BUSN 5000

Measurement Error and Sample Selection

Chris Cornwell

Terry College of Business

Fall 2023

Section 1

Preliminaries

Week 4 :: Reading

Reading: Bueno de Mesquita and Fowler, ch 16

IRL: "How Bad Data Traps People in the US Justice System", *Ted Talks*, April 2022.

Due: Homework 4 – Mon, Sep 18

R highlights

Packages

- `kableExtra` for a general approach to making pretty tables
- `vtable` for quickly documenting your data

Functions

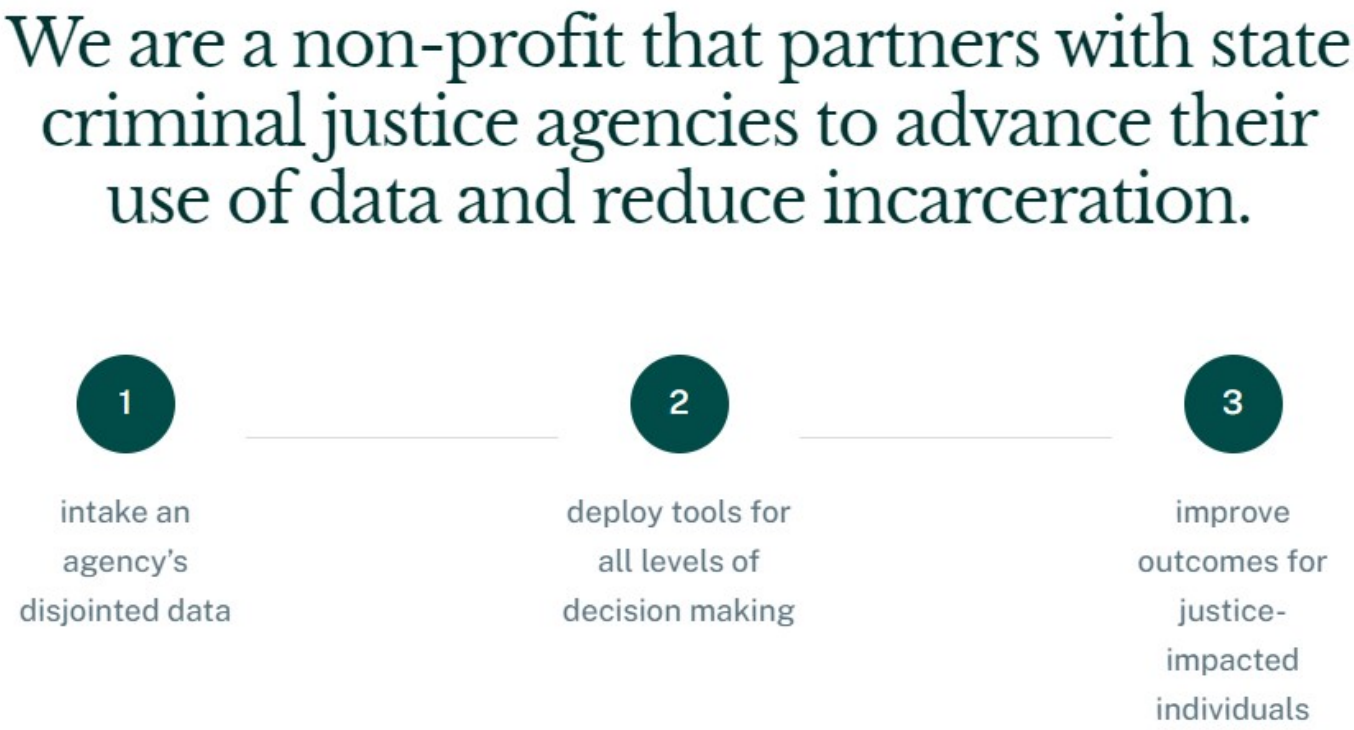
- `data` loads specified data sets and listing the available data sets
- `kable` is a simple table generator
- `st` is another way to produce tables of summary statistics

The main things

1. Measurement error model
2. Classical measurement error
3. Classification error
4. Selection mechanisms
5. MCAR
6. MAR
7. Endogenous selection
8. Imputation

ATACing criminal justice data

"Recidiviz helps criminal justice data help people."



Section 2

Class Size and Student Achievement

What do we want to learn?

We want to learn whether class size matters for student achievement:

$$E(\textit{achievement}|\textit{class size}).$$

"Much of the uncertainty in the literature derives from the fact that the appropriate specification – including the functional form, level of aggregation, relevant control variables, and identification – of the 'education production function' is uncertain. . . . Many of these specification issues arise because of the possibility of omitted variables, either at the student, class, school, or state level." (Krueger (1999))

Tennessee STAR class-size experiment

The Tennessee Student/Teacher Achievement Ratio (STAR) experiment was the largest randomized experiment of its kind at the time.

