# Synthetic Data for Small Area Estimation in the U.S. Federal Statistical System


Joseph W. Sakshaug

Institute for Social Research, University of Michigan

Institute for Employment Research, Nürnberg, Germany

Ludwig-Maximilian University of Munich, Germany



UNECE/Eurostat Work Session on Statistical Data Confidentiality

October 27, 2011

# Outline

- Background
- Small Area Estimation Programs in the United States
  - Pros/Cons
- Synthetic Microdata for Small Area Estimation
- Applications
  - American Community Survey
  - National Health Interview Survey
- Conclusions
- Future directions

# Background/Motivation

- Increasing demand for small area estimates (Tranmer et al., 2005)
  - states, counties, cities/towns, neighborhoods, etc.
- Small area effects can impact policy decisions and interventions at local levels
- Microdata for small areas typically not released to the public due to disclosure concerns
- Thus, statistical agencies are responsible for producing the majority of small area information

# Small Area Estimation Programs in the United States

- SAIPE – U.S. Census Bureau
  - County-level estimates on income and poverty rates
- SAHIE – U.S. Census Bureau
  - County-level estimates of health insurance coverage
- National Cancer Institute
  - County-level prevalence of smoking, mammography, pap smear test
  - Combines two surveys (Raghunathan et al., 2007)
    - National Health Interview Survey (NHIS)
    - Behaviorial Risk Factor Surveillance Survey (BRFSS)

# Pros/Cons of Small Area Programs

- Advantages
  - Provides important estimates at the local level
  - Sufficient for basic analytic purposes
  - Often used to inform policy decisions
  - Data confidentiality is maintained
- Disadvantages
  - No microdata available for small areas
  - Customized analyses not feasible
    - Variable recodes, subgroup analysis, alternative definitions of construct of interest
  - Multivariate estimates usually not released
    - Correlations, regression coefficients, etc.

# Microdata for Small Geographic Areas

Two main data dissemination approaches:

1) Release microdata files for areas with > 100,000 residents (U.S. Census Bureau)
    - 626 (out of 3141) counties meet this minimum
    - Other counties are combined with larger counties until threshold is met
2) Access restricted data via Research Data Centers
    - Limited # of Census RDCs in U.S. (13 locations)
    - Proposal and special sworn status required

# Alternative Method: Release Synthetic Microdata (Rubin, 1993)

- Treat unsampled portion of population as missing data
- Replace missing data with imputed (or synthetic) values drawn from a posterior predictive distribution
- Release samples of synthetic data which comprise the public-use microdata
- Apply standard combining rules to obtain valid inferences (Raghunathan, Reiter, and Rubin, 2003)
- Released data need not contain any observed records

# Previous Applications of Synthetic Data

- IAB Establishment Panel (Drechsler et al., 2008)
- SIPP/SSA/IRS (Abowd et al., 2006)
- ACS Group Quarters (Rodriguez, 2007)
- Longitudinal Business Database (Kinney & Reiter, 2008)

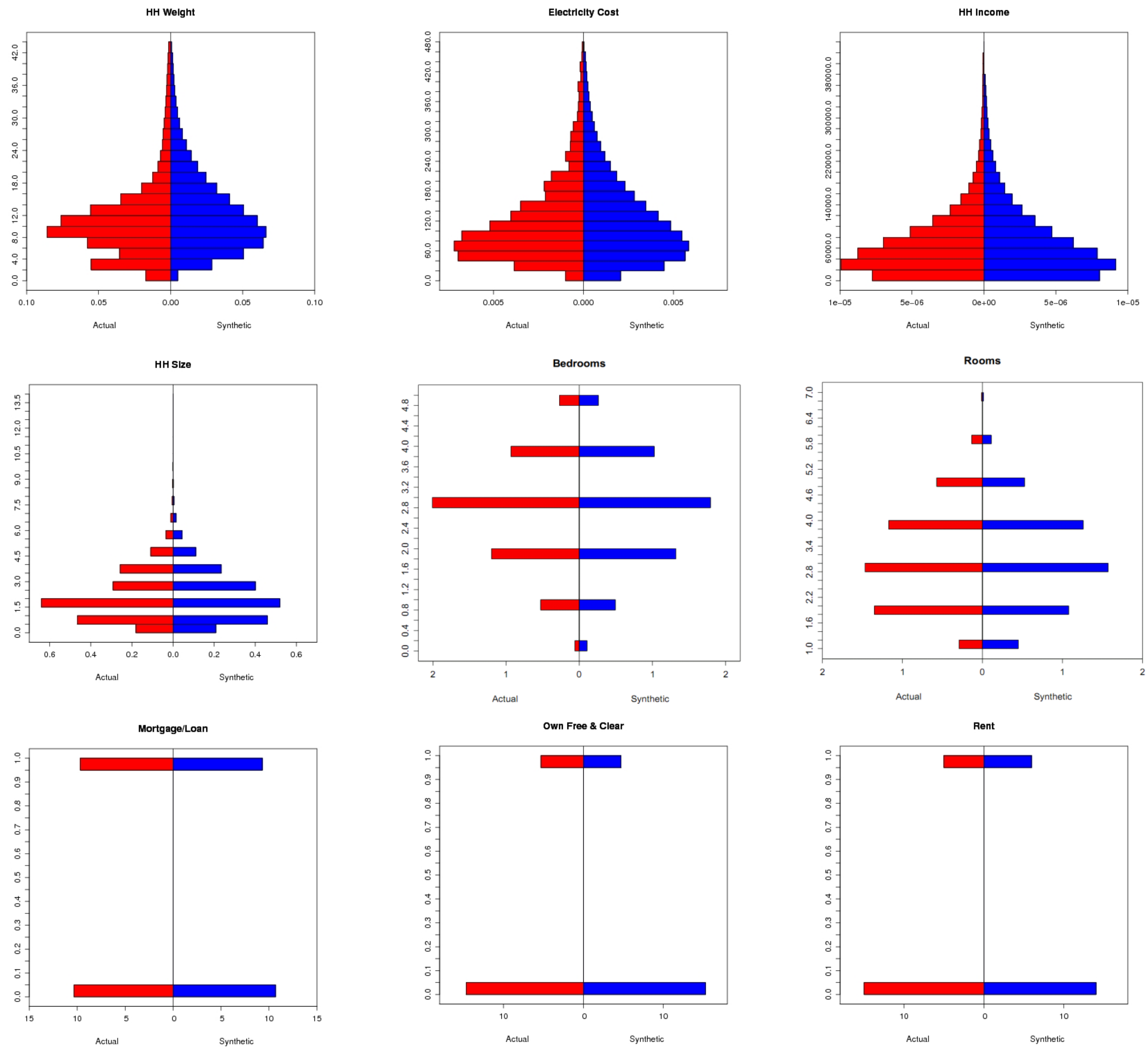
- Applications focus on preserving statistics about the entire sample, but ignores small area statistics.

# Synthetic Small Area Microdata Project

- Jointly funded by the U.S. Census Bureau and the Centers for Disease Control and Prevention
- Project goals
- 1) Develop synthetic data generation method
  - Hierarchical Bayesian model
- 2) Generate synthetic microdata for counties
  - American Community Survey (Northeast region)
  - National Health Interview Survey (sampled/nonsampled areas)
- 3) Compare inferences obtained from synthetic and actual data
  - Descriptive and analytic statistics

# Selected Items

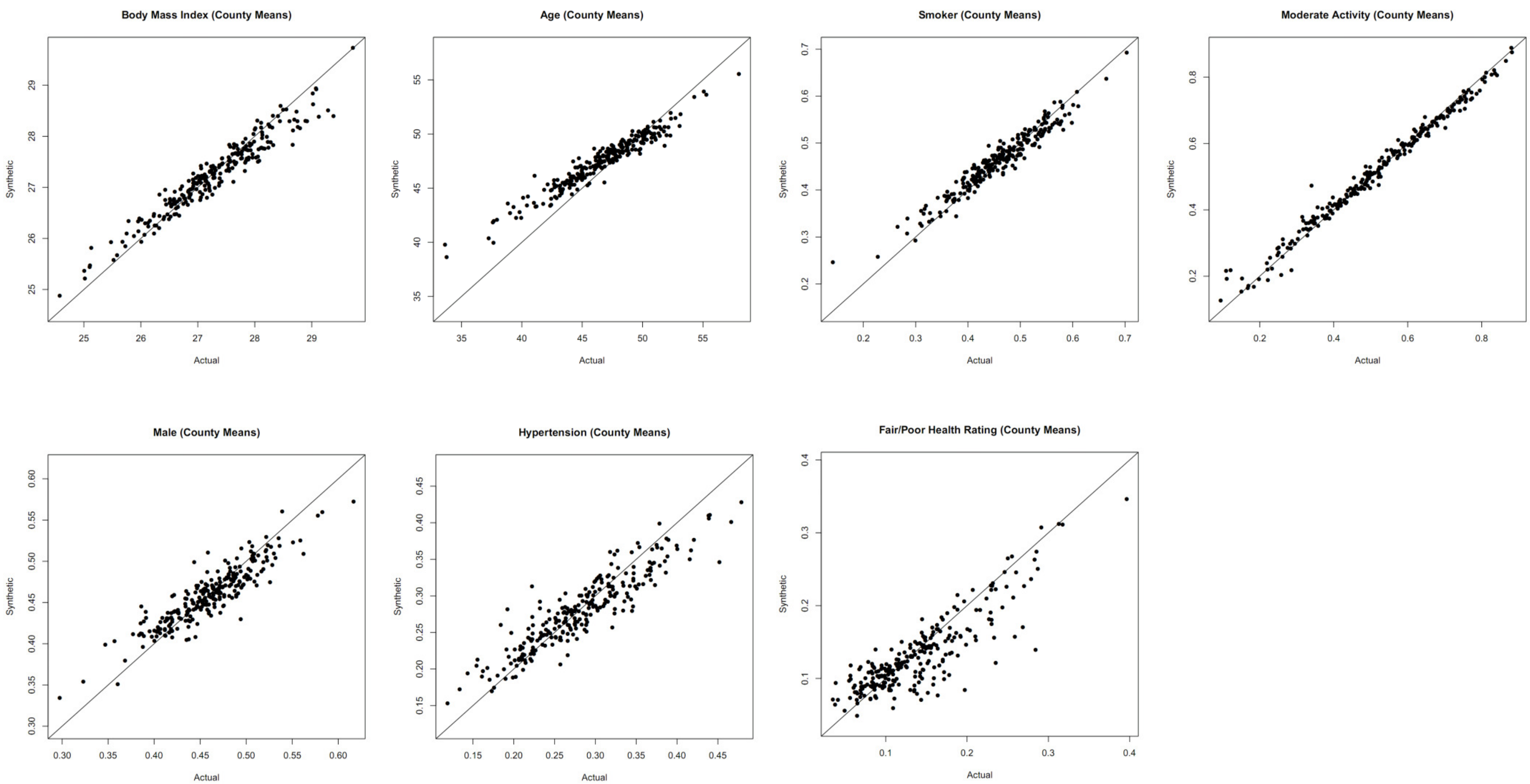
- Household items:
  - Household size, income, tenure (own, mortgage, rent), electricity payment, number of rooms
- Person items:
  - Age, sex, race, ethnicity, poverty status, self-reported health status, body mass index, smoking status, moderate activity, hypertension


HH Weight
Electricity Cost
HH Income
HH Size
Bedrooms
Rooms
Mortgage/Loan
Own Free & Clear
Rent
Actual
Synthetic

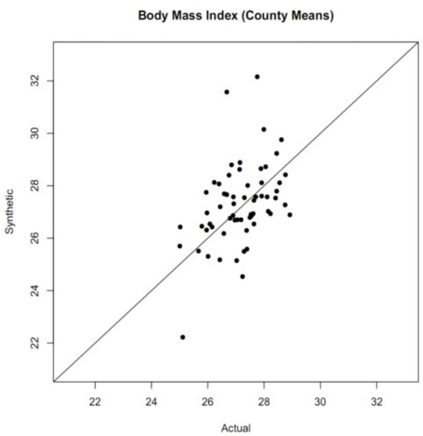
# Average County-Level Estimates: Household

