the predicting power of the chosen model.

Stepwise Selection using SAS

- Including too many explanatory variables will decrease the degree of freedom, but including fewer important variables may reduce the explanatory power of the model. So, a stepwise selection combining forward selection and backward elimination is applied.
- The rule of forward selection is, adding variables to the model once at a time until no significant variable can be found to increase R-square of the previous model (Default significant level for individual variable is 50%).
- The rule of backward elimination is, deleting unimportant variables (decrease R-square the least or has the highest p-value) from the full model until any variable left is significant at default level (10%).
- Stepwise regression combines features of forward selection and backward elimination, which means, every time we add a variable to the model, we ask whether any of the variables added earlier can be omitted. (Default individual variable's significance level is 15%)

Data Transformation

- Knowing that PPH is a population count average by housing units, taking log of the count variables may reduce the model variation and increase model stability.

- Moreover, since PPH is a count variable weighted by housing units, weighting independent variables also by housing units may also increase the explanatory power of regression models.

- Weighting variables by housing units may cause another issue. Since housing unit stock data for non-census years are estimates by BBER, it may be inconvenient to use by outside data users, comparing to the access of administrative data.

Stepwise Selected Models

	Count Model	Log Model	HU Weighted Model
Dep. Var.	PPH	PPH	PPH
Ind. Var.	AI	-	HUW_AI
	-	Ln_OASDI_rw	HUW_OASDI_rw
	-	Ln_Birth	HUW_Birth
	-	Ln_G7_9	HUW_G10_12
	-	Ln_EmpEND	HUW_EmpAVE
	-	Ln_HispSE	HUW_HispSE

F-Test, AIC/BIC screening

Model	Count Model	Log Model	HU Weighted Model
Pseudo Chow Test	-13.15	0.85	3.06**
AIC Test	-2.97	-4.38	-4.87
BIC Test	-79.47	-172.32	-204.22

