Introduction to Statistical Methods
SOC-GA 2332

Lecture 12
Reliability, Weighting, and Missing Data

Siwei Cheng

Today's Topics

- Reliability and how to measure it.
- What are we weighting for?
- Missing data.

Reliability

- Reliability is closely related to measurement errors.
- Our measured value of a variable is a combination of the "true value" plus "measurement error":

$$\text{Observed value} = \text{True value } + \text{Error} \tag{1}$$

Reliability and Validity

- We can think of two types of errors: **systematic** error and **idiosyncratic** error.
- Systematic error occurs when the observed value is "**biased**" towards a certain direction.
- For example, when interviewing hiring agents for a company about discriminatory practices in the hiring process, *social desirability bias* may lead to an understatement of the actual discriminatory practices. Another example is the under-reporting of "unintended pregnancy" in retrospective surveys.
- Systematic errors raise questions as to whether the measurement is *valid*.

Reliability and Validity

- On the other hand, idiosyncratic error affects individuals or other cases in unique ways that are unlikely to be repeated in just the same way. For example, people's reporting of hours of Internet usage per day in surveys may contain large errors in recalling and reporting.

- Large idiosyncratic errors will cause problems to the reliability of the measure.

Reliability

- Alwin (2007) provides a helpful illustration of the sources of errors in surveys:

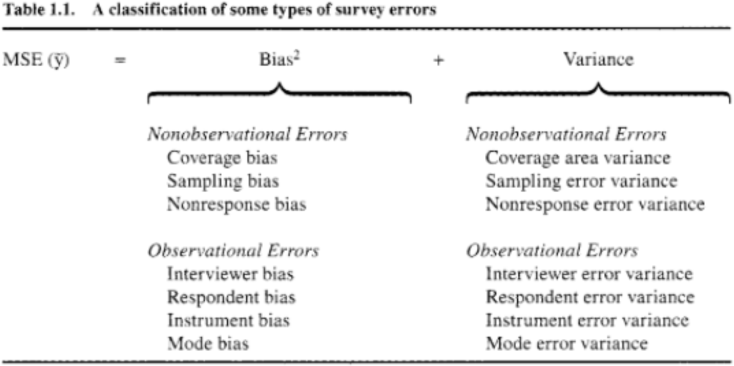


Table 1.1. A classification of some types of survey errors

MSE ($\bar{y}$)	=	Bias^2	+	Variance
		Nonobservational Errors		*Nonobservational Errors*
		Coverage bias		Coverage area variance
		Sampling bias		Sampling error variance
		Nonresponse bias		Nonresponse error variance
		Observational Errors		*Observational Errors*
		Interviewer bias		Interviewer error variance
		Respondent bias		Respondent error variance
		Instrument bias		Instrument error variance
		Mode bias		Mode error variance

Review: Measurement Error in the Independent Variable

- Why do we care about measurement errors?
- In earlier lectures, we have shown that measurement errors in explanatory variables cause an important concern in regression analysis.
- Let's consider a "true" regression model:

$$Y = \beta_0 + \beta_1 X^* + \epsilon. \tag{2}$$

- Let's assume that there is a measurement error in X:

$$e_x = X - X^* \tag{3}$$

- Let's further assume that the measurement error is uncorrelated with the true value of X:

$$Cov(e_x, X^*) = 0 \tag{4}$$

- It thus follows that $Var(X) = Var(X^*) + Var(e_x)$.

Review: Measurement Error in the Independent Variable

- Now, let's plug Equation (3) into Equation (2):

$$\begin{aligned} Y &= \beta_0 + \beta_1(X - e_x) + \epsilon. && (5) \\ &= \beta_0 + \beta_1 X + (\epsilon - \beta_1 e_x) && (6) \end{aligned}$$

- As you can see, the measurement error becomes part of the error term in the equation.

