GOV 1000/ 2000e/ 2000/ E-2000 Section 9

Solé Prillaman

Harvard University

October 31, 2013

Logistics - This Week

Problem Set 7 - Due by 1pm on Tuesday to course dropbox.

Reading Quiz - Due by 9pm on Sunday on course website.

What have we covered?

- Summarizing and describing data: both univariate and multivariate populations (no uncertainty)
- Sampling as source of randomness and uncertainty
  - Probability / random variables
  - Sample statistics and sampling distributions
- Revisit regression in context of sampling
  - Standard errors, hypothesis testing, and confidence intervals for estimated regression coefficients
- Verifying, diagnosing, and correcting for assumptions of OLS regression
- Dealing with **missing data**

Where are we going?

Causal inference is a missing data problem!

Credit Card Expenditures Data

Let's return to our data on credit card expenditures but with missing data: `ccarddata_missing.csv`.

- Outcome variable: credit card expenditure
- Covariates:
  - Age
  - Household income (monthly in thousands of dollars)
  - Dummy for home ownership

Missing Data

- Missing observations of *outcome* variable only

Understanding Missingness in Our Data

The first 6 observations in our dataset are:

ccexpend	age	income	homeowner
124.98	38	4.52	1
	33	2.42	0
15.00	34	4.50	1
	31	2.54	0
546.50	32	9.79	1
92.00	23	2.50	0

Missing values shaded in orange.

Understanding Missingness in Our Data

How was this missingness generated?

We can characterize the missingness mechanism as:

- **Missing completely at random (MCAR)**: missingness purely random; unrelated to variables in data or any unobserved variables
- **Missing at random (MAR)**: missingness related to *observed* data
- **Not missing at random (NMAR)**: missingness related to *unobserved* data

Understanding Missingness in Our Data

How can we characterize the missingness in our credit card data?

We can break the data into two groups: observations with missing values and observations without missing value.

Then we can summarize our data for each of these groups *separately*. What we want is balance - for our missing data group to look the same as our non-missing data group.

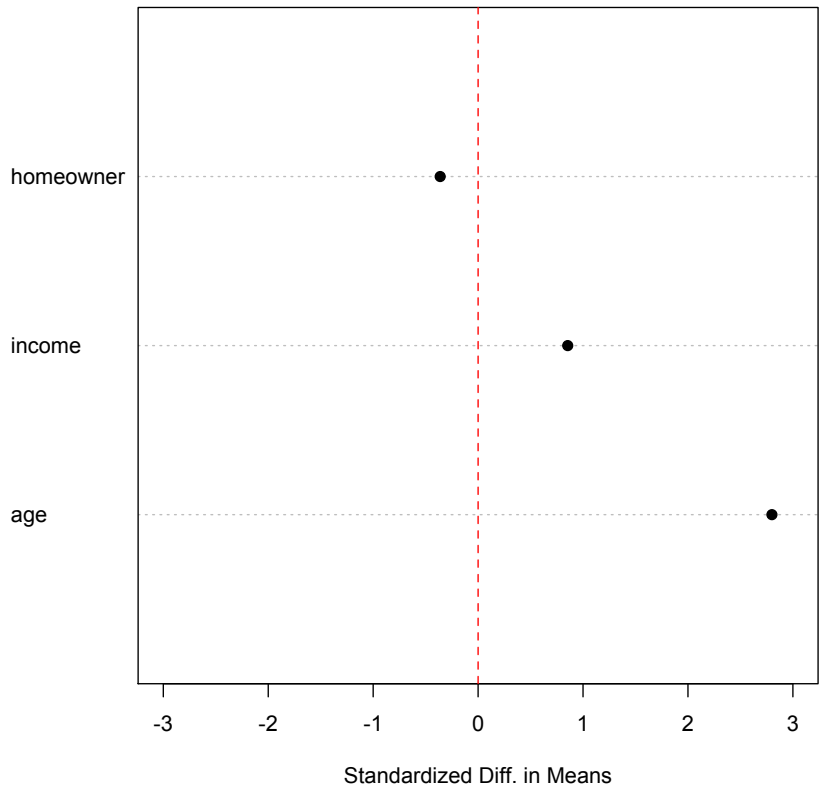
What would we expect if we believe our missingness is **MCAR? MAR? NMAR?**

Understanding Missingness in Our Data

	$\bar{X}_{\text{missing}}$	$\bar{X}_{\text{non-missing}}$	$\bar{X}_{\text{missing}} - \bar{X}_{\text{non-missing}}$	t-stat
`age`	33.68	29.13	4.54	2.80
`income`	3.62	3.27	0.34	0.85
`homeowner`	0.35	0.39	-0.04	-0.36

Missingness in Our Data

We can also look at the missingness mechanism graphically:



Does it look like we have balance across the covariates?

Dealing with Missingness

A few common ways to deal with missingness:

- Complete case analysis (listwise deletion)
- Mean imputation
- Regression imputation
- Multiple imputation (Gov 2001)

Complete Case Analysis

ccexpend	age	income	homeowner
124.98	38	4.52	1
	33	2.42	0
15.00	34	4.50	1
	31	2.54	0
546.50	32	9.79	1
92.00	23	2.50	0

In R:

```
# R automatically row-deletes observations with missing
    data
lm(ccexpend ~ income + homeowner + age, data=cc.missing)
```

Complete Case Analysis

What are the consequences of complete case analysis if we have:

MCAR? Unbiased inference but fewer observations

MAR? Unbiased inference *if* missingness covariates included (**ignorable**) but fewer observations

NMAR Possibly biased inference and fewer observations

Main Concern: Likely to induce bias unless MCAR.

Mean Imputation

ccexpend	age	income	homeowner
124.98	38	4.52	1
$\bar{y}$	33	2.42	0
15.00	34	4.50	1
$\bar{y}$	31	2.54	0
546.50	32	9.79	1
92.00	23	2.50	0

In R, mean of outcome is:

```
mean(cc.missing$ccexpend, na.rm=TRUE)
# mean is 209.4542
```

Mean Imputation

ccexpend	age	income	homeowner
124.98	38	4.52	1
209.45	33	2.42	0
15.00	34	4.50	1
209.45	31	2.54	0
546.50	32	9.79	1
92.00	23	2.50	0

Mean Imputation

What are the consequences of regression imputation if we have:

MCAR? Unbiased inference

MAR? Possibly biased inference

NMAR Possibly biased inference

Main Concern: Distorted distribution of imputed variable, unbiased mean but underestimated standard deviation, attenuated correlations. Better for univariate than multivariate statistics.

Regression Imputation

ccexpend	age	income	homeowner
124.98	38	4.52	1
$\hat{y}_2$	33	2.42	0
15.00	34	4.50	1
$\hat{y}_4$	31	2.54	0
546.50	32	9.79	1
92.00	23	2.50	0

In R, we can predict missing values:

```
lm.cc.missing <- lm(ccexpend ~ income + homeowner + age,
     data=cc.missing)
missing.df<- cc.missing[is.na(cc.missing$ccexpend),c("
    income","homeowner","age")]
predict(lm.cc.missing,missing.df)
```

Regression Imputation

ccexpend	age	income	homeowner
124.98	38	4.52	1
94.41	33	2.42	0
15.00	34	4.50	1
109.15	31	2.54	0
546.50	32	9.79	1
92.00	23	2.50	0

Regression Imputation

What are the consequences of regression imputation if we have:

MCAR? Unbiased inference

MAR? Unbiased inference

NMAR Possibly biased inference

Main concern: Does not account for uncertainty around fitted value, i.e. residual for imputed observations will always be 0.

Questions?

Happy Halloween!