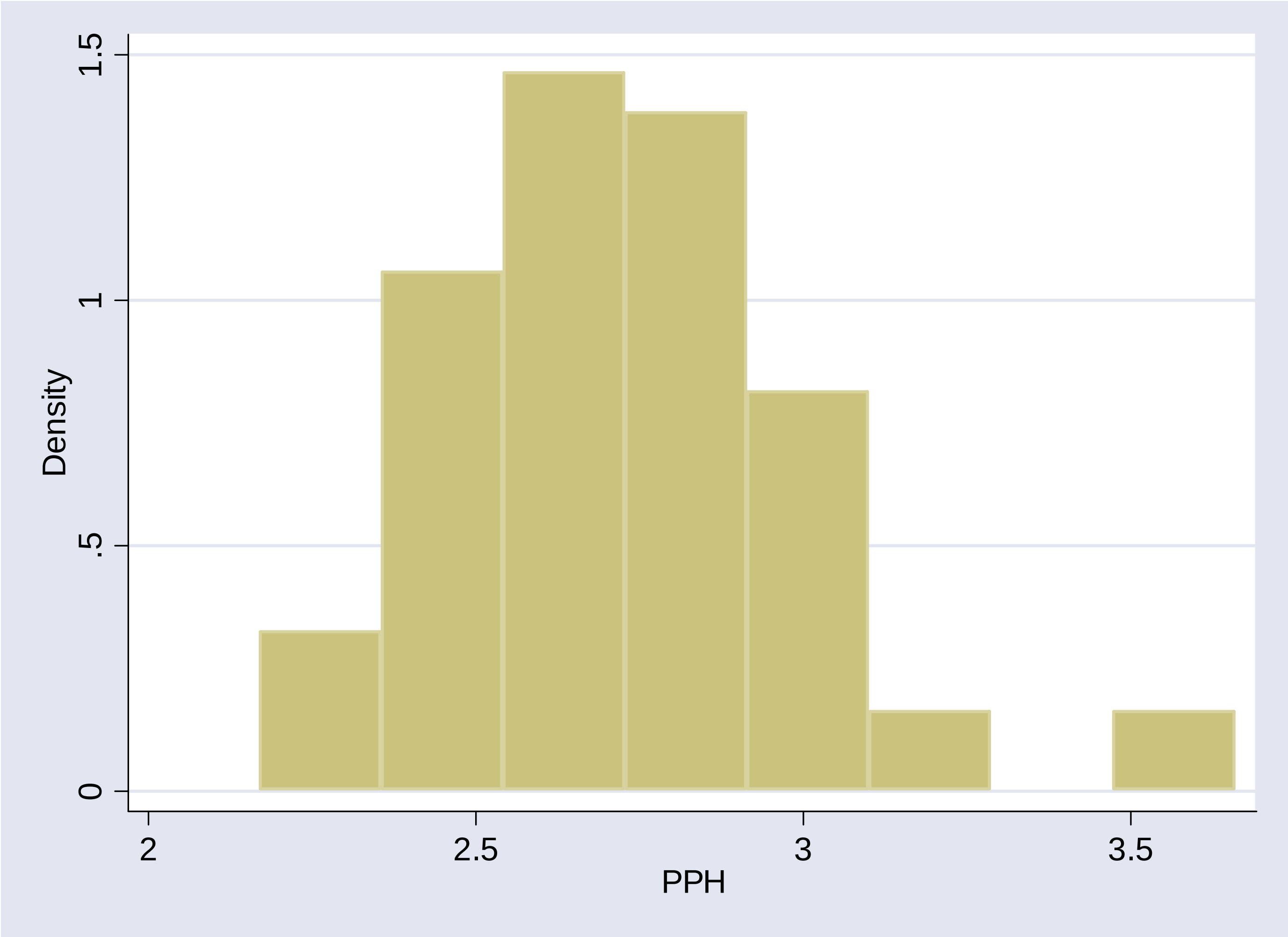
$$AIC = 2k + n[\ln(RSS/n) + 1]$$

$$BIC = n * \ln(RSS/n) + k * \ln(n)$$

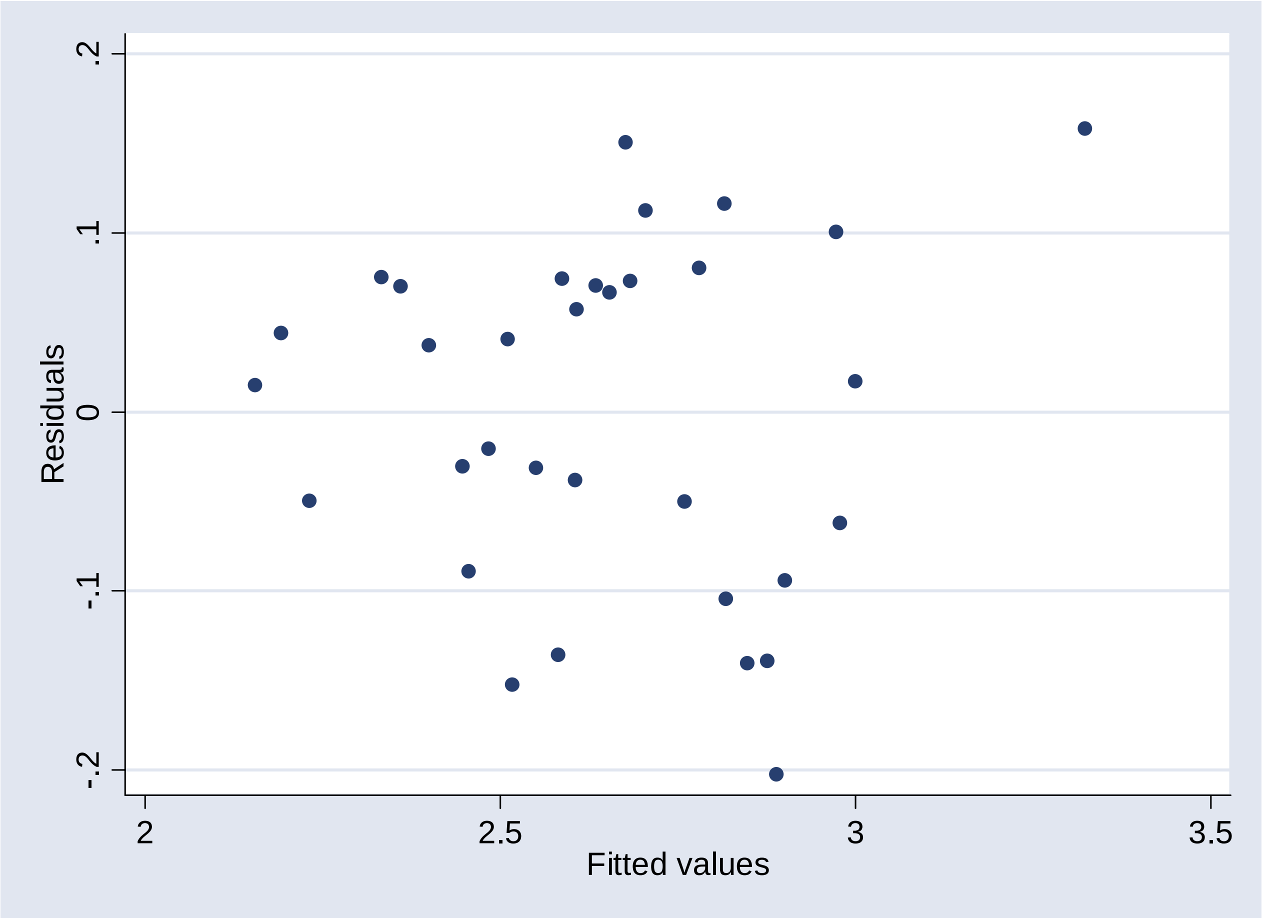
Descriptive Statistics

	Var. Name	Obs	Mean	S.D.
Dep. Var.	PPH	33	2.65	0.28
Log Model	Ln_OASDI_rw	33	7.81	1.19
	Ln_Birth	33	5.72	1.58
	Ln_G7_9	33	6.97	1.37
	Ln_EmpEND	33	8.86	1.54
	Ln_HispSE	33	7.60	1.43
HU Weighted Model	HUW_Indian	33	0.04	0.09
	HUW_OASDI_rw	33	0.22	0.05
	HUW_Birth	33	0.03	0.01
	HUW_G10_12	33	0.09	0.03
	HUW_EmpAVE	33	0.63	0.23
	HUW_HispSE	33	0.20	0.10