Review: Measurement Error in the Independent Variable

- What would be the OLS estimator for β_1 in this case?

$$\begin{aligned}\hat{\beta}_1 = \frac{Cov(Y,X)}{Var(X)} = \frac{Cov(\beta_0 + \beta_1 X + (\epsilon - \beta_1 e_x), X)}{Var(X)} \\ = \frac{\beta_1 Var(X) - \beta_1 Cov(X, e_x)}{Var(X)} \\ = \beta_1 \cdot (1 - \frac{Cov(X, e_x)}{Var(X^*) + Var(e_x)})\end{aligned} \tag{7}$$

- Note that $Cov(X, e_x) = Cov(X^* + e_x, e_x) = Var(e_x)$.
- It thus follows that:

$$\hat{\beta}_1 = \beta_1 \cdot (1 - \frac{Var(e_x)}{Var(X^*) + Var(e_x)}) \tag{8}$$

Review: Measurement Error in the Independent Variable

- Or, equivalently:

$$\hat{\beta}_1 = \beta_1 \cdot \frac{Var(X^*)}{Var(X^*) + Var(e_x)} \tag{9}$$

- What does this mean? This means that the "estimated" $\hat{\beta}_1$ is always closer to zero than is the true β_1!
- This is called the **"attenuation bias"** in the OLS estimator due to the measurement error in the explanatory variable.
- $\frac{Var(X^*)}{Var(X^*)+Var(e_x)}$ can be seen as the "signal-to-noise" ratio.
- If the variance of X is very large relative to the variance in the measurment error, then the inconsistency in OLS will be small.

Reliability

- How to quantify reliability? A recent article by Hout and Hastings (2016) on the reliability of the core items in the General Social Survey is a good example to look at.
- The authors summarized two ways to quantify reliability (or unreliability) of a GSS survey item.
- The first is related to the correlation of items for the same person. Consider the case of a two-time measurement, and first and second measurements are denoted by y_1 and y_2.
- If we assume the two times of measurement are capturing the same "true" Y value, then we know that the difference between y_1 and y_2 must be due to measurement error, and the common part of these two measures must be due to their association with the "true value."

Reliability: Example

- Furthermore, we can use the product of the standard deviations of y_1 and y_2 as an estimate of the total variance $\sigma_Y^2 + \sigma_\epsilon^2$.
- Then, we can write our estimate of reliability as:

$$\rho = \frac{Cov(y_1, y_2)}{s_1 s_2} \tag{10}$$

- In cases with three-time measurements, the reliability can be written as:

$$\rho = \frac{Cor(y_1, y_2) Cor(y_2, y_3)}{Cor(y_1, y_3)} \tag{11}$$

Reliability: Example

- Their second way that the authors adopted to quantify reliability is based on a **multilevel model**, which gives a complementary decomposition of the variance in repeated survey items.
- Recalled that in earlier lectures, we discussed the classic measurement model:

$$y_{it} = Y_i + \epsilon_{it} \tag{12}$$

- They also accommodate change over time by adding dummy variables for years:

$$y_{it} = Y_i + \Sigma_{t=2008}^{2014} \tau_t + \epsilon_{it} \tag{13}$$

Reliability: Example

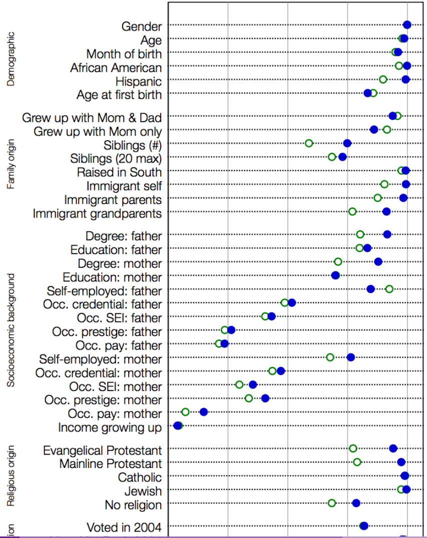
- We can think of the **between-person** variance as the variance of the true score in the population, and **within-person** variance as the variance of the measurement error.
- Hence, the reliability can be written as:

$$\rho = \frac{\hat{\sigma}_Y^2}{\hat{\sigma}_Y^2 + \hat{\sigma}_\epsilon^2} \tag{14}$$