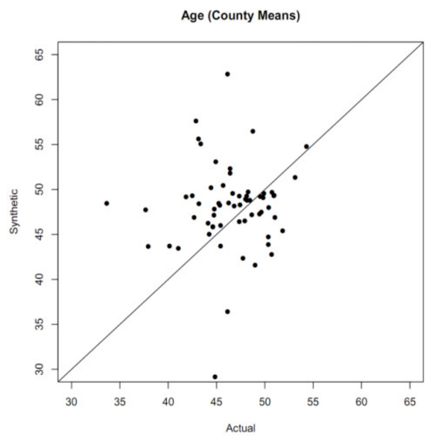
| | **Avg. County Means** | | **Avg. County Standard Errors** | |
|---|---|---|---|---|
| **Household variables** | Actual | Synthetic | Actual | Synthetic |
| Household size | 2.12 | 2.12 | 0.02 | 0.01 |
| Sampling weight | 9.99 | 10.20 | 0.11 | 0.11 |
| Number of bedrooms | 2.88 | 2.82 | 0.01 | 0.01 |
| Electricity cost/month | 118.89 | 119.37 | 1.25 | 1.10 |
| Number of rooms | 3.23 | 3.18 | 0.02 | 0.02 |
| Income | 67983.89 | 67382.38 | 1067.29 | 692.56 |
| Tenure: Mortgage/loan (%) | 49.00 | 47.03 | 0.82 | 0.74 |
| Tenure: Own free & clear (%) | 31.12 | 30.37 | 0.77 | 0.72 |
| Tenure: Rent (%) | 19.88 | 22.60 | 0.63 | 0.63 |

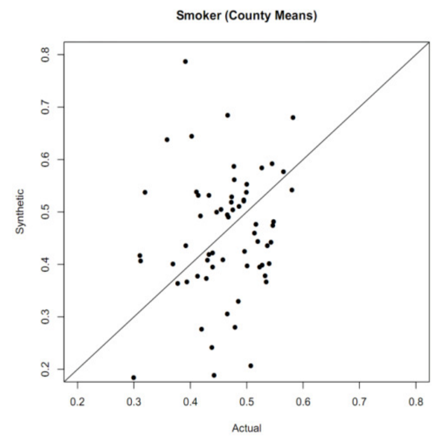
# County Means: Actual vs. Synthetic

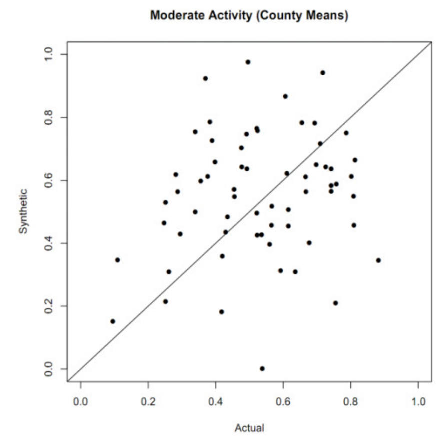


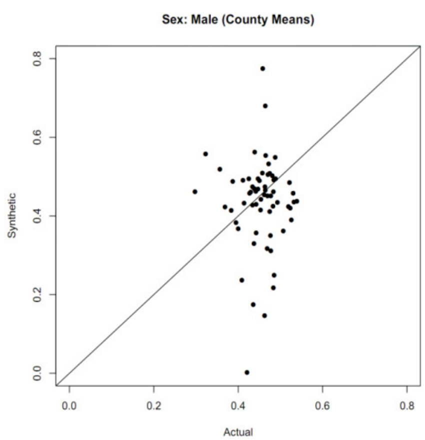
# Cross-Validation Study: Nonsampled County Means (N=63)

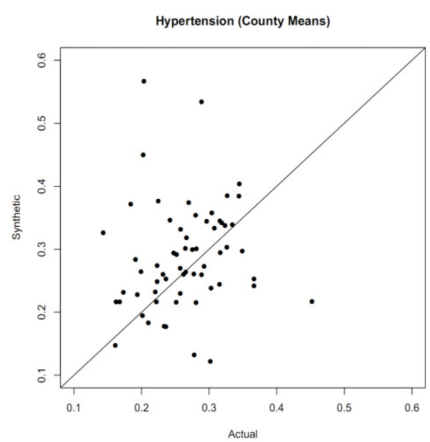


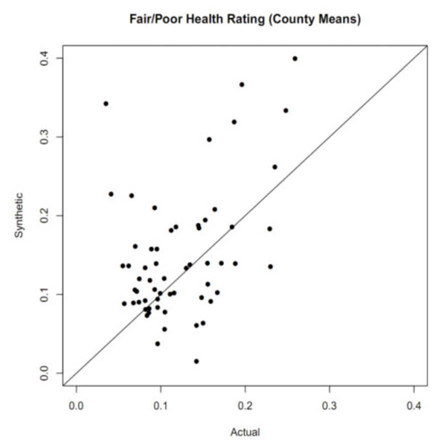



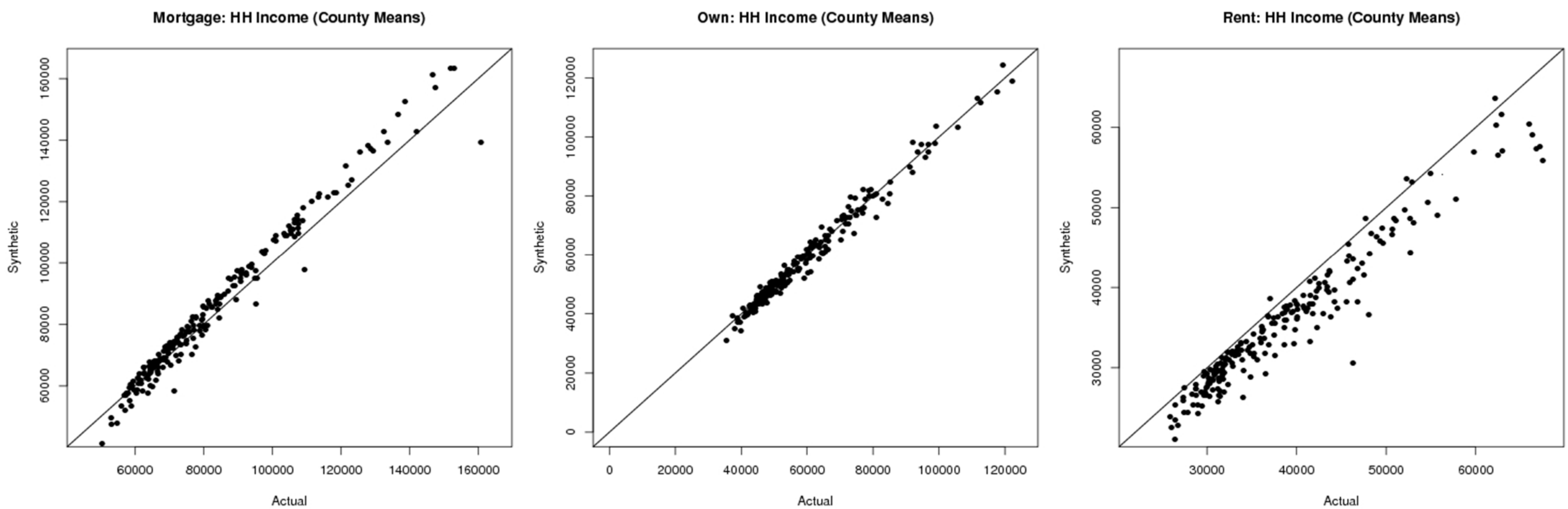



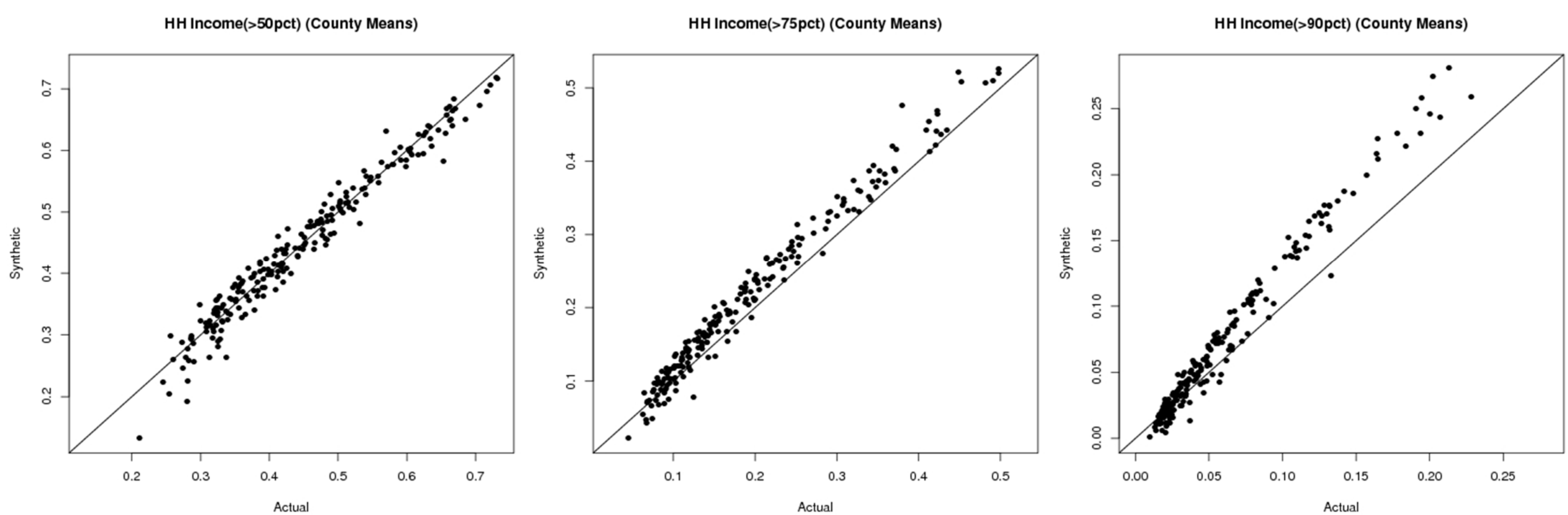



## A) Mean Income by Household Tenure




## B) Household Income > 50$^{th}$, 75$^{th}$, and 90$^{th}$ percentiles



# Average County-Level Regression Estimates

| | Avg. County Coefficients | | Avg. County Standard Errors | |
|---|---|---|---|---|
| ***Linear regression of household income on*** | Actual | Synthetic | Actual | Synthetic |
| Intercept | 24.34 | 24.26 | 1.11 | 1.09 |
| Household size | 1.52 | 1.44 | 0.14 | 0.14 |
| Sampling weight | -0.04 | -0.05 | 0.24 | 0.26 |
| Number of bedrooms | 1.15 | 1.23 | 0.19 | 0.18 |
| Electricity cost/month | 0.99 | 1.04 | 0.18 | 0.17 |
| Number of rooms | 1.25 | 1.26 | 0.14 | 0.13 |
| Tenure: Mortgage/loan | Ref | Ref | Ref | Ref |
| Tenure: Own free & clear | -3.47 | -3.05 | 0.37 | 0.34 |
| Tenure: Rent | -6.01 | -6.84 | 0.44 | 0.47 |

# Conclusions

- Synthetic data preserves small area statistics reasonably well in most cases
  - Univariate/multivariate, subgroup estimates
- Modeling approach could be improved
  - non-standard distributions, multinomial distributions
- Practical Strengths
  - Easy to implement; doesn’t require MCMC
  - Data can presumably be released to the public without restriction (needs disclosure risk analysis)
  - Method could be adopted for large scale survey projects

Thanks for your attention!

joesaks@umich.edu

# Modeling Approach

- Extension of SRMI (Raghunathan et al., 2001)
- Hierarchical Bayesian Model
  - Two levels; e.g., counties nested within states
- Fit sequential regression models within each small area,

$$f\left(Y_{cs,1}\right), f\left(Y_{cs,2}|Y_{cs,1}\right),\ldots,f\left(Y_{cs,P}|Y_{cs,1},\ldots,Y_{cs,P-1}\right)$$