Project STAR began in the 1985-86 school year randomly assigning kindergarteners and teachers to one of three class configurations – small, regular and regular+aide – and then followed them for four years.

The sample included 11,600 students from 80 schools.

Krueger (1999) provides a comprehensive description and analysis of the experiment.

Test-score distributions by class size

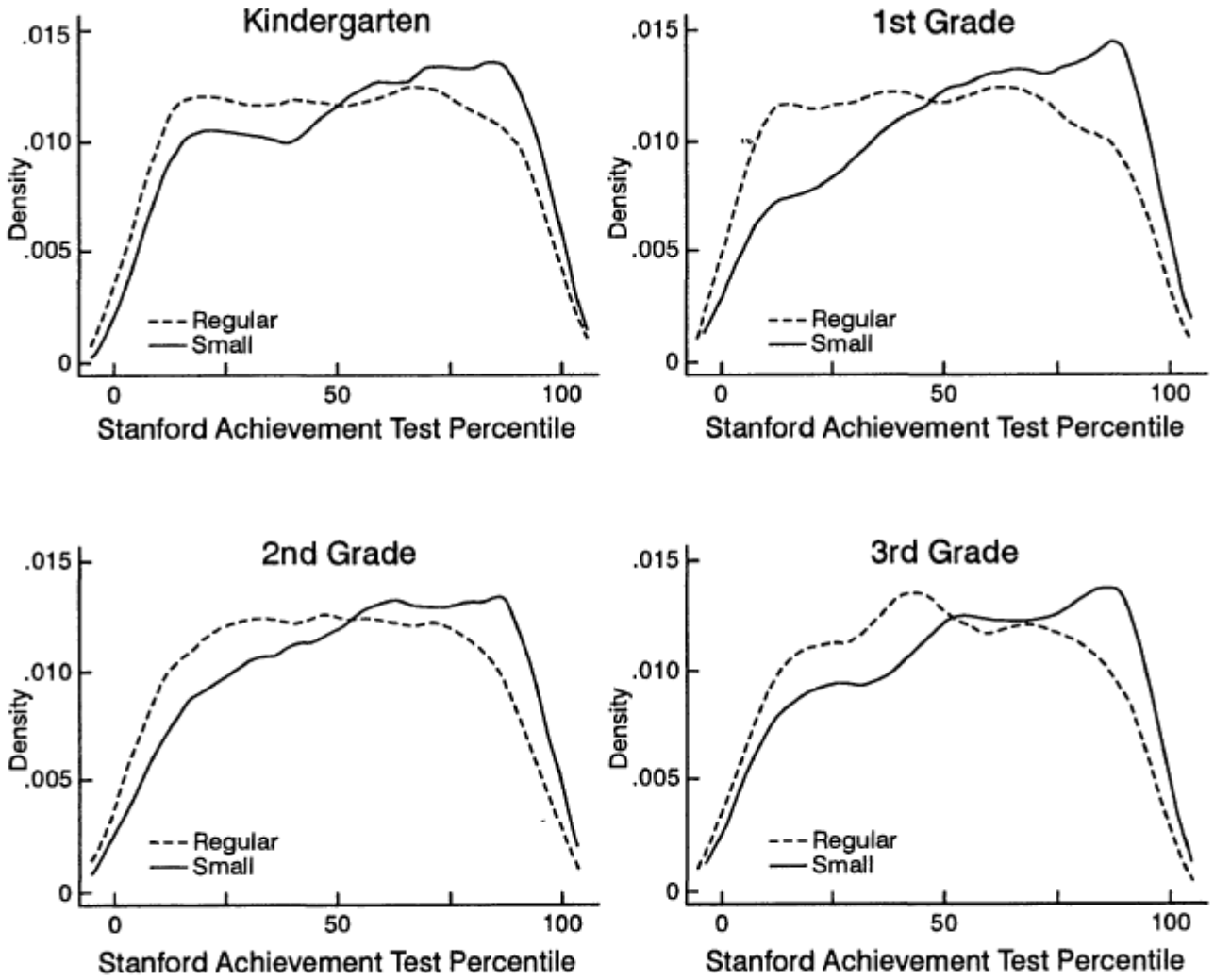


FIGURE I
Distribution of Test Percentile Scores by Class Size and Grade

The AER STAR data

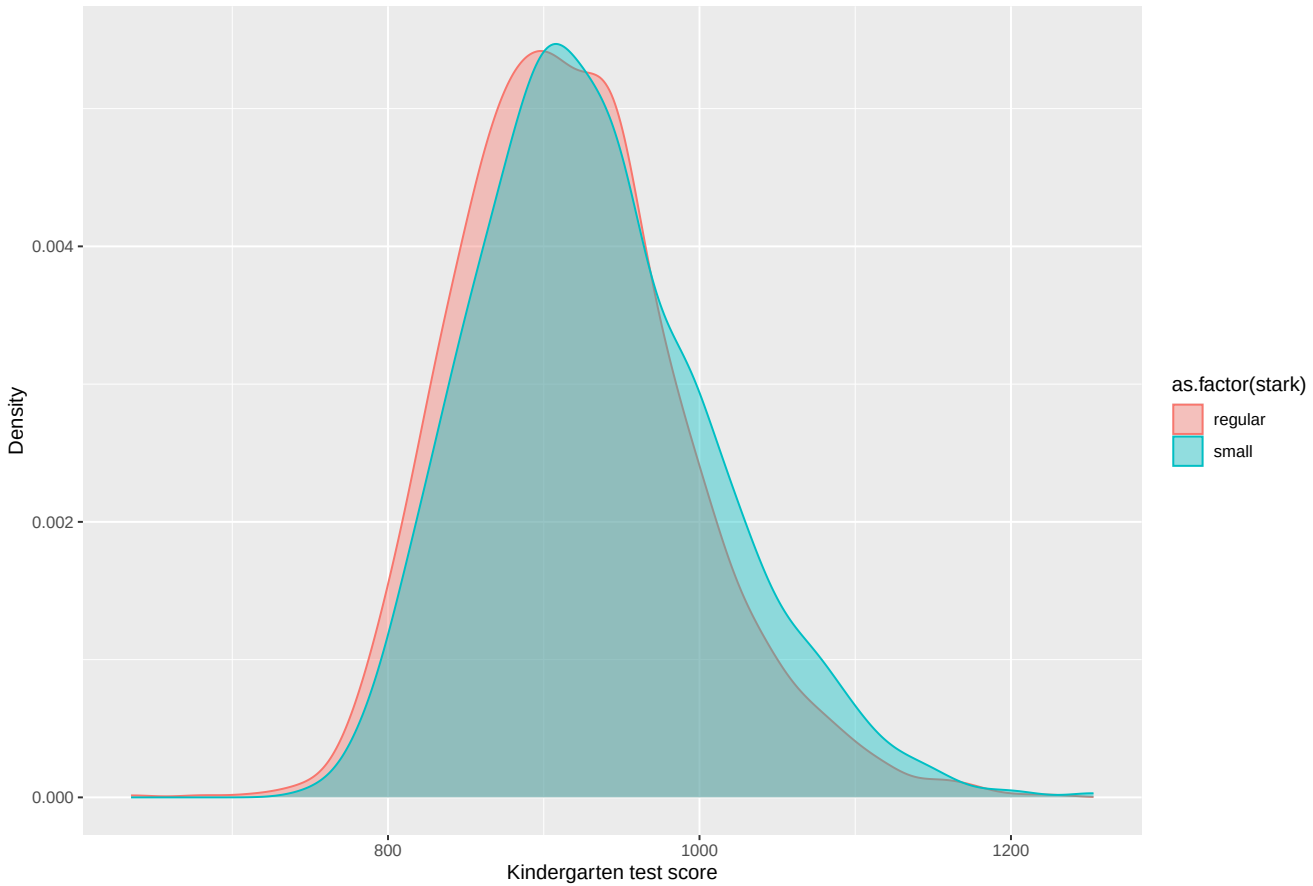
Our version of the STAR data comes from the AER package and it documented here.

```
data(STAR)
STAR2 <- STAR %>%
  mutate(
    scorek = readk + mathk,
    score1 = read1 + math1,
    score2 = read2 + math2,
    score3 = read3 + math3
    )
distk <- ggplot(data = subset(STAR2, stark=="small" | stark=="regular"),
  aes(x=scorek, fill=as.factor(stark),
      color=as.factor(stark))) +
  geom_density(alpha=0.4) +
  labs(title="Figure 1. Empirical test score distributions, kindergarten",
       x="Kindergarten test score",
       y="Density")
```

Kindergarten test score distributions

Because `AER` provides raw test scores, not the percentile measure Krueger uses, the scale of our estimated effects will be different.