PPH Histogram

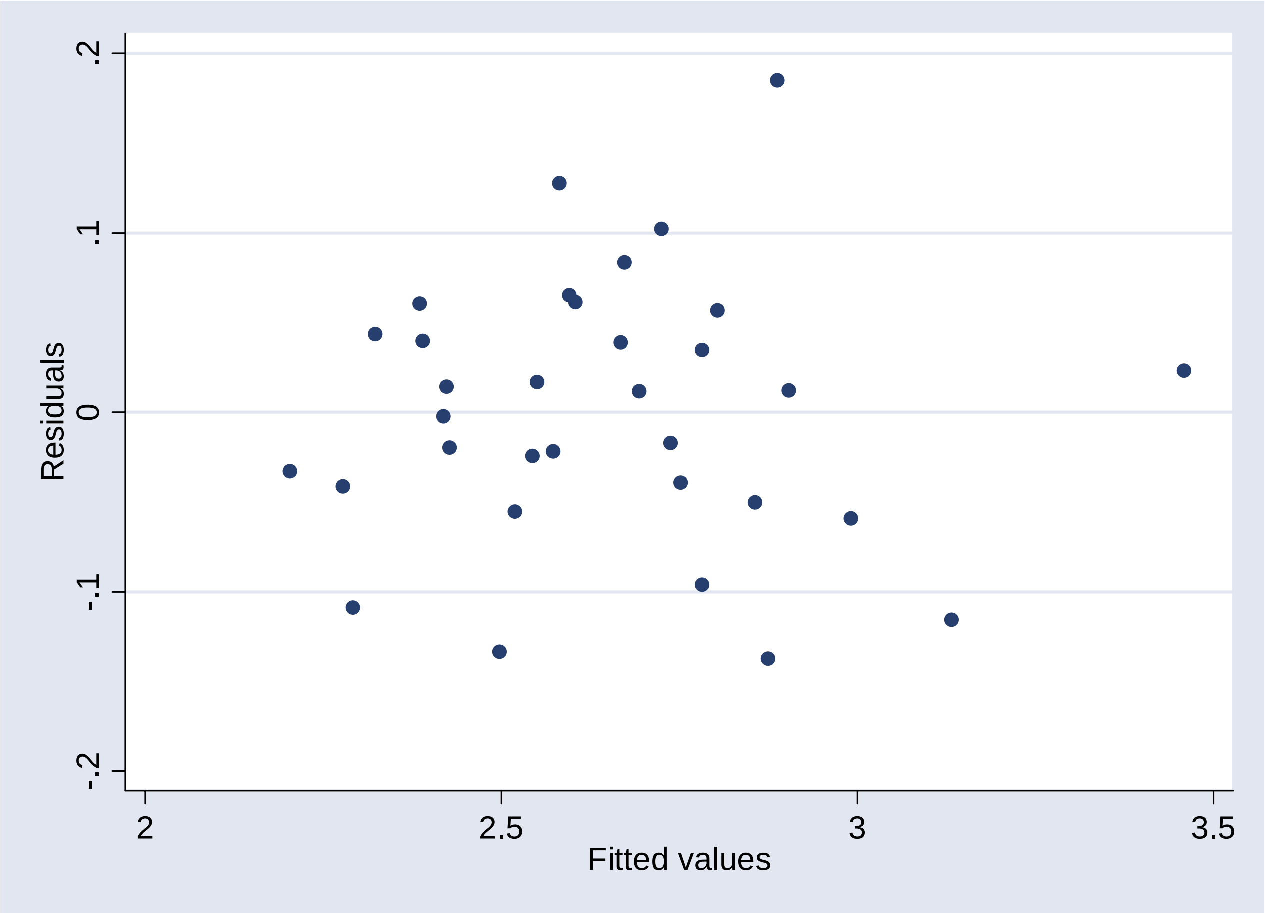


Heteroscedasticity Check Model 1



Log Model Residual Plot

Heteroscedasticity Check Model 2



HU Weighted Model Residual Plot

Estimated Equation

Stepwise Selected Log Model

$$PPH = \underset{(0.257)^{***}}{4.26386} - \underset{(0.070)^{***}}{0.36158}*Ln\_OASDI\_rw + \underset{(0.084)^{***}}{0.54668}*Ln\_Birth + \underset{(0.085)^{**}}{0.19909}*Ln\_G7\_9$$

$$- \underset{(0.083)^{***}}{0.28872}*Ln\_EmpEND - \underset{(0.039)^{**}}{0.09882}*Ln\_HispSE$$

$R^2 = 0.8795$, $F(5,27) = 39.40^{***}$

Stepwise Selected HU Weighted Model

$$PPH = \underset{(0.098)^{***}}{2.25202} - \underset{(0.336)^{**}}{0.75932}*HUW\_OASDI\_rw + \underset{(2.409)^{***}}{15.49605}*HUW\_Birth + \underset{(0.644)^{**}}{1.52901}*HUW\_G10\_12$$

$$- \underset{(0.080)^{***}}{0.22723}*HUW\_EmpAVE + \underset{(0.254)^{***}}{1.00188}*HUW\_Indian + \underset{(0.207)^{*}}{0.40881}*HUW\_HispSE$$

$R^2 = 0.9306$, $F(6,26) = 58.15^{***}$

Validation

- The comparison of the fitted values against original data can tell us how good the model fits the data. But it may not reveal the ability to predict other place or other time spot.

- The model's ability to predict depends on both the model specification and the quality of the data in validating years.

- As we have to use ACS PPH estimates instead of other solid administrative records to validate our models, the differences between the predicted PPH and the ACS estimates may not reveal the true distance to the real PPH values.