- Approximate distribution of design-based parameter estimates,

$$\hat{\beta}_{cs,p} \sim MVN\left(\beta_{cs,p},\hat{V}_{cs,p}\right)$$

- Assign proper prior to the unknown population parameter ,

$$\beta_{cs,p} \sim MVN\left(\hat{\beta}_p Z_s,\hat{\Sigma}_p\right)$$

- Draw unknown parameter from posterior distribution,

$$\tilde{\beta}_{cs,p} \sim MVN\left[\left(\hat{V}_{cs,p}^{-1}+\hat{\Sigma}_p^{-1}\right)^{-1}\left(\hat{V}_{cs,p}^{-1}\hat{\beta}_{cs,p}+\hat{\Sigma}_p^{-1}\hat{\beta}_p Z_s\right),\left(\hat{V}_{cs,p}^{-1}+\hat{\Sigma}_p^{-1}\right)^{-1}\right]$$

# Model Setup

- Estimate weighted $\hat{\beta}_{cs,p}$ by fitting sequential regression models

$$\hat{\beta}_{cs,p} \sim MVN\left(\beta_{cs,p}, \hat{V}_{cs,p}\right) \qquad [1]$$

$$\beta_{cs,p} \sim MVN\left(\beta_{s,p} Z_{cs}, \Sigma_{s,p}\right) \qquad [2]$$

$$\hat{\beta}_{s,p} \sim MVN\left(\beta_{s,p}, \hat{V}_{s,p}\right) \qquad [3]$$

$$\beta_{s,p} \sim MVN\left(\beta_p Z_s, \Omega_p\right) \qquad [4]$$

- Hyperparameters estimated using EM algorithm (Dempster et al., 1977)

$$\tilde{\beta}_{s,p} \sim MVN\left[\left(\hat{V}_{s,p}^{-1} + \hat{\Omega}_p^{-1}\right)^{-1}\left(\hat{V}_{s,p}^{-1}\hat{\beta}_{s,p} + \hat{\Omega}_p^{-1}\hat{\beta}_p Z_s\right), \left(\hat{V}_{s,p}^{-1} + \hat{\Omega}_p^{-1}\right)^{-1}\right]$$

$$\tilde{\beta}_{cs,p} \sim MVN\left[\left(\hat{V}_{cs,p}^{-1} + \hat{\Sigma}_{s,p}^{-1}\right)^{-1}\left(\hat{V}_{cs,p}^{-1}\hat{\beta}_{cs,p} + \hat{\Sigma}_{s,p}^{-1}\hat{\beta}_{s,p} Z_{cs}\right), \left(\hat{V}_{cs,p}^{-1} + \hat{\Sigma}_{s,p}^{-1}\right)^{-1}\right]$$

- Models [2] and [4] used to simulate coeffs for nonsampled areas
- Simulated coefficients used as inputs for drawing synthetic values

# Generating Synthetic Values

- Simulating a synthetic variable $\tilde{Y}_{cs}$ is achieved by drawing from the posterior predictive distribution in sequential order,

$$f\left(\tilde{Y}_{cs,1}|\tilde{\beta}_{cs}\right), f\left(\tilde{Y}_{cs,2}|\tilde{Y}_{cs,1},\tilde{\beta}_{cs}\right), \ldots, f\left(\tilde{Y}_{cs,P}|\tilde{Y}_{cs,1},\tilde{Y}_{cs,2},\ldots,\tilde{Y}_{cs,P-1},\tilde{\beta}_{cs}\right)$$

- For continuous variables,

$$\tilde{Y}_{cs,p} \sim N\left[\left(X_{cs},\tilde{Y}_{cs,1},\tilde{Y}_{cs,2},\tilde{Y}_{cs,p-1}\right)\tilde{\beta}_{cs},\hat{\sigma}_{cs}^{2}\right]$$

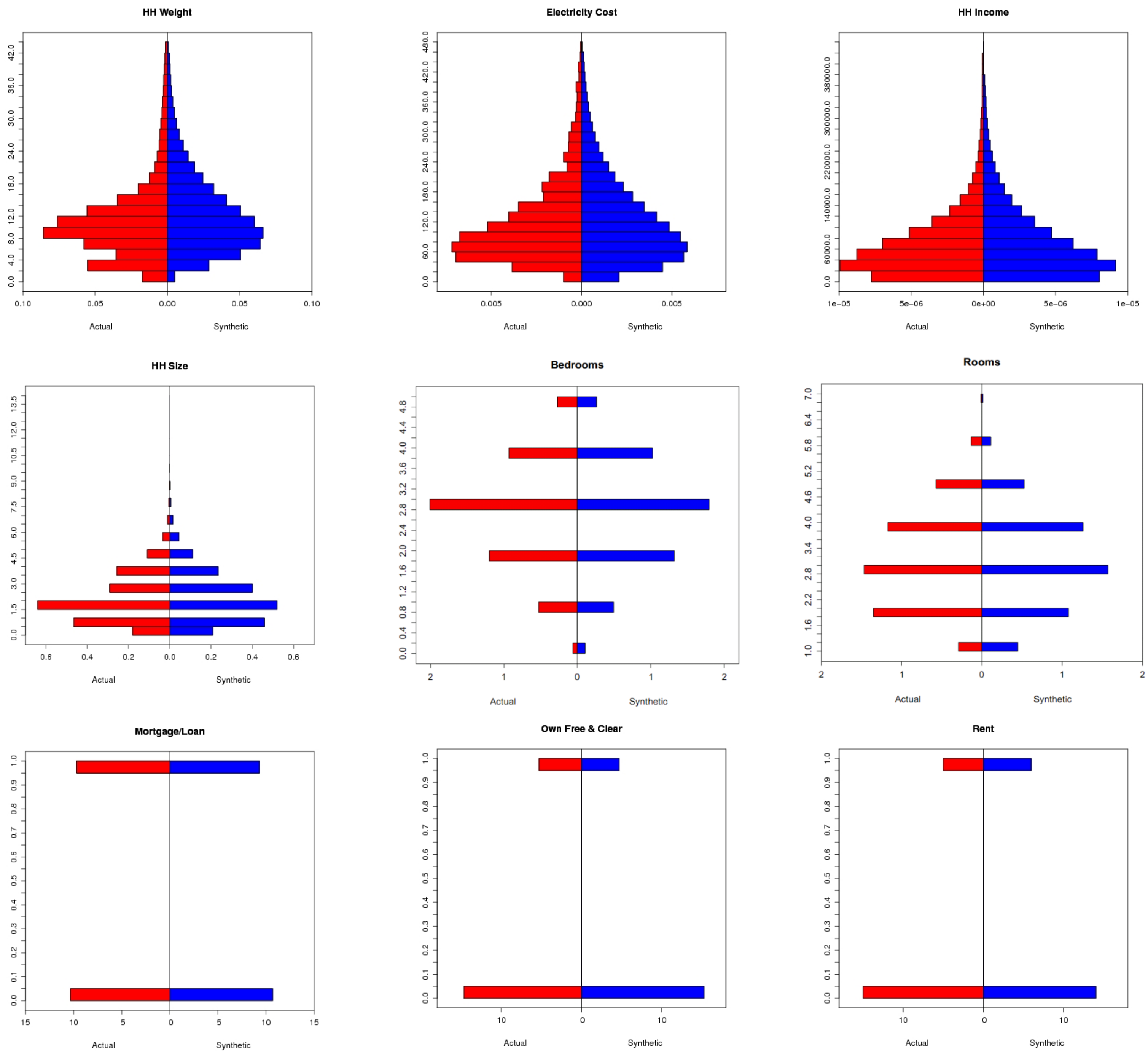
- For binary variables,

$$\tilde{Y}_{cs,p} \sim Bin\left[1,\hat{p}\left\{\left(X_{cs},\tilde{Y}_{cs,1},\tilde{Y}_{cs,2},\ldots,\tilde{Y}_{cs,p-1}\right)\tilde{\beta}_{cs}\right\}\right]$$

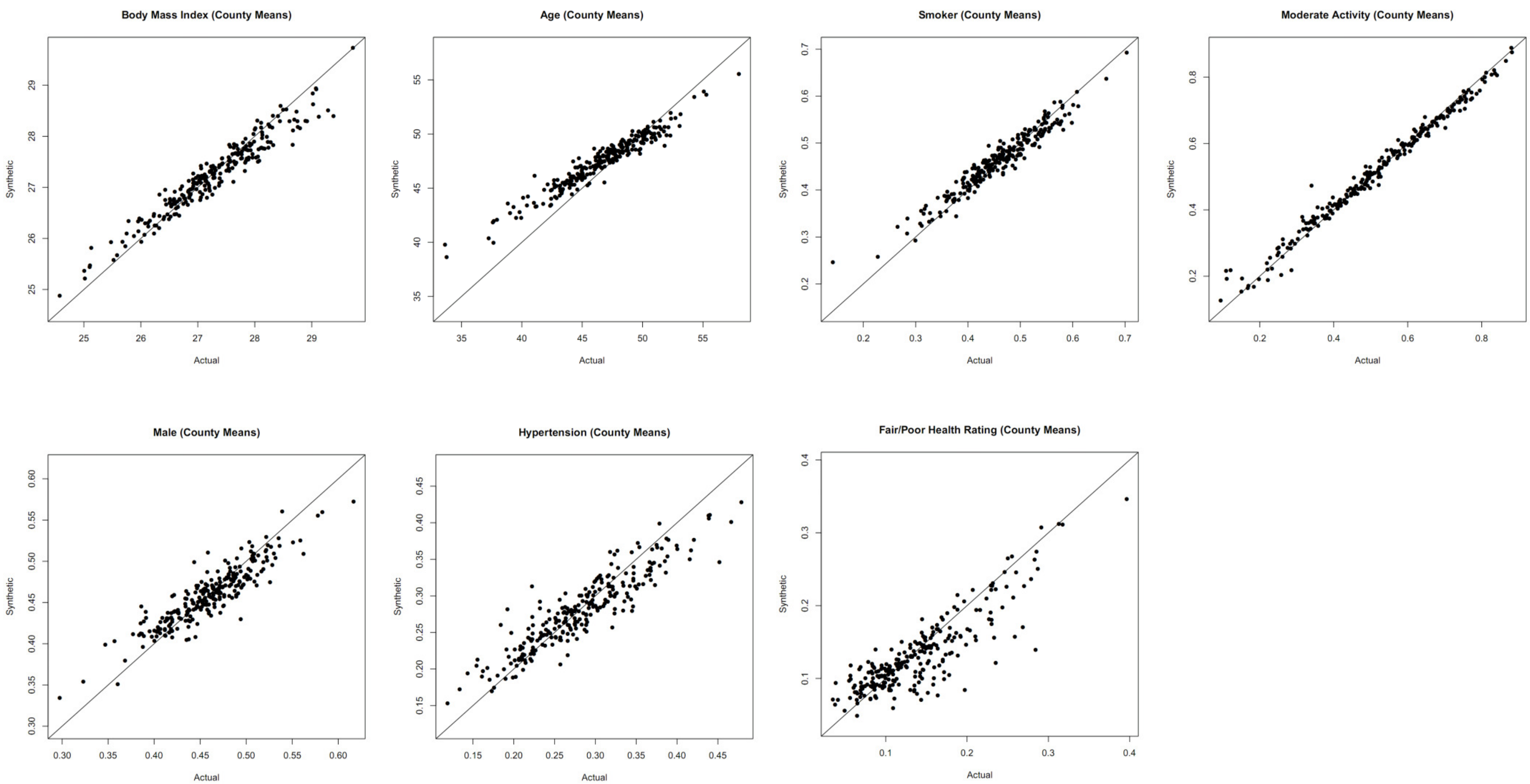
- Extension to count and multinomial distributions is possible
- Process is repeated $M$ times to produce $M$ synthetic data sets
  - $M$ = 5 (or 10) is usually sufficient to obtain valid inferences (Reiter, 2005)
- Valid inferences obtained using standard combining rules (Raghunathan, Reiter, Rubin, 2003)