Figure 1. Empirical test score distributions, kindergarten



Average kindergarten test scores by class type

```
datasummary(
  stark*scorek ~ N + Mean + SD,
  data=STAR2,
  title="Test scores by class type, kindergarten"
  )
```

Table 1: Test scores by class type, kindergarten

stark		N	Mean	SD
regular	scorek	2005	918.04	73.14
small	scorek	1738	931.94	76.36
regular+aide	scorek	2043	918.36	71.31

Section 3

Sources of Bias

Identifying "features" of the DGP

One more time – in general, we can think of an estimator being composed of three things:

$$\textit{estimator} = \textit{estimand} + \textit{bias} + \textit{sampling error}.$$

We'd like to think whatever estimator we're using *identifies* the underlying estimand we care about. This is the essence of *consistency* – bias and sampling error vanish and the estimator collapses onto the estimand as sample size grows.

The foundation of statistical inference we've laid out depends on it.

Unfortunately, the real world does not always jibe with this foundation.

Bias from measurement error

Measurement error is exactly what it sounds like: sometimes the data we have are not totally accurate measurements of the phenomenon we are interested in.

For example, survey asks about weekly earnings, but respondents' reports are inaccurate. Or, you are measuring IQ, but there is test variation and normalizing is imprecise.

Measurement error can also infect *classification*. For example, students who switch classes in the Tennessee STAR experiment may be misclassified in terms of class type.

To understand the effects of measurement error, we need a model.

Bias from missing data

We can often frame bias in terms of *missing data*. (We'll see in Part II that you can even frame measurement error this way.)

Data can be missing because respondents were asked about their earnings but refused to provide it, as in a survey. This is an example of *item nonresponse*.

But more often, data are missing because it was not available to collect in the first place. For example, we see the effects of a drug trial or a training program only on people who *selected* into the drug trial or maintained their enrollment in a training program. These are examples of *unit nonresponse*.

The question, of course, is how to account for selection. Which is to say, we need a model.

Section 4

Measurement Error

Conceptualizing measurement

There is the thing we want to measure (call it the "true" thing) and there is the measurement we make. The difference, if there is any, we can call *measurement error*:

$$Measurement = Truth + Error.$$

All three should be understood as random variables.

We'll see that how they are related to each other and to other variables will determine how we think about the effects of measurement error on the estimation of estimands we want to learn about.

Modeling measurement error

Let's start with the idea that we want to learn about the relationship between a Y and an X, as captured by the CEF, $E(Y|X)$.

To learn about this relationship, we collect data on the random variables, Y and X. Now, let's say that what we measure as X is not the "true X", which we'll call X^*. A simple *model* for how X and X^* are related is

$$X = X^* + e,$$

where e represents measurement error.

So, it's really $E(Y|X^*)$ we want to learn about, but the measurement error is getting in the way. To see how, we'll need to bear down on this model a little.

Learning about $E(X^*)$

Let's first consider how measurement error affects what we can learn about the features of the distribution of X^* (the so-called "truth").

Take $E(X^*) = \mu_X$. With a random sample on X, we can estimate μ_x (the estimand) by the sample mean, which we know can be decomposed as

$$\overline{X} = \mu_X + \textit{bias} + \textit{sampling error}.$$

Now, $E(\textit{sampling error}) = 0$, so the bias created by the measurement error boils down to

$$\textit{bias} = E(\overline{X} - \mu_X) = E(X^* + e) - \mu_X = E(e).$$

It shouldn't be surprising that computing the sample average to estimate the "true" mean of the population distribution will biased unless $E(e) = 0$.

Learning about var(X^*)

It should also not be surprising that measurement error will cause bias in the estimation of the "true variance" as well:

$$\text{var}(X) = \text{var}(X^*) + \text{var}(e) + 2 \times \text{cov}(X^*, e).$$

In general, var(X) can be higher or lower than var(X^*) because cov(X^*, e) could be positive or negative.

More troublesome is the fact that if X^* and e are related, the measurement error depends on the "true X" and it is not possible to learn anything about about the distribution of X^* using the observed data.

To make progress, we will need to make assumptions about the distribution of measurement error and the joint distribution X^* and e.

The classical assumptions

The classical take on measurement error assumes

1. $E(e) = 0$
2. X^* and e are independent ($f(e|X^*) = f(e)$)

We get a lot of mileage from these assumptions. The first implies

$$E(\overline{X}) = E(X^*) = \mu_X.$$

The second implies $\text{cov}(X^*, e) = 0$ and

$$\text{var}(X) > \text{var}(X^*).$$

Learning about $E(Y|X)$

Now, let's turn to the relationship between Y and X. Using the classical assumptions, let's determine how measurement error in X affects our ability to learn about the correlation between Y and X^*:

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)}\sqrt{\text{var}(Y)}}.$$

First, note that the classical assumptions also include

3. $\text{cov}(Y, e) = 0$,

which means

$$\text{cov}(X, Y) = \text{cov}(X^*, Y).$$

Combined with the fact that $\text{var}(X) > \text{var}(X^*)$, it should be clear that $\text{corr}(X, Y)$ is biased down:

$$\text{corr}(X, Y) < \text{corr}(X^*, Y)$$

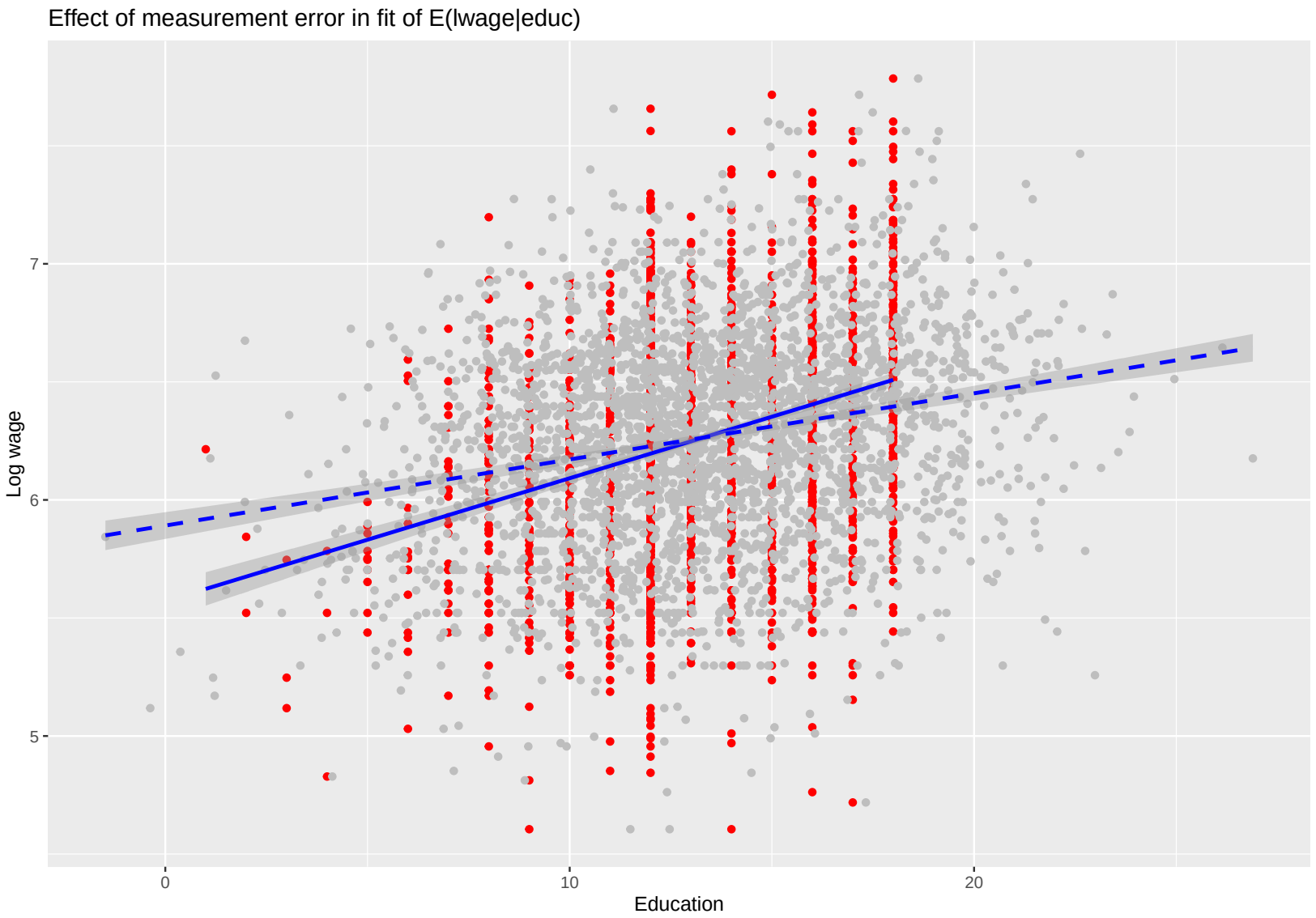
Measurement error in linear models for $E(Y|X)$

The downward bias in $\text{corr}(X, Y)$ foreshadows the effect of measurement error on linear fits of $E(Y|X)$.