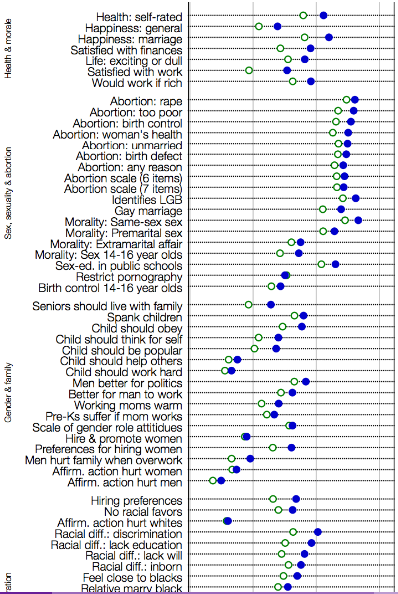
- For example, reliability is high when a large share of the total variance is due to between-person differences.

Reliability: Example

- See some examples from their findings:



Reliability: Example



What are we weighting for?

What are we weighting for?

- Something we have not yet covered in the previous lectures is **weighting**.
- When you set out to conduct your own quantitative research (or when you review others' work), you may sometimes find it a little bit confusing when it comes to weighting, because people use weighting (the term itself, and the method) in very different ways.
- A paper by Solon, Haider, and Wooldridge (2015) tries to dispel confusion and dismay by clarifying the issues surrounding weighting.
- Let's take some of their discussion as our starting point.

What are we weighting for?

- First, when estimating population descriptive statistics, weighting is needed to make the analytic sample representative of the target population.
- For example, the Panel Study of Income Dynamics oversamples low-income households. Without any weighting to adjust for this oversampling, one would estimate the U.S. poverty rate at 26 percent. The weighted poverty rate is 12 percent (Solon et al 2015).

Example: Mturk Sample versus Knowledge Networks Sample

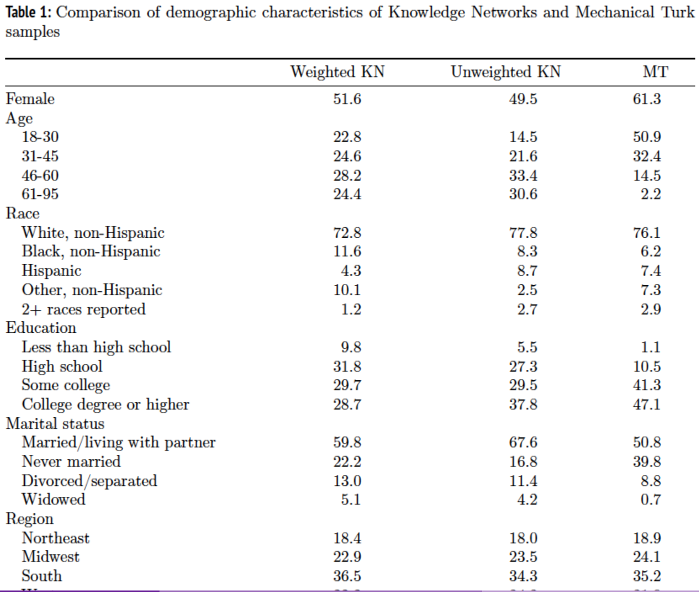


Table 1: Comparison of demographic characteristics of Knowledge Networks and Mechanical Turk samples