# Application 1: American Community Survey

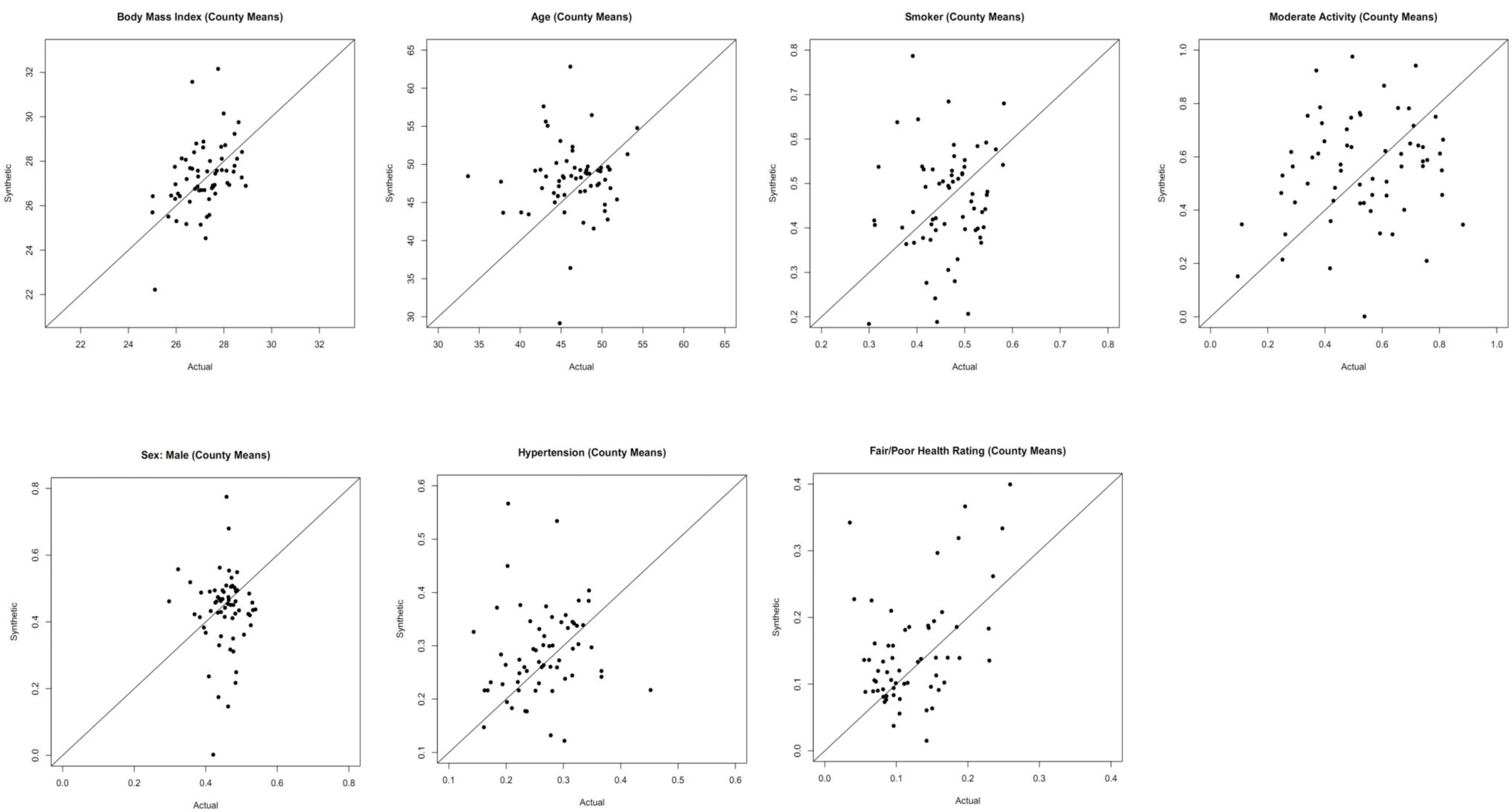
- American Community Survey (2005-2009)
- "Small areas" defined as counties
- Northeast Region (N=217 counties)
- N = 599,450 households; 1,506,011 persons
- $M = 10$ synthetic data sets
- Household-level variables
  - Sampling weight, electricity cost/mo., income, household size, # bedrooms, # total rooms, household tenure (mortgage/loan, own free & clear, rent)

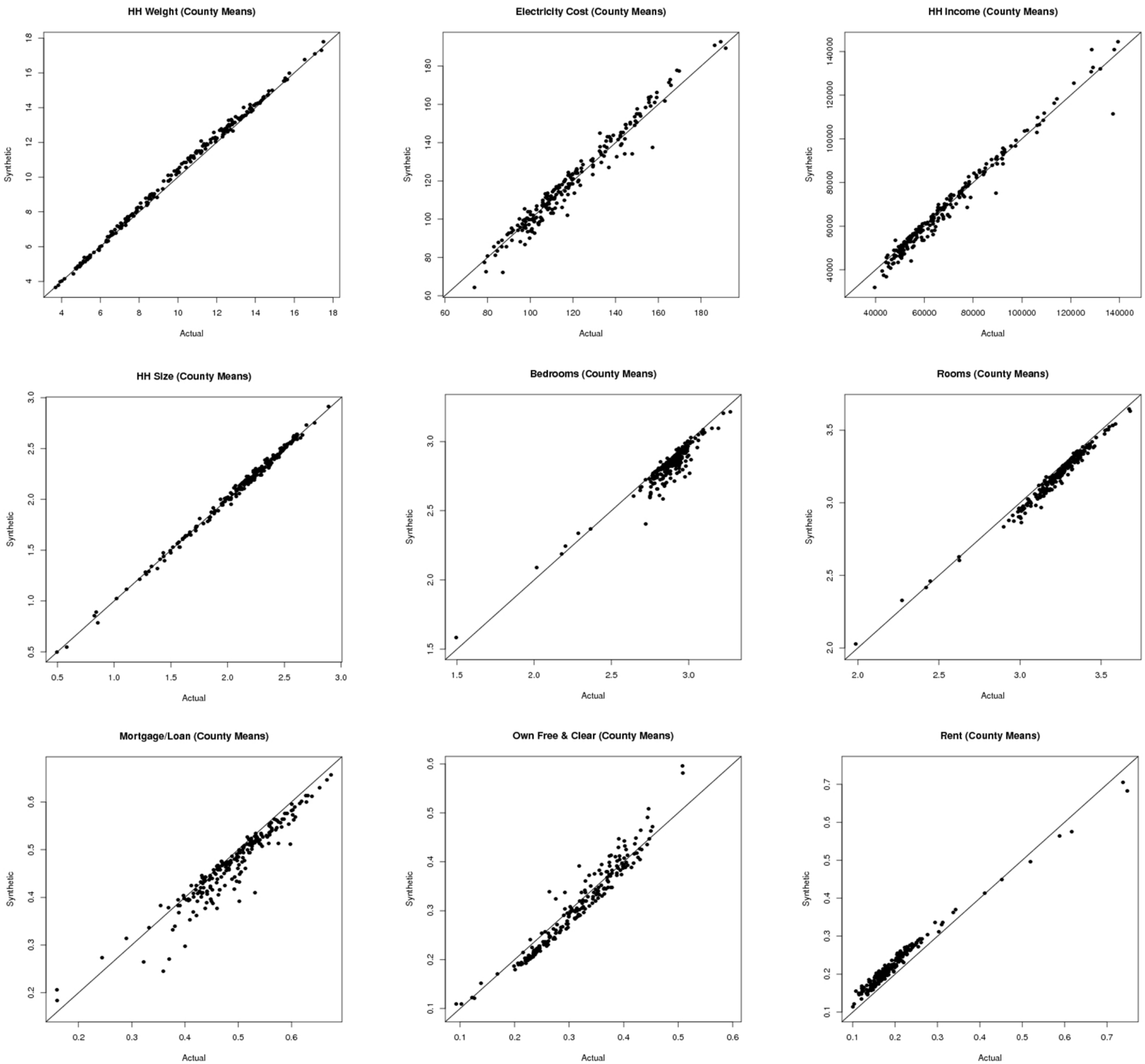

HH Weight
Actual
Synthetic
Electricity Cost
Actual
Synthetic
HH Income
Actual
Synthetic
HH Size
Actual
Synthetic
Bedrooms
Actual
Synthetic
Rooms
Actual
Synthetic
Mortgage/Loan
Actual
Synthetic
Own Free & Clear
Actual
Synthetic
Rent
Actual
Synthetic

# PSU Means: Actual vs. Synthetic (sampled)



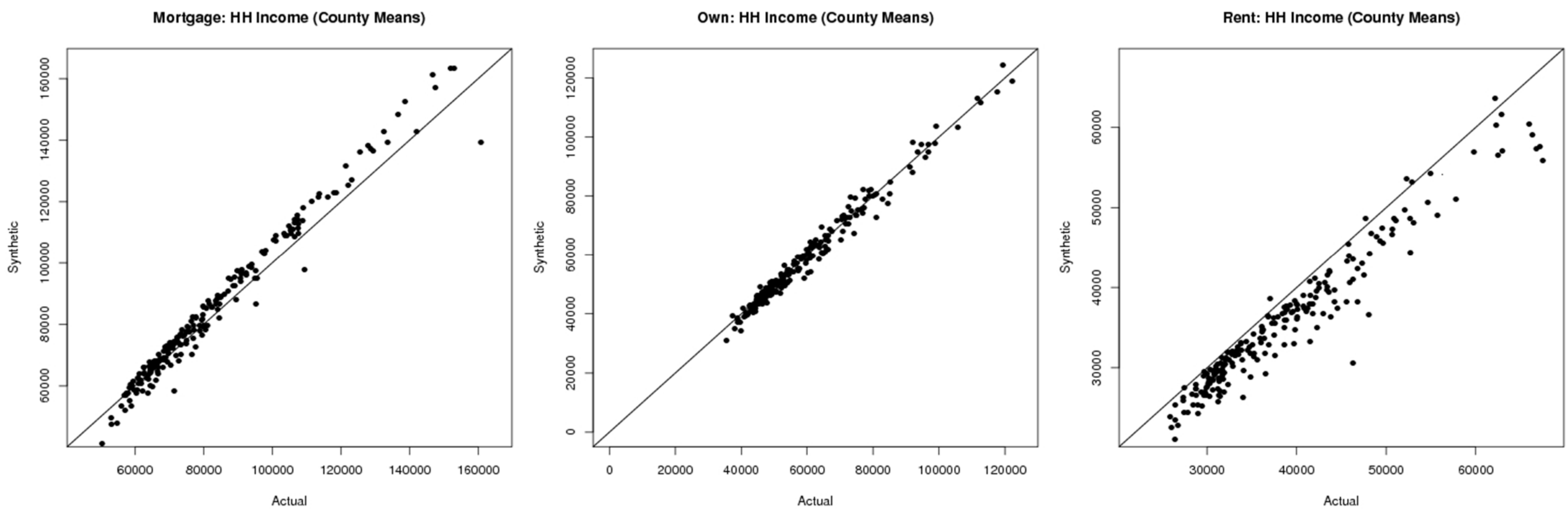
# Cross-Validation Study: Nonsampled County Means (N=63)



