Lecture 8: Assessing Regression Studies(II)

Introduction to Econometrics,Spring 2025

Zhaopeng Qu

Business School,Nanjing University

April 17 2025

1. Review of previous lectures
2. Internal Validity: Measurement error
3. Simultaneous Causality
4. Functional form misspecification
5. Missing Data and Sample Selection
6. Sample Selection Models
7. Sources of Inconsistency of OLS Standard Errors
8. Magnitude of β_1

Review of previous lectures

Accessing Regression Studies

- The validity of regression studies
  - **Internal** and **External** validity
  - The population and settings are studies and the **generalizability** of the results.
  - The **internal validity** of a regression study is the top priority in causal inference studies.

Internal Validity in OLS Regression

- Suppose we are interested in the causal effect of X_1 on Y and we estimate the following multiple regression model

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + ... + \beta_k X_{k,i} + u_i, i = 1, ..., n$$

- Internal validity has three components:

1. The estimators of β_1 should be **unbiased and consistent**. This is the most critical aspect.
2. Hypothesis tests and confidence intervals should have the **desired significance level** (at least 5% significant).
3. The value of β_1 should be **large enough** to be meaningful or economically significant.

Threats to Internal Validity

- Threats to internal validity:
  1. **Omitted Variables**
  2. **Misspecification**
  3. **Measurement Error**
  4. **Simultaneous Causality**
  5. **Missing Data and Sample Selection**
  6. **Heteroskedasticity and/or Correlated error terms**
  7. **Significant coefficients or marginal effects**
- In a narrow sense,
  - **Internal Invalidity = endogeneity in the estimation** which is caused by the above 1-5 threats.
- In a broad sense,
  - **Internal Invalidity = 1-5 threats + 6-7 threats**

OLS Regression Estimators in partitioned regression

- OLS estimator in Multiple OLS

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + ... + \beta_k X_{i,k} + u_i, i = 1, ..., n$$

- The OLS estimator of β_j is

$$\hat{\beta}_j = \frac{\sum_{i=1}^{n} \tilde{X}_{ij}, Y_i}{\sum_{i=1}^{n} (\tilde{X}_{ij})^2}$$

- The asymptotic OLS estimator of β_j

$$plim \hat{\beta}_j = \frac{Cov(\tilde{X}_{ij}, Y_i)}{Var(\tilde{X}_{ij})}$$

- Where $\tilde{X}_{ij}$ is the fitted OLS **residual** of regressing X_{ij} on other regressors,thus

$$X_{ij} = \hat{\gamma}_0 + \hat{\gamma}_1 X_{i1} + \hat{\gamma}_2 X_{i2} + ... + \hat{\gamma}_{j-1} X_{i,j-1} + \hat{\gamma}_{j+1} X_{i,j+1} + ... + \hat{\gamma}_k X_{i,k} + \tilde{X}_{ij}$$

the Standard Error of $\hat{\beta}$

The Variance of $\hat{\beta}_j$ under Homoskedasticity

$$Var\left(\hat{\beta}_j\right) = \sigma^2_{\hat{\beta}_j} = \frac{\sigma_u^2}{(n-1)s_j^2(1-R_j^2)}$$

- How does the variance of $\hat{\beta}_j$ change with the following factors?

Factors	symbols	$Var\left(\hat{\beta}_j\right)$
the variance of u_i	$\sigma_u^2 \uparrow$	$\uparrow$
the sample variance of X_j	$s_j^2 \uparrow$	$\downarrow$
the R_j^2	$R_j^2 \uparrow$	$\uparrow$
the sample size	$n \uparrow$	$\downarrow$

Control Variables: W

- We will discuss the potential bias or the precise of the OLS estimators when the control variable is as follows:

W	**Irrelevant** to Y	**Relevant** to Y
Uncorrelated with X	**Irrelevant Variables**	**Non-Omitted Variables**
Correlated with X	**Irrelevant Variables**	**Omitted Variables**
Highly-Correlated with X	**(Worse)Irrelevant Variables**	**Omitted Variables and Multicollinearity**

Control Variables: Guides

- **Irrelevant Variables**: Drop
- **Relevant Variables**
  - **Non-Omitted Variables**: Keep
  - **Omitted Variables**: It depends.
- **High Correlation with X**: Be cautious

Good Control v.s Bad Control

- **DAGs** can help us to identify
- Building Blocks
  - **Chains**
  - **Confounders**
  - **Colliders**
- **Good Control**: Block the backdoor path
  - Confounders
- **Bad Control** : Open the backdoor path
  - Colliders and Chains

Tips for Control Variables

1. Identify the variable's category to decide.
2. Use economic theory as a guide.
3. Use **Directed Acyclic Graphs (DAGs)** to visualize and evaluate variable relationships.
4. Gain insights from papers published in reputable journals.