Validation Continued

	Log Model		HUW Model	
	2000	**2005**	**2000**	**2005**
# County	33	7	33	7
ME	0.0001	0.0512	0.0000	-0.0955
MPE	0.0011	0.0248	0.0008	-0.0267
MAPE	0.0304	0.0809	0.0215	0.0657

Bayesian Basic Concepts for Simulation

- The basic idea behind Bayesian Modeling is a simple statistic idea: conditional probability.
- Conditional Probability Example: Flu test for 151 students

	Flu	No Flu	
Test +	46	15	61
Test -	4	86	90
	50	101	151

	Flu	No Flu	
Test +	0.30	0.10	0.40
Test -	0.03	0.57	0.60
	0.33	0.67	1.00

$$P(T^+/F) = 46/50 = 0.30/0.33 = P(T^+\&F) / P(F)$$
$$= P(F/T^+)P(T^+)/P(F) = (0.3/0.4)*0.4/0.33$$

- Question: If one student was tested positive, what is his probability to have a real flu?

Parallel Bayesian Question in our Case

- If the predictors are estimated like the equation below, with such a specified variance, what would they really look like once we have more than enough observations?

PPH = 4.26386-0.36158*Ln_OASDI_rw+0.54668*Ln_Birth+0.19909*Ln_G7_9

(0.257)*** (0.070)*** (0.084)*** (0.085)**

-0.28872*Ln_EmpEND-0.09882*Ln_HispSE

(0.083)*** (0.039)**

R^2 = 0.8795, F(5,27) = 39.40***

WinBUGS do Bayesian Automatically

- What one needs to do is to specify the model first,
- Input data second, and then
- Specify the prior information and let the program run simulation

The OPEN SOURCE WinBUGS website has a demonstration movie

http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml

First Try of Bayesian using Prior Estimates

Bayesian Simulated Log Model

PPH = 4.267-0.3656*Ln_OASDI_rw+0.5434*Ln_Birth+0.2015*Ln_G7_9

(0.269)** (0.072)** (0.088)** (0.089)**

-0.2874*Ln_EmpEND-0.09627*Ln_HispSE

(0.086)** (0.040)**

Bayesian Simulated HU Weighted Model

PPH = 2.248-0.7529*HUW_OASDI_rw+15.44*HUW_Birth+1.528*HUW_G10_12

(0.102)** (0.352)** (2.512)** (0.678)**

-0.2209*HUW_EmpAVE+0.9972*HUW_Indian+0.4116*HUW_HispSE

(0.084)** (0.268)** (0.218)**

Bayesian Model Validation

	Log Model		HUW Model	
	2000	2005	2000	2005
# County	33	7	33	7
ME	0.0009	0.1199	0.0001	0.0372
MPE	0.0015	0.0496	0.0008	0.0192
MAPE	0.0304	0.0890	0.0215	0.0904

2005 PPH Sensitivity Analysis

County	Fitted 2005 PPH	ACS 2005 PPH	Diff.	Pop Based on Fitted	Pop based on ACS	Pop Gap
Bernalillo	2.55	2.37	0.18	635745	591860	43885
Dona Ana	2.76	2.75	0.01	185340	184377	964
McKinley	3.38	2.59	0.79	66919	51308	15611
Sandoval	2.63	2.75	-0.12	102224	106692	-4468
San Juan	3.01	3.29	-0.28	114344	125106	-10761
Santa Fe	2.35	2.61	-0.26	124332	137805	-13473
Valencia	2.68	2.74	-0.06	65523	67108	-1585
					Total	30173

Future Work for Model Specification

- Look for previous census year data that can be used in regression models
- Compare fitted values to BBER Pop estimates.
- Use only those 7 MSA counties to do a structural change test.
- Use ACS 2006 data as reference instead of 2005 data.
- Try simultaneous equations

Appendix

- SAS code
- WinBUGS Bayesian Modeling Code

Thank you!

Contact Xiaomin Ruan
xmruan@unm.edu
505-277-3541