We'll demonstrate this using the card data, focusing on $E(lwage|educ)$. We'll plot the "true" relationship and one where we add measurement error to the true education variable (educ).

```
card <- card %>%
  mutate(num_obs = length(lwage),
         e       = rnorm(n = num_obs, mean = 0, sd = 2.6),
         educ_e  = educ + e)
meas_err_plot <- ggplot(data=card) +
  geom_point(aes(x=educ,   y=lwage), color = "red") +
  geom_point(aes(x=educ_e, y=lwage), color = "grey") +
  geom_smooth(mapping = aes(x=educ, y=lwage),
              method=lm, se=TRUE, color="blue") +
  geom_smooth(mapping = aes(x=educ_e, y=lwage),
              method=lm, se=TRUE, linetype="dashed", color="blue") +
  labs(title="Effect of measurement error in fit of E(lwage|educ)",
       x="Education",
       y="Log wage")
```

Downward bias in the linear fit



Classification error

If X^* is categorical, then measurement error is about *misclassification* and the classical assumptions don't apply.

To avoid confusion, let's not talk in terms of X. Instead, let's define a dummy variable, D, which is coded as 1 or 0 to indicate whether the unit of observation is in a category or not. Let's take it one step further – call the category "treated" as in a treatment group of an experiment or an A/B test. This will be particularly salient in Part II.

Then, we'll call D^* the "true indicator" and write D in terms of D^* like this:

$$D = D^* + e,$$

where $e \in \{-1, 0, 1\}$. In this setup, the measurement error *must* be negatively correlated with the truth. Hence, it is *non-classical*.

Classification error in the STAR data

"As a partial check on these potential 'reactive' effects, I examined the relationship between class size and student achievement just among students assigned to regular-size classes. . . . Based on these estimates, an eight-student reduction in class size is associated with a three-to-four-percentile increase in test scores . . . And given that much of the variability in class size in the control group may be due to measurement errors (e.g., students moving in and out of class during the school year), it is noteworthy that these regressions find any evidence of class-size effects." (Krueger (1999))

Dissecting classification error

If we know something about the nature of the classification errors, we can account for it in our analysis.

Think about the Tennessee STAR context. Say D indicates whether a student was *assigned* to a small class and D^* indicates whether a student was *enrolled* in a small class. The idea is that students may switch classes, which the researcher cannot observe, creating errors in assignment.

A full analysis of the effects of the STAR program should account for the probability a student is recorded as assigned to a small class even though she is enrolled in a large one, $Pr(D = 1|D^* = 0)$ and the probability a student is recorded as assigned to a large class even though she is enrolled in a small one, $Pr(D = 0|D^* = 1)$.

The first misclassification is a *false positive* and the latter as a *false negative*.

What we need to know

By the *Law of Total Probability (LTP)* we can decompose the probability a student was assigned to a small class like this:

$$P(D=1) = P(D=1|D^*=1)P(D^*=1) + \underbrace{P(D=1|D^*=0)}_{\text{false positive}} P(D^*=0).$$

This recognizes that a student can be assigned to a small class whether they are actually enrolled or not. The second term reflects is the sort of classification error we are worried about.*false positive*

The researcher wants to know $P(D^*=1)$ but can only observe $P(D=1)$ (the share of students assigned to small classes).

As it turns out, it is possible to make progress on $P(D^*=1)$ if there is auxiliary information on the false positive and false negative rates.

Sorting it out

Let's try to sort this out. For convenience, let's adopt a simpler notation for the conditional probabilities and assemble terms in a classification table. For example, the probability of a false positive, $P(D = 1|D^* = 0)$, is denoted π_{10}.

	Assigned Small	Assigned Large	Marginal Prob
Enrolled Small	π_{11}	π_{01}	$P(D^* = 1)$
Enrolled Large	π_{10}	π_{00}	$P(D^* = 0)$
Marginal Prob	$P(D = 1)$	$P(D = 0)$	

Let's say we can find out the false positive rate, π_{10}, and false negative rate, $\pi_{01} = 1 - \pi_{11}$ from school enrollment records. Substituting into the LTP expression, and using the fact that $P(D^* = 0) = 1 - P(D^* = 1)$, we have

$$\begin{aligned} P(D = 1) &= \pi_{11}P(D^* = 1) + \pi_{10}[1 - P(D^* = 1)] \\ &= \pi_{10} + (\pi_{11} - \pi_{10})P(D^* = 1) \\ &= \pi_{10} + (1 - \pi_{01} - \pi_{10})P(D^* = 1). \end{aligned}$$

Then, it's a short step to an expression for $P(D^* = 1)$ in terms of the classification error rates:

$$P(D^* = 1) = \frac{P(D = 1) - \pi_{10}}{1 - \pi_{01} - \pi_{10}}.$$

The bottom line

Three things emerge from what we've dissected and sorted out:

1. The sample share of students assigned to small classes, $\hat{P}(D = 1)$, will be a biased estimator of $P(D^* = 1)$, with the size and sign of bias depending on the relative size of the false positive and false negative rates.
2. If you knew the false positive and false negative rates, you could get an unbiased estimator by correcting the observed share assigned to small classes for the misclassification.
3. The classification errors cause the variance of your estimate of $P(D^* = 1)$ will be greater than the variance of the estimate $P(D = 1)$.