	Weighted KN	Unweighted KN	MT
Female	51.6	49.5	61.3
Age			
18-30	22.8	14.5	50.9
31-45	24.6	21.6	32.4
46-60	28.2	33.4	14.5
61-95	24.4	30.6	2.2
Race			
White, non-Hispanic	72.8	77.8	76.1
Black, non-Hispanic	11.6	8.3	6.2
Hispanic	4.3	8.7	7.4
Other, non-Hispanic	10.1	2.5	7.3
2+ races reported	1.2	2.7	2.9
Education			
Less than high school	9.8	5.5	1.1
High school	31.8	27.3	10.5
Some college	29.7	29.5	41.3
College degree or higher	28.7	37.8	47.1
Marital status			
Married/living with partner	59.8	67.6	50.8
Never married	22.2	16.8	39.8
Divorced/separated	13.0	11.4	8.8
Widowed	5.1	4.2	0.7
Region			
Northeast	18.4	18.0	18.9
Midwest	22.9	23.5	24.1
South	36.5	34.3	35.2

What are we weighting for?

- In your own research, if you are trying to provide descriptive statistics that represent a target population, it will be helpful to report weighted statistics.
- It is also helpful to check for yourself whether and in what direction do the numbers differ in your weighted sample and unweighted sample.
- Indicate in your footnote of the descriptive statistics table whether your statistics are weighted or unweighted, and if possible, how is it weighted.

What are we weighting for?

- **Second**, weighting can be used to correct for some important problems in making causal inferences.

1. Weighting can be used to correct for heteroskedasticity. This will generally increase the precision of estimates (unless in some specific cases it could decrease precision, see Solon et al (2015) for details).

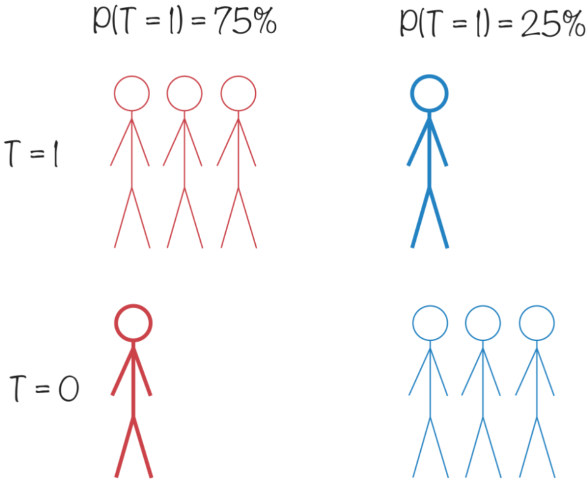
   For example, when estimating the effect of unilateral divorce laws on state-level divorce rate, some researchers weighted the sample by state/year population in the estimation as correction for population -size-related heteroskedasticity in the state/year error terms.

What are we weighting for?

2. Weighting can be used to correct for endogenous sampling (selection). Some researchers will weight the sample by the **inverse** probability of being selected into the treatment (control) group. (consider the example of college premium: they will weight the sample by the inverse probability of obtaining a college degree).

- **Intuition:** we want to weigh the sample by the inverse probability of being treated, because individuals who were assigned to the treatment group even though they were much more likely to be assigned to the control group are a rare, and valuable breed.

Inverse Probability Weighting: Illustration



What are we weighting for?

3. Weighting can be used to account for heterogeneous treatment effect and for estimating different average partial effects.

- For example, if we know the effect of becoming a father on men's wages depends of whether the father and the child are living together, but our data oversampled coresident fathers. In this case, we may want to reweight the sample so that our estimated effect of fatherhood on earnings is representative of the population of fathers.

Missing Data

Missing Data

- Missing data are regrettably common among data sets used by social scientists.
- In almost all previous lectures, we have basically assumed that there are no missing data involved, or deleted cases in which one or more variables are missing.
- Here, we are going to think more carefully about missing data. We will start with a discussion on the basic distinctions concerning the processes that generate missing data.
- Then, we will move on to very briefly introduce methods for dealing with missing data.