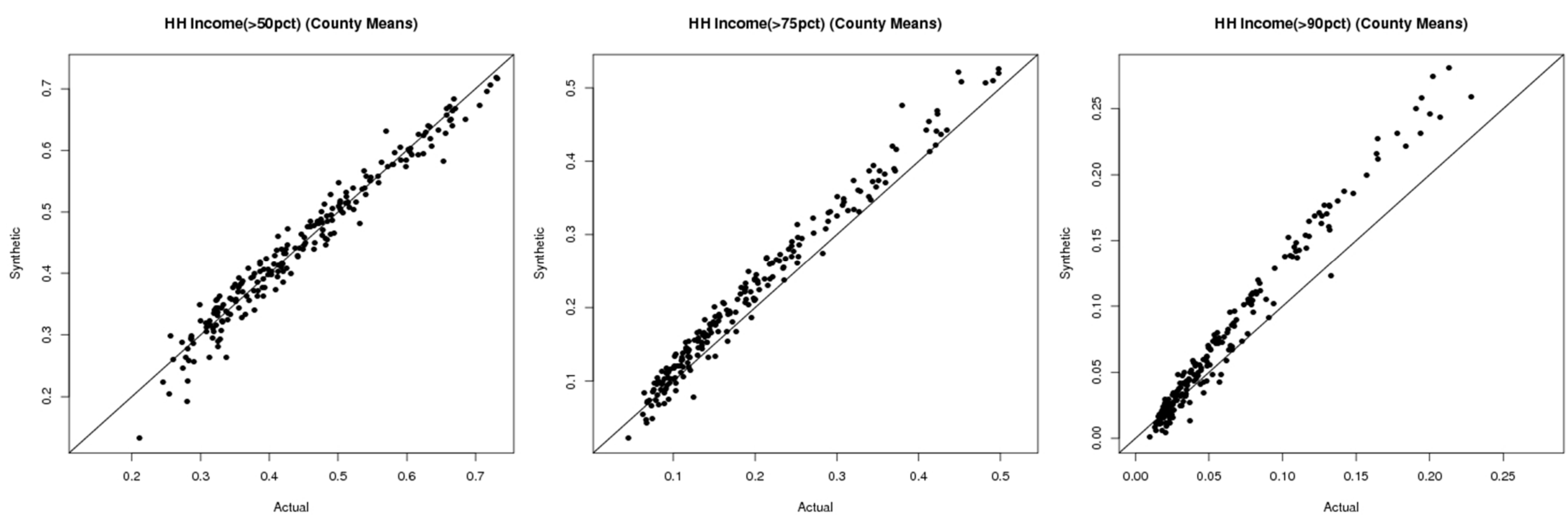


HH Weight (County Means)
Electricity Cost (County Means)
HH Income (County Means)
HH Size (County Means)
Bedrooms (County Means)
Rooms (County Means)
Mortgage/Loan (County Means)
Own Free & Clear (County Means)
Rent (County Means)
Synthetic
Actual

## A) Mean Income by Household Tenure

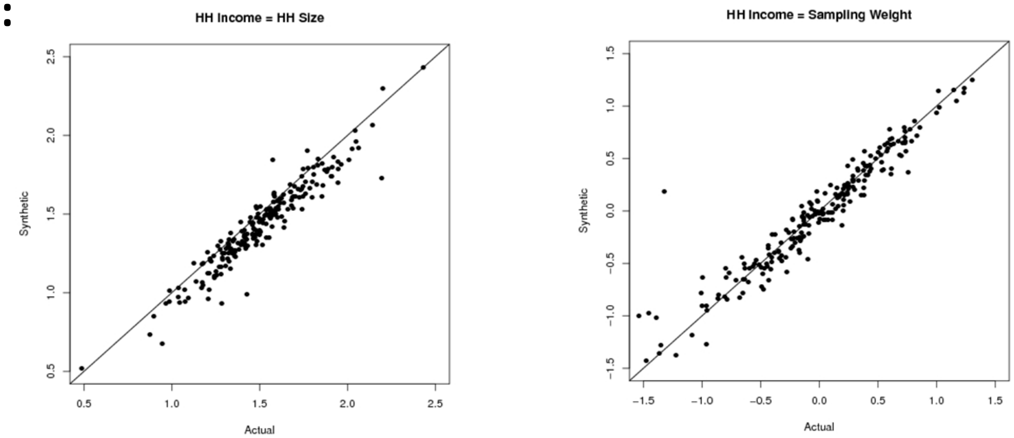


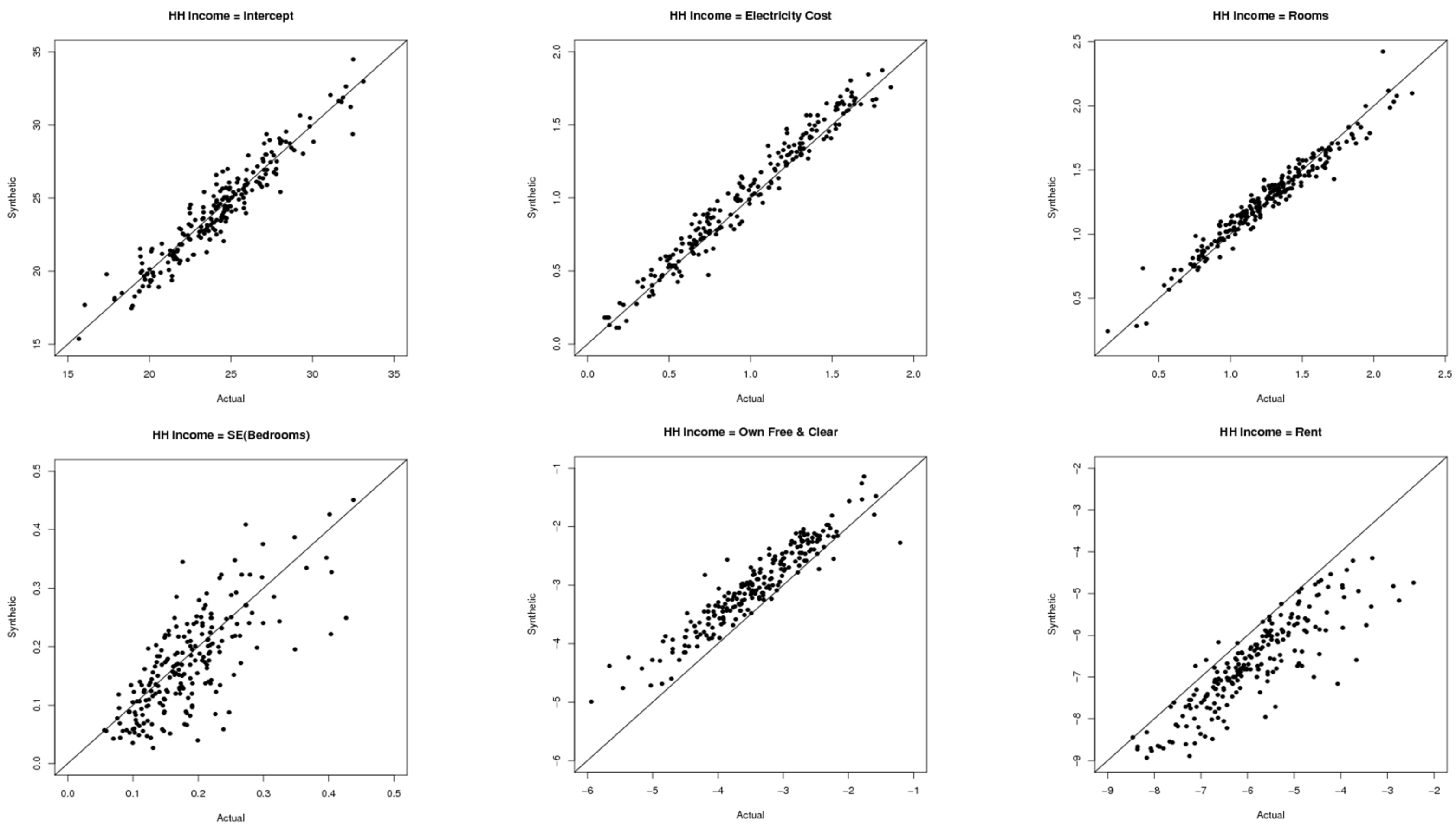

## B) Household Income > 50th, 75th, and 90th percentiles



**Linear Regression of Income on:**

HH Size
Sampling weight
Electricity cost/mo.
# Bedrooms
# Total rooms
Tenure: Own Free & Clear
Tenure: Rent



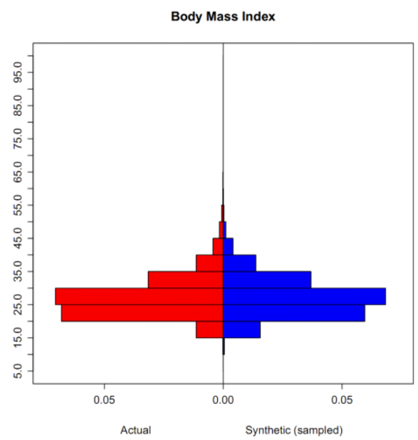
# Simulation Study: CI Coverage – PUMA Means

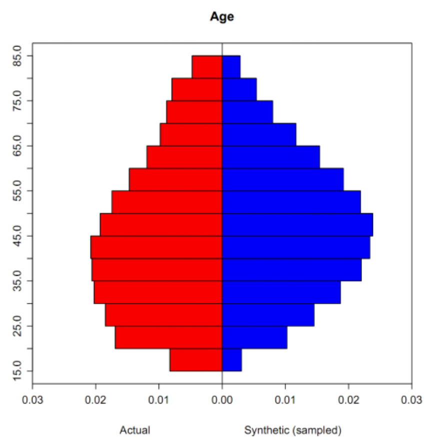
| | **Conditional** | **Unconditional** | |
|---|---|---|---|
| | | Synthetic | Actual |
| Household size | 0.99 | 0.98 | 0.98 |
| Sampling weight | 0.95 | 0.99 | 0.98 |
| Number of bedrooms | 0.89 | 0.93 | 0.98 |
| Electricity cost/month | 0.86 | 0.91 | 0.98 |
| Number of rooms | 0.97 | 0.98 | 0.98 |
| Income | 0.90 | 0.94 | 0.98 |
| Tenure: Own free & clear | 0.93 | 0.96 | 0.98 |
| Tenure: Rent | 0.94 | 0.96 | 0.98 |
| **Coverage mean** | **0.93** | **0.96** | **0.98** |

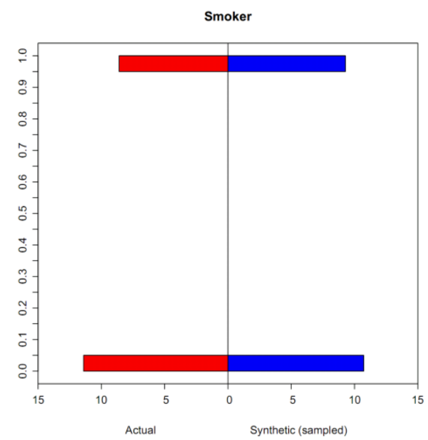
# Application 2: National Health Interview Survey

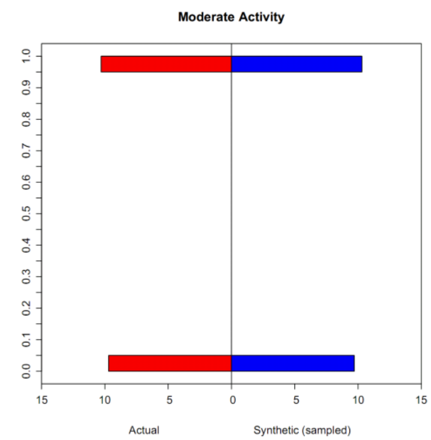
- NHIS 2003-2005; Complex sample survey
- Treat PSUs as "small areas" nested within strata
- Generate synthetic data for both sampled PSUs and nonsampled counties
- Incorporate PSU/county-level and stratum/state-level covariates into hierarchical model
- N=93,606 sampled adults
- Continuous and binary items
  - Age, body mass index, smoker, sex, moderate activity, hypertension, fair/poor health rating

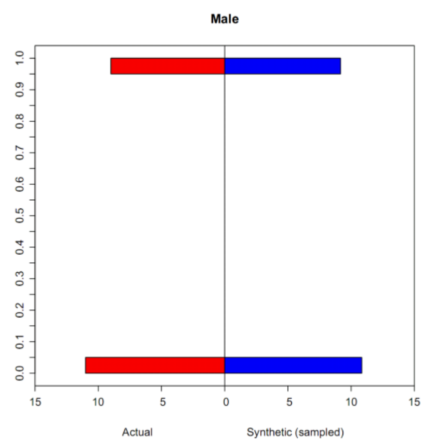
# Actual vs. Synthetic (sampled/nonsampled)