Section 5

Sample Selection

Selected samples

When we say a sample is "selected", we generally mean one that is not randomly drawn from the population.

Selected samples can arise from a variety of *selection mechanisms*, some of which are related to sample design, while others are related the behavior of the sampled units (think nonresponse or attrition).

We should say upfront that sample selection only becomes an issue after the population of interest is specified.

For example, if we propose a model for the subset of a larger population, we are fine to randomly sample only from the subpopulation and proceed with our analysis. Not having randomly sampled from the larger population does not preclude us from conducting inference on the subpopulation.

Attrition in the STAR data

```
st(STAR2, vars=c('scorek','score1','score2','score3'),
   group='stark',
   title="Test scores by class type, all grades")
```

Table 3: Test scores by class type, all grades

stark	regular			small			regular+aide		
Variable	N	Mean	SD	N	Mean	SD	N	Mean	SD
scorek	2005	918	73	1738	932	76	2043	918	71
score1	1456	1057	91	1339	1076	95	1503	1054	91
score2	1201	1179	83	1080	1189	85	1183	1175	83
score3	1047	1247	70	937	1258	73	1021	1247	73

Modeling selection

Let's start with a Y and an X, random variables for which we will collect data by randomly sampling from the population to learn about $E(Y|X)$.

To this setup, we'll introduce a selection indicator, S, which determines whether an unit is included in the analysis.

Now, a random draw from the population for Y and X will also include S, i.e., the sample will consist of

$$\{(X_i, Y_i, S_i) : i = 1, \ldots, N\},$$

where $S_i = 1$ if unit i is included in the sample and 0 if not.

To understand the effects of selection we need to know about the distribution of S and its dependence on Y and X.

Missing completely at random

We say that data are *missing completely at random (MCAR)* if S is *independent* of Y and X.

If this is true, the probability the data are selected has nothing to do with the data itself. Technically, this means $P(S = 1|Y, X) = P(S = 1)$. In other words, the selected data and missing data are just different samples from the same distribution.

The upshot of MCAR is that we can proceed with our analysis using only the *complete cases*.

Missing at random

MCAR is a strong assumption and often not realistic.

More interesting and more realistic is the *missing at random (MAR)* case, where selection can depend on X but not other unobserved factors. In terms of the CEF, this means

$$E(Y|X,S) = E(Y|X).$$

The label MAR can be confusing because the selection process actually depends on X. Instead, you might say that selection mechanism is *exogenous*.

As you might guess, if the data are MCAR they are MAR, but not the other way around.

MAR or not?

Suppose we want to learn about $E(earnings|age)$, but say our data were based on a survey of individuals between 35-54, leaving us a nonrandom sample of the working-age population.

This is not a problem – i.e. MAR would apply – if the CEF is the same for any subset of the working-age population.

Now, suppose we want to learn about $E(lwage|IQ)$, but IQ is *more likely* missing for men with lower IQ scores. There is no fixed rule here, as in the age-based survey. Instead, the issue is that probability of selection rises with IQ score. The questions is whether $P(S = 1|IQ)$ is the whole story or whether we should be concerned that other unobserved factors could be influencing selection.

Endogenous selection

When selection is based on Y, it is *endogenous* and this is always a problem for learning about $E(Y|X)$.

Again, suppose we want to learn about $E(earnings|age)$, but our data were based on a survey of individuals who earned less that \$400,000 per year. Simply put, the CEF, $E(earnings|age)$, is not the same as the CEF conditional on earnings being less than \$400,000.

Note this is not a topcoding problem. In that case, high-earners are surveyed but their earnings are recorded at the topcode value. Topcoding is rather a *censoring* problem and is another source of bias in estimating the CEF. Fortunately, there are strategies for modeling the likelihood of a topcoded value.

Handling missing values

When a unit is included in the sample but values are missing on one or more variables, we have *item non-response*. This is different from *unit non-response* where a unit from the population is omitted entirely.

There are several ways to handle item nonresponse:

1. Drop the units with missing values (Is MCAR or MAR plausible?)
2. Impute the missing values (What is your model for imputation?)
3. Control for the missing values using regression (How to specify?)

Imputing missing values

Imputation is a process of filling in the missing values based on data you do observe.