Missing Data

- There are three basic types of missing data:
  1. **MCAR (missing completely at random)**: the missing data can be regarded as a simple random sample of the complete data.
  2. **MAR (missing at random)**: Missingness is random conditional on the observed data. For example, income can be missing at random if conditional on observed covariates, such as education, occupation, gender, and part-time/full-time work status, the decision to report income is independent of income itself.
  3. **MNAR (missing not at random), or nonignorable**: the probability that a data value is missing depends on missing data, even when the information in the observed data is taken into account. For example, conditional on the observed variables, individuals with higher incomes are more likely to withhold information in household surveys.

Dealing with Missing Data

- How to deal with missing data?
- Let's begin with some traditional methods.
- When we consider these methods, we ask whether the method provides unbiased/consistent estimates of population parameters and whether the method use the observed data efficiently.
- There are three traditional methods for dealing with missing data.

Dealing with Missing Data

1. **Complete-case analysis** (listwise/casewise deletion): simply deleting all cases where any of the variables is missing. (pro: simplicity; con: discards some valid data and does not use the information in the data efficiently).
2. **Available-case analysis**: using all nonmissing observations to compute each statistic of interest.
3. **Filling in the missing data by**: *(a) unconditional mean imputation* (making the distribution less variable and tends to weaken relationships between variables); *(b) conditional-mean imputation* (but imputed observations will still tend to be less variable than real data because they lack residual variation, and also we failed to account for **uncertainty** in the estimation of the regression coefficients used to obtain the imputed values).

Missing Data

- Next, we move to a brief introduction of more advanced methods in dealing with missing data: Maximum-Likelihood Estimation for MAR; Multiple Imputation for MAR; and models for non-ignorable selection.
- We briefly mentioned the maximum-likelihood estimation earlier in this course. Here, maximum likelihood can be applied to parameter estimation in the presence of missing data.

Maximum-likelihood (ML) estimation for MAR

- The first method, Maximum-likelihood (ML) estimation requires making **assumptions** about the distribution of the complete data and about the process producing missing data.
- This is reflected in the joint distribution of X_{obs} and X_{mis}.
- Let $p(X;\theta) = p(X_{obs}, X_{mis};\theta)$, where X_{obs} and X_{mis} represent the observed missing components in the data, respectively.
- If MAR, the ML estimate $\hat{\theta}$ of θ can be obtained from the marginal distribution of the observed data.

$$p(X_{obs};\theta) = \int p(X_{obs}, X_{mis}, \theta) dX_{mis} \tag{15}$$

Missing Data

- That is, **the marginal distribution of observed data is integrated over the missing data.** Parametric assumptions are needed to specify how the observed and missing components are associated.
- The expectation-maximization (EM) algorithm can be used to generate $\hat{\theta}$ (not covering in this course).

Multiple Imputation (MI) for MAR

- The second method, (Bayesian) multiple imputation, or MI for short, also deals with missing data under MAR.
- Essential idea of multiple imputation:

1. It reflects the **uncertainty** associated with missing data by imputing several values for each missing value, each imputed value drawn from the **predictive distribution** of the missing data. Therefore, it produces not one but **several** completed data sets.
2. Point estimates of the population regression coefficients are obtained by averaging across g imputations:

$$\tilde{\beta}_j = \frac{\sum_{l=1}^{g} B_j^l}{g} \tag{16}$$

Missing Data

3. The standard errors of the estimated coefficients are obtained by combining information about within - and between-imputation variation in the coefficients:

$$\tilde{S}E(\tilde{\beta}_j) = \sqrt{V_j^W + \frac{g+1}{g} V_J^B} \tag{17}$$

and the within-imputation component is:

$$V_j^W = \frac{\Sigma_{l=1}^{g} SE^2(B_j^l)}{g} \tag{18}$$

where the between-imputation component is:

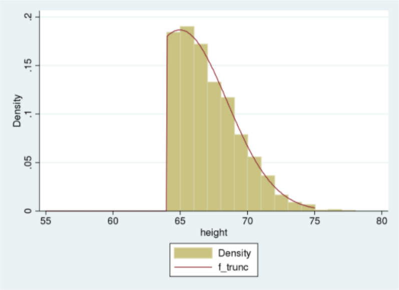
$$V_j^B = \frac{\Sigma_{l=1}^{g} (B_j^l - \tilde{\beta}_j)^2}{g-1} \tag{19}$$