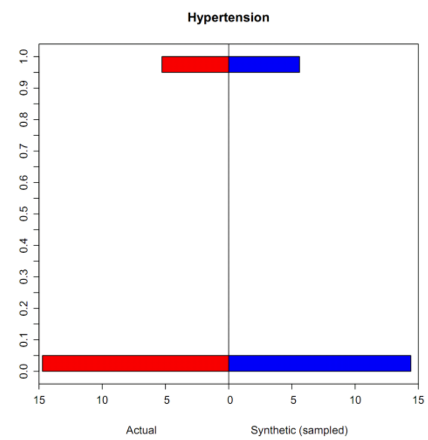


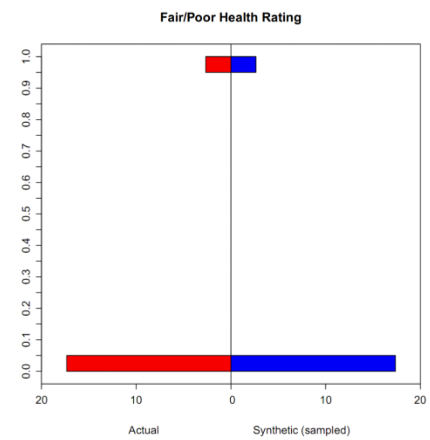



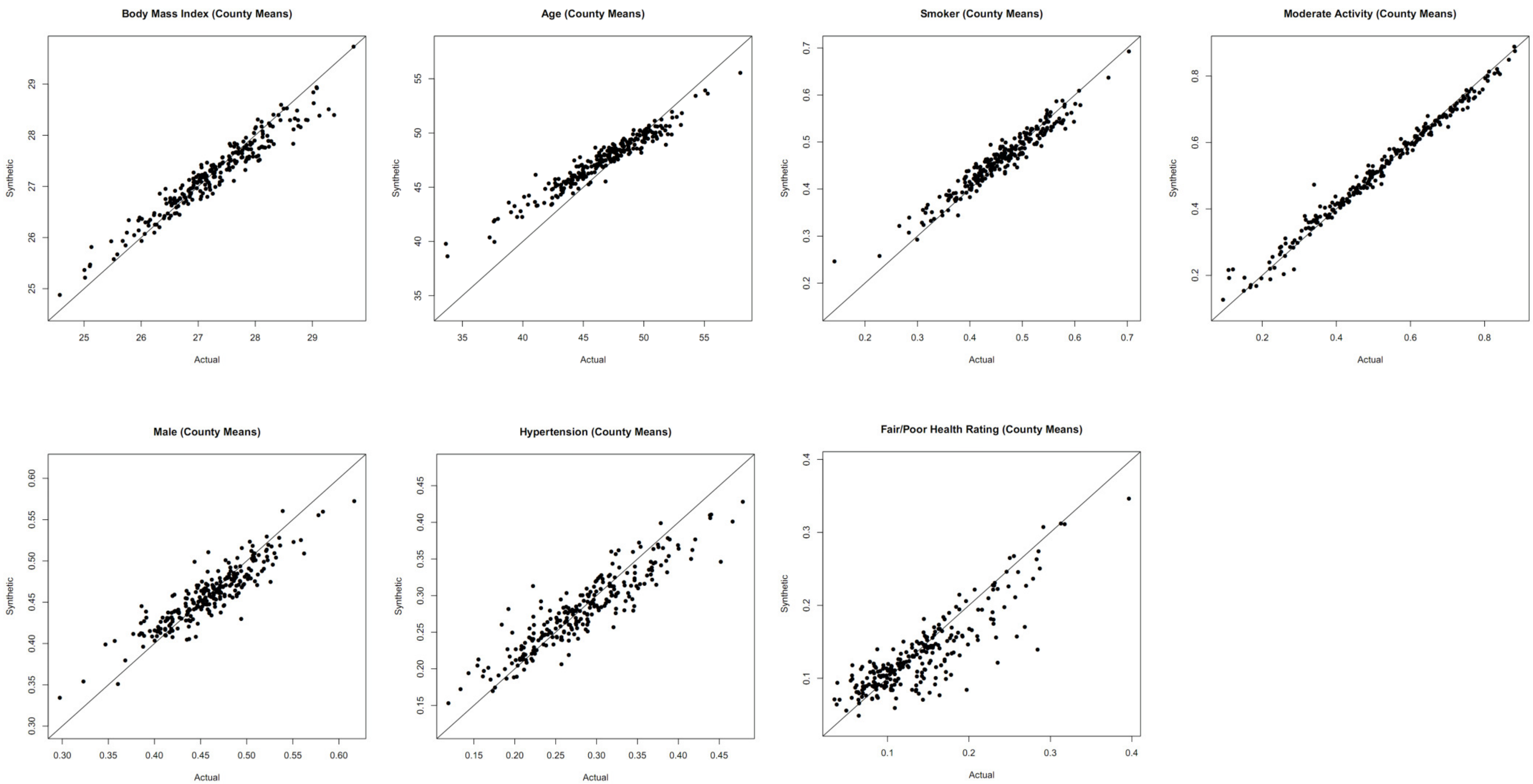



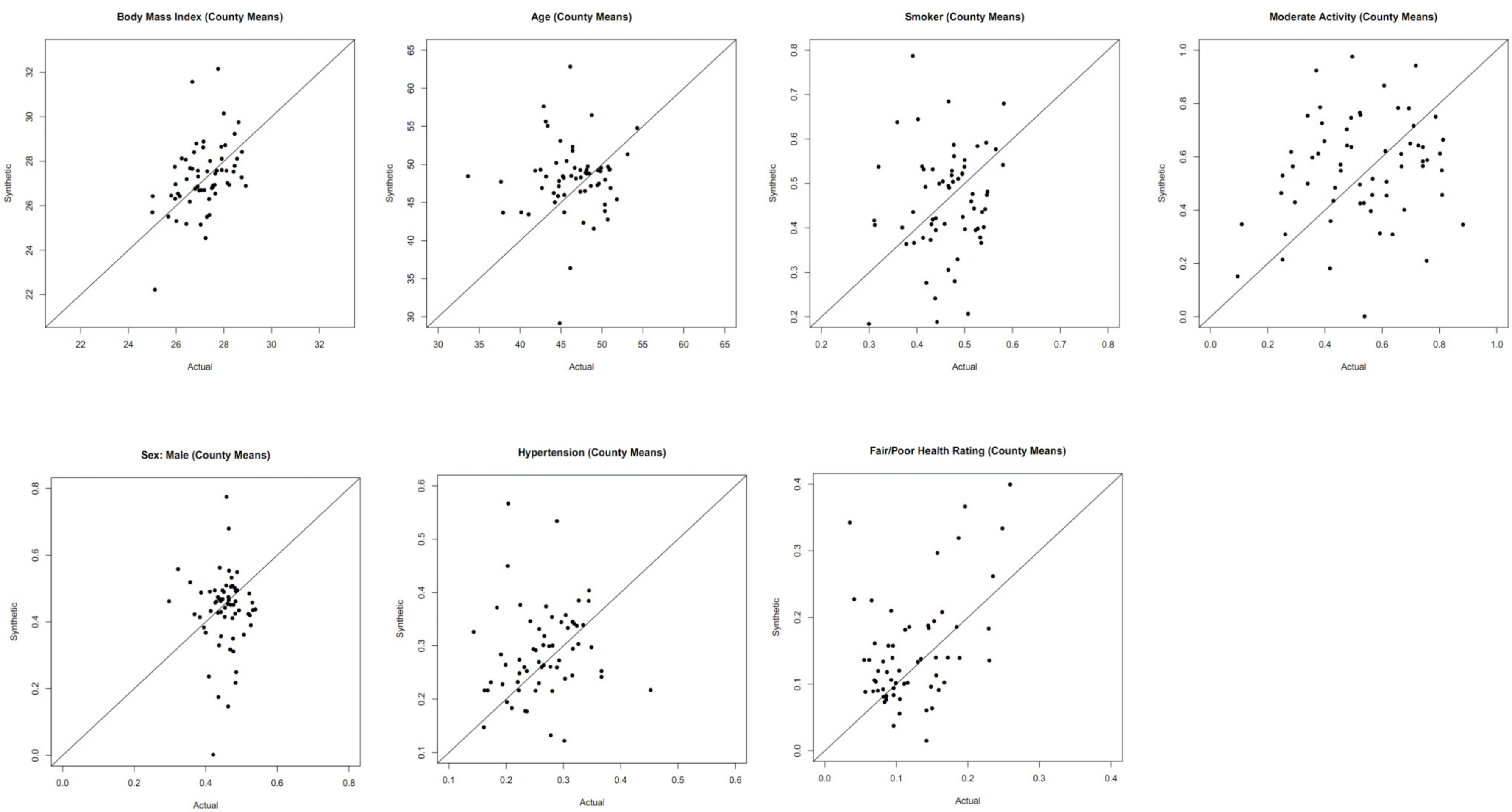



# PSU Means: Actual vs. Synthetic (sampled)



# Cross-Validation Study: Nonsampled County Means (N=63)



# Simulation Study: CI Coverage (Means & Regression Coefficients)

| | Conditional Inference | Unconditional Inference | |
|---|---|---|---|
| | CIC | CIC | Actual |
| **BMI** | 0.99 | 0.99 | 0.97 |
| **Age** | 0.91 | 0.99 | 0.98 |
| **Smoker** | 0.99 | 0.99 | 0.98 |
| **Moderate activity** | 0.99 | 0.99 | 0.98 |
| **Male** | 0.99 | 0.99 | 0.98 |
| **Hypertension** | 0.99 | 0.99 | 0.97 |
| **Fair/poor health** | 0.99 | 0.99 | 0.97 |

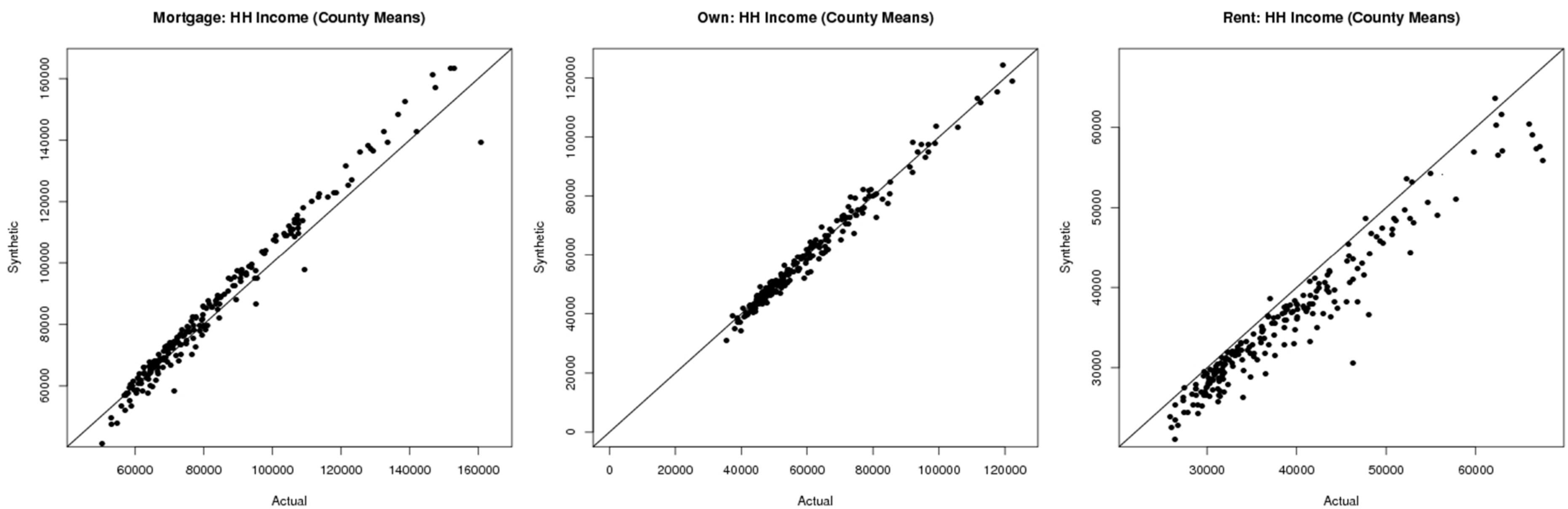
| | Conditional Inference | Unconditional Inference | |
|---|---|---|---|
| **Covariates** | CIC | CIC | Actual |
| **Regression of BMI(log) on** | | | |
| **Intercept** | 0.99 | 0.99 | 0.97 |
| **Age** | 0.99 | 0.99 | 0.97 |
| **Smoker** | 0.99 | 0.99 | 0.98 |
| **Moderate activity** | 0.99 | 0.99 | 0.97 |
| **Male** | 0.99 | 0.99 | 0.98 |
| **Hypertension** | 0.99 | 0.99 | 0.98 |
| **Fair/poor health** | 0.99 | 0.99 | 0.96 |