There are several ways to do it:

- Using other data from the same survey (*relational imputation*)
- Using related information for the same unit from a different survey (*longitudinal imputation*)
- Using a statistical model (*prediction task*)

Recognizing missing values

To handle missing values, you need to be able to recognize them. There are three broad ways missing values will be evident in data:

1. There is no information for that variable for a particular observation. This will show up in R as `NA`.
2. The data have specific codes to indicate the missing information. For example, the CPS uses values of -1, -2, or -3 indicate that the respondent did not provide that information.
3. The data appear to be complete, but a second variable has "flags" that indicate whether the response for that variable was missing.

Best practices include:

- Check documentation to see how missing values are handled
- Check each variable to see what possible values it can take on
- Decide which values you want to treat as missing

CPS allocation flags

In the CPS, the main variables are typically complete (there are no missing values), but flags on the data file indicate that sometimes those apparently complete values were not reported by the respondent, but were based on some kind of imputation. On the next slide is an extract from the CPS codebook that shows how the typical "flag" variable is coded.

Missing items and imputation in the CPS

```
ALLOCATION FLAGS

                00 VALUE  NO CHANGE
                01 BLANK  NO CHANGE
                02 DON'T KNOW  NO CHANGE
                03 REFUSED  NO CHANGE
                10 VALUE TO VALUE
                11 BLANK TO VALUE
                12 DON'T KNOW TO VALUE
                13 REFUSED TO VALUE
                20 VALUE TO LONGITUDINAL VALUE
                21 BLANK TO LONGITUDINAL VALUE
                22 DON'T KNOW TO LONGITUDINAL VALUE
                23 REFUSED TO LONGITUDINAL VALUE
                30 VALUE TO ALLOCATED VALUE LONG.
                31 BLANK TO ALLOCATED VALUE LONG.
                32 DON'T KNOW TO ALLOCATED VALUE LONG.
                33 REFUSED TO ALLOCATED VALUE LONG.
                40 VALUE TO ALLOCATED VALUE
                41 BLANK TO ALLOCATED VALUE
                42 DON'T KNOW TO ALLOCATED VALUE
                43 REFUSED TO ALLOCATED VALUE
                50 VALUE TO BLANK
                52 DON'T KNOW TO BLANK
                53 REFUSED TO BLANK

ALLOCATED VALUE means the value was imputed using a statistical technique called the "Hot Deck".
ALLOCATED VALUE LONG  means the value was imputed in a previous month for the same household / person and carri
```

The details of allocation flag coding and missing value imputation can be found here.

Section 6

Back to class size effects

Testing the null of no difference in test scores

Let's start with the kindergartners and the brute force BUSN 3000 t test calculation.

$$
\begin{aligned}
t &= \frac{\bar{y}_{small} - \bar{y}_{regular}}{\text{se}(\bar{y}_{small} - \bar{y}_{regular})} \\
&= \frac{\bar{y}_{small} - \bar{y}_{regular}}{\sqrt{\frac{s^2_{small}}{N_{small}} + \frac{s^2_{regular}}{N_{regular}}}} \\
&= \frac{931.94 - 918.04}{\sqrt{\frac{76.36^2}{1738} + \frac{73.14^2}{2005}}} \\
&= \frac{13.9}{2.45} = 5.66
\end{aligned}
$$

Repeat with t.test

```
##
##  Welch Two Sample t-test
##
## data:  scorek by stark
## t = -5.6635, df = 3616, p-value = 0.00000001598
## alternative hypothesis: true difference in means between group regular a
## 95 percent confidence interval:
##  -18.710595  -9.087394
## sample estimates:
## mean in group regular   mean in group small
##              918.0429                931.9419
```

Testing the null in each grade

```
First  <- t.test(score1~stark, data=STAR2_2)
Second <- t.test(score2~stark, data=STAR2_2)
Third  <- t.test(score3~stark, data=STAR2_2)
results <- data.frame(
  Test = c("K", "1st", "2nd", "3rd"),
  Estimate  = c(diff(K$estimate), diff(First$estimate),
                diff(Second$estimate), diff(Third$estimate)),
  t  = c(K$statistic, First$statistic,
         Second$statistic, Third$statistic),
  df = c(K$parameter, First$parameter,
         Second$parameter, Third$parameter),
  p  = c(K$p.value, First$p.value, Second$p.value, Third$p.value)
  )
```

t test table

```
knitr::kable(results, caption = "Summary of t tests")
```

Table 4: Summary of t tests

Test	Estimate	t	df	p
K	13.898995	-5.663532	3616.041	0.0000000
1st	19.782214	-5.595915	2746.461	0.0000000
2nd	9.599936	-2.716076	2238.394	0.0066569
3rd	11.607821	-3.614289	1939.060	0.0003089

Is STAR a bonafide RCT?

We've highlighted:

- Classification error
- Attrition

What else?