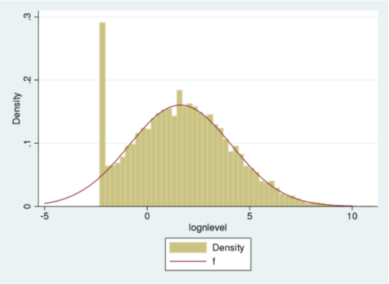
Missing Data

- Lastly, we move on to the more complicated case of MNAR — that is, the selection into missing is **not independent** of the value of Y.
- For example, (1) only wages above minimum wage are observed; (2) higher-wage workers are more likely to be non-respondents.
- There is a large and ongoing literature on addressing MNAR, and we are not going to the details in this course. But we will briefly introduce the **selection** problem.

Truncated and Censored Distributions

- The key difference between truncated- and censored-normal distributions:





Missing Data

- For **truncated data**, there is no observation below the truncation point *a*, so the density function p(z) of Z still must enclose an area of 1, and so this density is given by:

$$p(z) = \frac{\phi(z)}{1 - \Phi(a)} \text{ , for} z \geq a \tag{20}$$

- For **censored data**, the area to the left of the threshold a will have zero density, but the distribution will have a large density at the threshold:

$$Z = \begin{cases} a, if & \xi <= a \\ \xi, if & \xi > a \end{cases} \tag{21}$$

Missing Data

- In analyzing truncated or censored data, the censored-regression model (or truncated regression model) basically taking into account the likelihood of falling below or above the censoring threshold. The log-likelihood therefore comprises terms for the censored and fully-observed observations.
- Another method for dealing with truncated/centored data is Heckman's two-stage selection model, which comes from economist James Heckman.

Missing Data

- The Heckman selection model (Heckman 1977) consists of two parts:
  1. A *regression equation* for a latent response variable ξ (e.g. potential earnings):

$$\begin{aligned}\xi_i &= \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + \epsilon_i && (22)\\ &= \eta_i + \epsilon_i && (23)\end{aligned}$$

  2. A *selection equation* that determines whether or not ξ is observed:

$$\begin{aligned}\zeta_i &= \gamma_0 + \gamma_1 Z_{i1} + \gamma_2 Z_{i2} + ... + \gamma_p Z_{ip} + \delta_i && (24)\\ &= \psi_i + \delta_i && (25)\end{aligned}$$

Missing Data

- Hence, the *observed response variable* is:

$$Y_i = \begin{cases} missing, if \quad \zeta_i <= 0 \\ \xi, if \quad \zeta_i > 0 \end{cases} \tag{26}$$

- It is assumed that the two error terms, ϵ and δ, follow a bivariate-normal distribution with zero means.
- We also assume the variance of δ is one, and the correlation between the two error terms is ρ.
- Because of the correlation of the two error variables, the missing data are not ignorable.

Missing Data

- After a few steps, it can be shown that:

$$E(Y_i|\zeta > 0) = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + \beta_\lambda \lambda_i + \nu_i \tag{27}$$

  where $\beta_\lambda = \sigma_\epsilon \rho$ and $\lambda_i = m(-\psi_i) = \frac{\phi(-\psi_i)}{1-\Phi(-\psi_i)}$ is called the Inverse Mills Ratio.

- If the errors from the regression and selection equations are uncorrelated, then $\beta_\lambda = 0$, and ignoring λ is inconsequential.

- There are different strategies for estimating this model. Two-step estimation, which was proposed by Heckman, basically fit a "selection equation" using a probit model first, and use the fitted values to estimate the inverse Mills ratio. Then, the inverse Mills ratio can be included as an auxiliary regression in the linear regression of Y_i on the Xs for the complete cases.