# Future Work

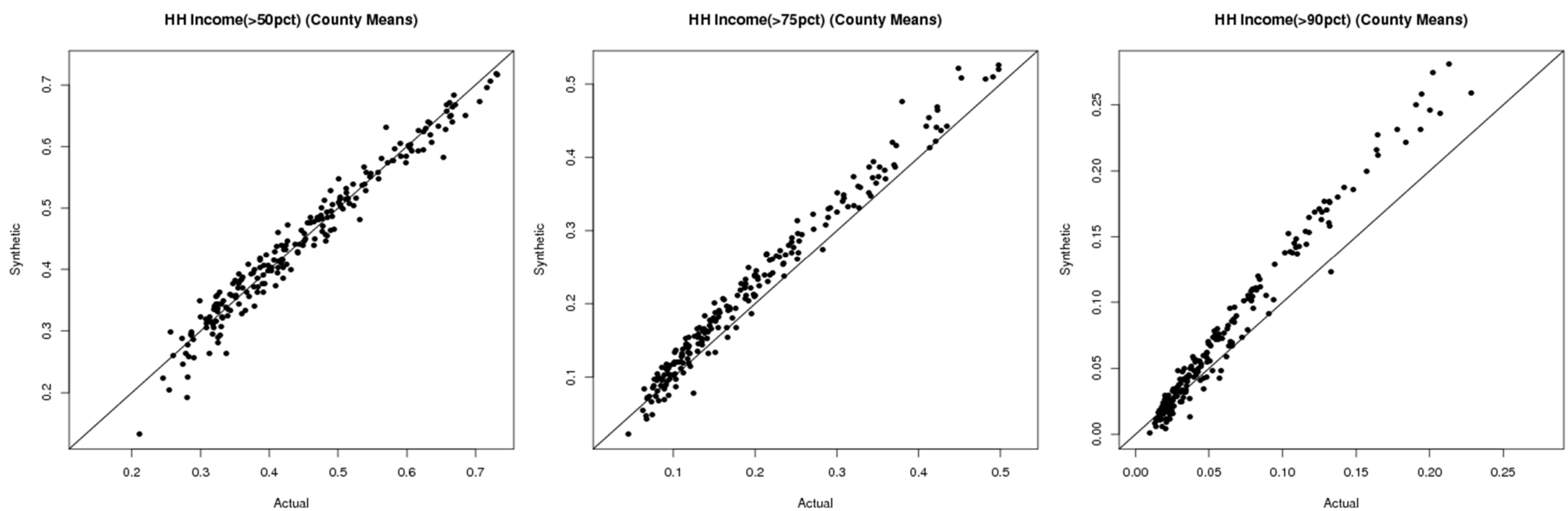
- Application to *smaller* areas is always desirable
  - Census tracts, block groups
- Additional variable distributions
  - E.g., mixed-type,
- Cross-classified tables
  - add more details in public-use summary files
- Incorporate auxiliary information to improve efficiency of synthetic data estimates
  - E.g., Administrative data, external survey data (e.g., BRFSS)
- Longitudinal small area estimates (e.g., HRS, PSID)

# PUMA Subgroup Means & Percentiles (Synthetic vs. Actual)

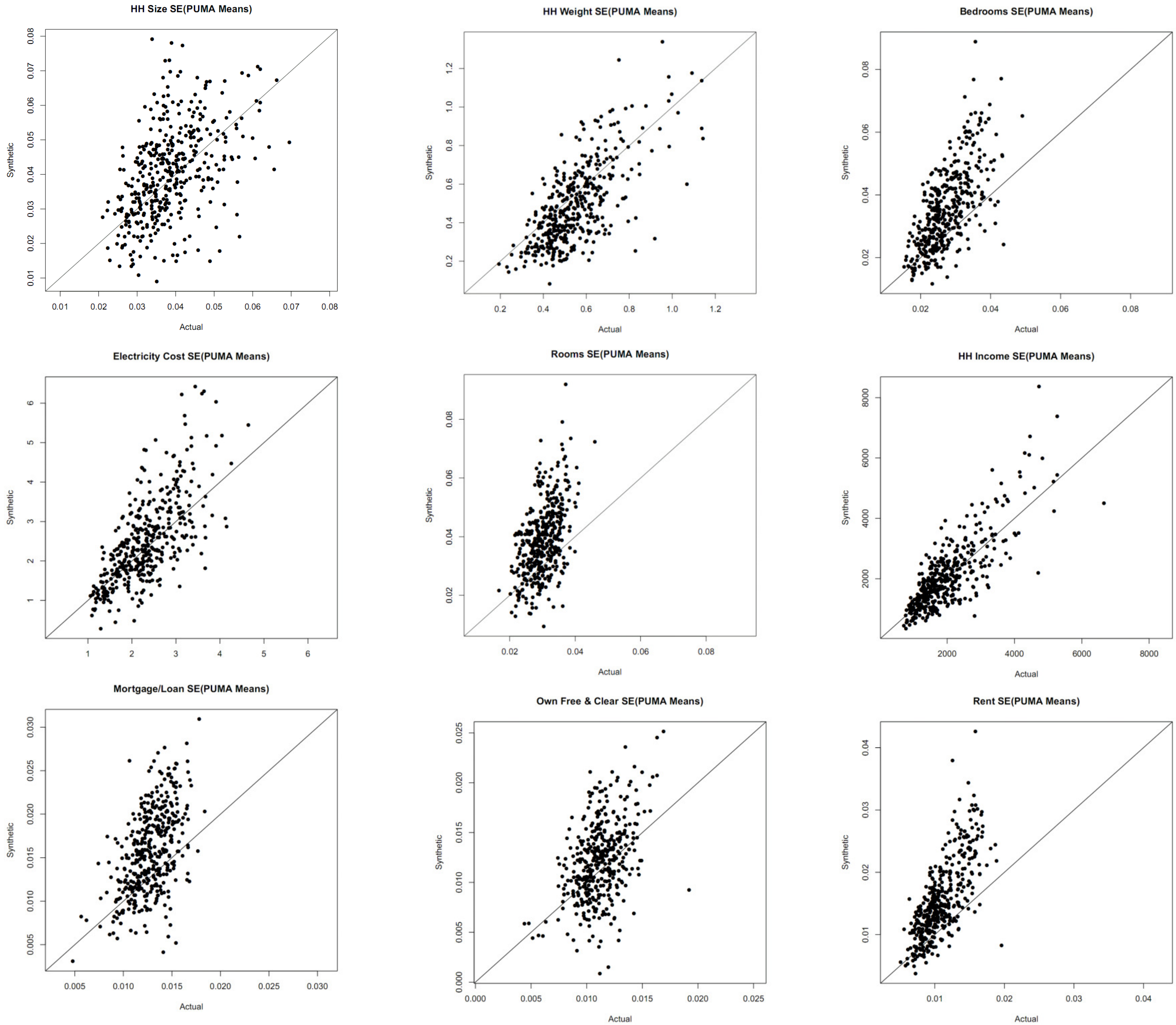
## A) Mean Income by Household Tenure




## B) Household Income > $50^{th}$, $75^{th}$, and $90^{th}$ percentiles

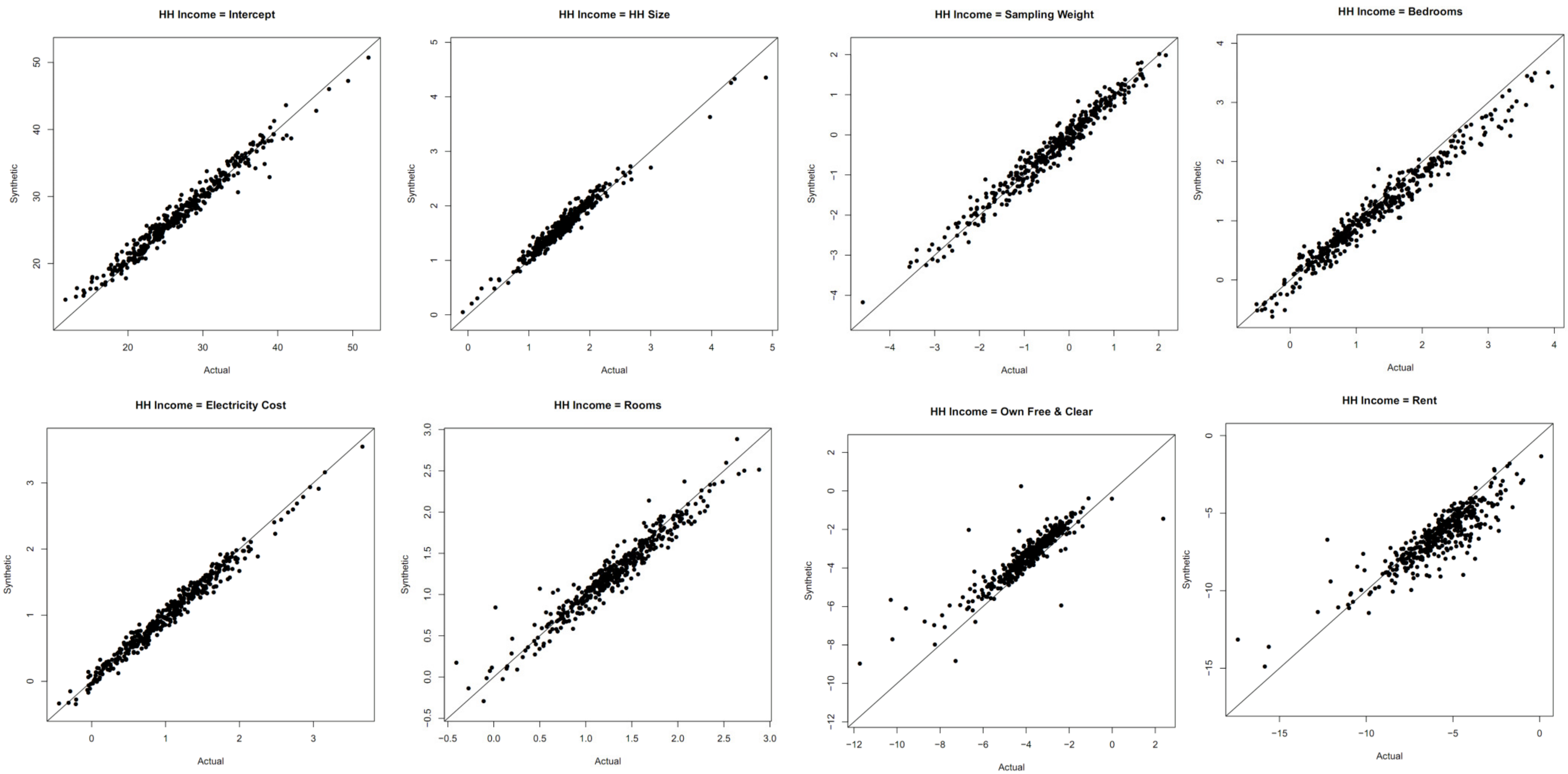


# Standard Errors of PUMA Means (Synthetic vs. Actual)

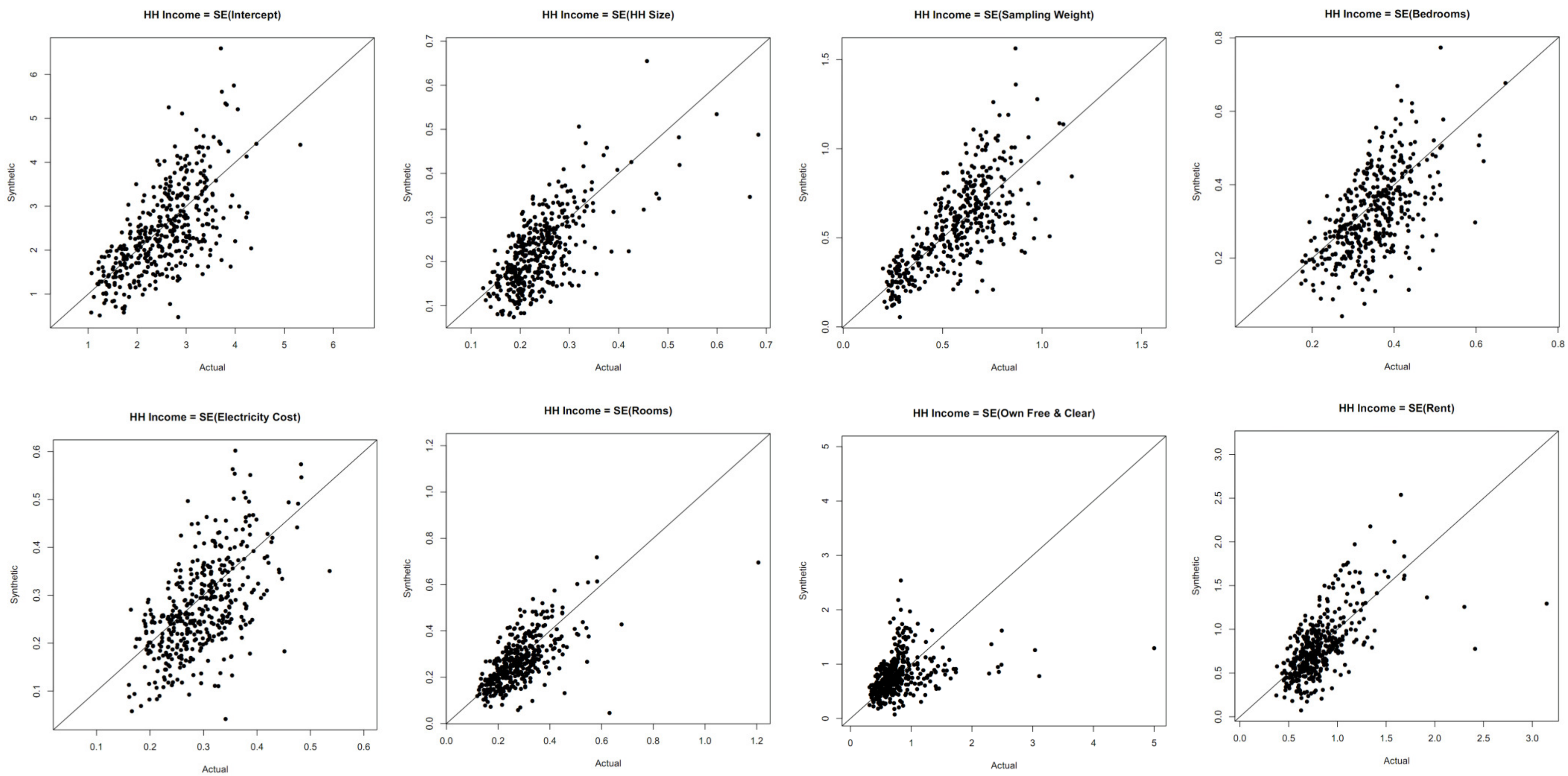

HH Size SE(PUMA Means)
HH Weight SE(PUMA Means)
Bedrooms SE(PUMA Means)
Electricity Cost SE(PUMA Means)
Rooms SE(PUMA Means)
HH Income SE(PUMA Means)
Mortgage/Loan SE(PUMA Means)
Own Free & Clear SE(PUMA Means)
Rent SE(PUMA Means)
Synthetic
Actual

# PUMA Regression Coefficients (Synthetic vs. Actual)

# HH Income (y) = Intercept + HH Size + Sampling Weight + Bedrooms + Electricity + Rooms + Own Free & Clear + Rent + Error



# Standard Errors
# (Household-Level Regression Coefficients)



# Average ACS County-Level Estimates: Household

| | **Avg. County Means** | | **Avg. County Standard Errors** | |
|---|---|---|---|---|
| **Household variables** | Actual | Synthetic | Actual | Synthetic |
| Household size | 2.12 | 2.12 | 0.02 | 0.01 |
| Sampling weight | 9.99 | 10.20 | 0.11 | 0.11 |
| Number of bedrooms | 2.88 | 2.82 | 0.01 | 0.01 |
| Electricity cost/month | 118.89 | 119.37 | 1.25 | 1.10 |
| Number of rooms | 3.23 | 3.18 | 0.02 | 0.02 |
| Income | 67983.89 | 67382.38 | 1067.29 | 692.56 |
| Tenure: Mortgage/loan (%) | 49.00 | 47.03 | 0.82 | 0.74 |
| Tenure: Own free & clear (%) | 31.12 | 30.37 | 0.77 | 0.72 |
| Tenure: Rent (%) | 19.88 | 22.60 | 0.63 | 0.63 |

# Average ACS County-Level Regression Estimates

| | **Avg. County Coefficients** | | **Avg. County Standard Errors** | |
|---|---|---|---|---|
| ***Linear regression of household income on*** | Actual | Synthetic | Actual | Synthetic |
| Intercept | 24.34 | 24.26 | 1.11 | 1.09 |
| Household size | 1.52 | 1.44 | 0.14 | 0.14 |
| Sampling weight | -0.04 | -0.05 | 0.24 | 0.26 |
| Number of bedrooms | 1.15 | 1.23 | 0.19 | 0.18 |
| Electricity cost/month | 0.99 | 1.04 | 0.18 | 0.17 |
| Number of rooms | 1.25 | 1.26 | 0.14 | 0.13 |
| Tenure: Mortgage/loan | Ref | Ref | Ref | Ref |
| Tenure: Own free & clear | -3.47 | -3.05 | 0.37 | 0.34 |
| Tenure: Rent | -6.01 | -6.84 | 0.44 | 0.47 |

# Synthetic Data Inference

- Estimating a scalar quantity $Q$ is achieved using standard combining rules (Raghunathan, Reiter, and Rubin, 2003)
- Point estimate $\bar{q}_M$ is obtained by averaging point estimates across the $M = (l = 1,2, \ldots, M)$ synthetic data sets,

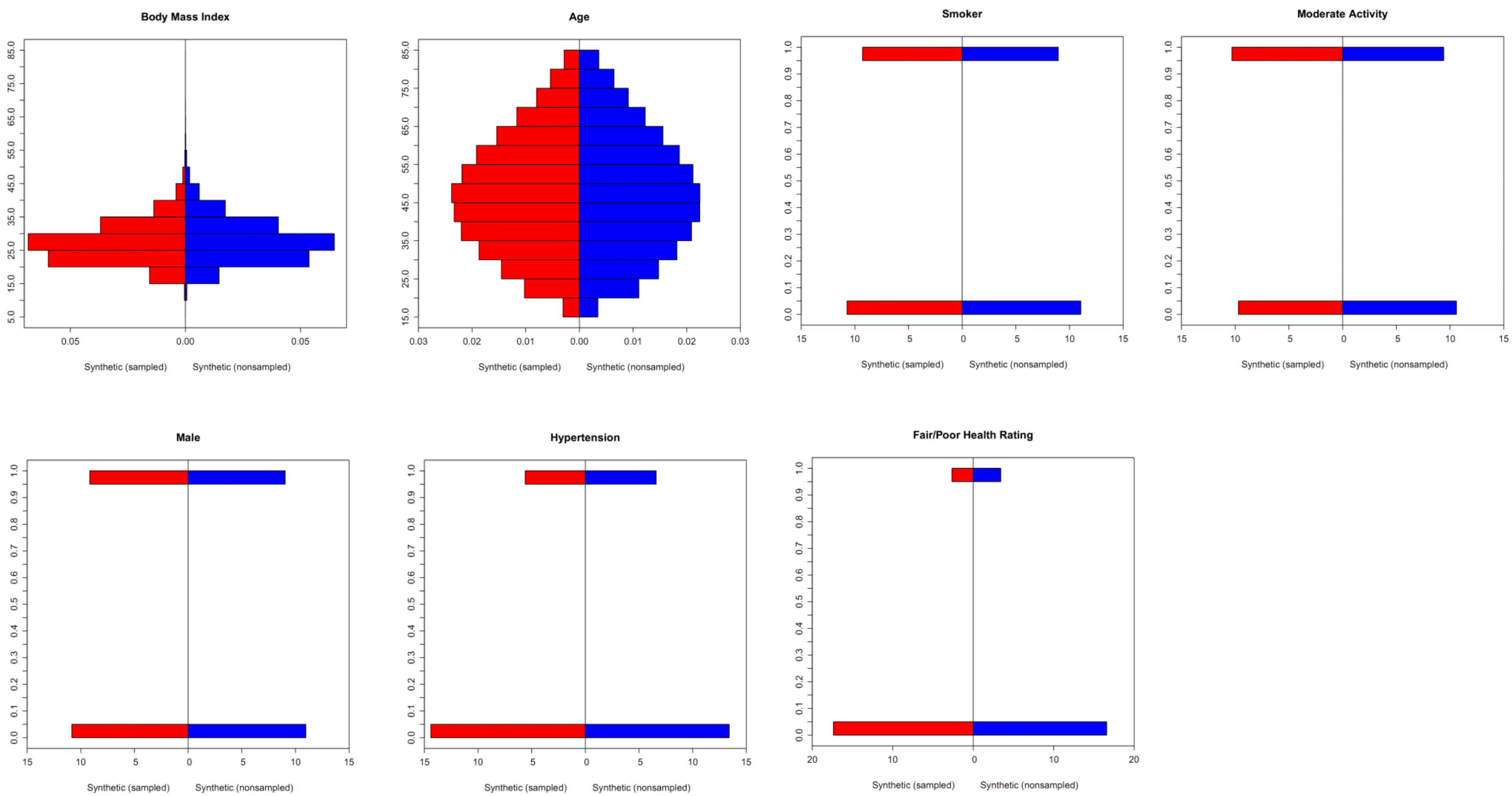
$$\bar{q}_M = \sum_{l=1}^{M} q^{(l)} / M$$

- Variance of point estimate $T_M$ consists of within- and between-variance components,

$$T_M = (1 + M^{-1}) b_M - v_m$$

  where $b_M = \sum_{l=1}^{M} \left( q^{(l)} - \bar{q}_M \right)^2 / (M-1)$ and $\bar{v}_M = \sum_{l=1}^{M} v^{(l)} / M$
- When $n$, $n_{syn}$, and $M$ are large, inferences for scalar $Q$ can be based on normal distributions.
  - For moderate $M$, inferences can be based on $t$-distributions

# Synthetic (samp) vs. Synthetic (nonsamp)



# Propensity Score Balance Check

- Actual and synthetic data sets stacked
- Fit logistic regression of belonging to actual data set
- Predicted probabilities sorted, grouped into deciles
- $\chi^2$ -test of equality of synthetic data proportions across deciles

| | **Mean** | **Min** | **Max** |
|---|---|---|---|
| **Estimated probabilities** $\hat{p}$ | 0.30 | 0.18 | 0.48 |
| $\chi^2$ **statistic** | 14.80 | 7.92 | 42.90 |
| **P-value** | 0.23 | 0.01 | 0.57 |