Internal Validity: Measurement error

Introduction

- When a variable is **measured imprecisely**,then it might make OLS estimator biased.
- This bias persists even in very large samples, so the OLS estimator is inconsistent if there is measurement error.
- for example: recall last year's earnings

Types of Measurement errors

There are different types of measurement error

1. Measurement error in the dependent variable Y

2. Measurement error in the independent variable X(**errors-in-variables bias**)

Measurement error in the dependent variable Y

- Suppose the true population regression model(Simple OLS) is

- Because Y is measured with errors, we can not observe Y_i but observe $\tilde{Y}_i$, which is a noisy measure of Y_i,thus

- The noisy part of $\tilde{Y}_i$, ω_i, satisfies

Measurement error in the dependent variable Y

- And we can only estimate

$$\widetilde{Y}_i = \beta_0 + \beta_1 X_i + e_i$$

  where $e_i = u_i + \omega_i$
- The OLS estimate $\hat{\beta}_1$ will be **unbiased** and **consistent** because $E[e_i|X_i] = 0$
- Nevertheless,the estimate will be less precise because

- Measurement error in Y is generally less problematic than measurement error in X.

Measurement error in X: classical measurement error

- The true model is

- Due to the **classical measurement error**,we only have X_{1i}^* thus $X_{1i}^* = X_{1i} + w_i$,we have to estimate the model is

- where $e_i = -\beta_1 w_i + u_i$

Measurement error in X: classical measurement error

- Similar to OVB bias in simple OLS model

Measurement error in X: classical measurement error

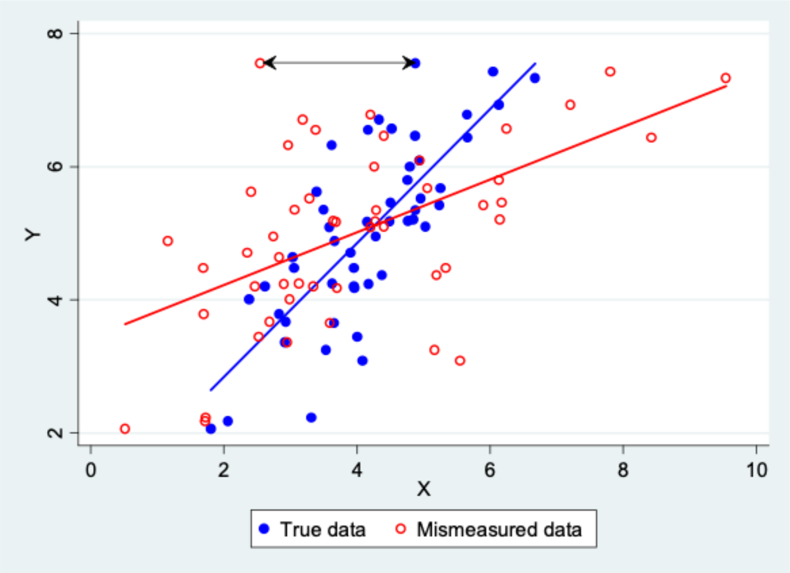
- Because

$$0 \leq \frac{\sigma^2_{X_{1i}}}{\sigma^2_{X_{1i}} + \sigma^2_w} \leq 1$$

- we have

- The classical measurement error β_1 is biased towards 0, which is also called **attenuation bias.**

Measurement error in X: classical measurement error



Solutions to errors-in-variables bias

- The best way to solve the errors-in-variables problem is to get an **accurate measure** of X.
  - Say nothing useful!
- **Instrumental Variables**
  - It relies on having another variable (the “instrumental” variable) that is correlated with the actual value X_i but is uncorrelated with the measurement error. We will discuss it later on.

Simultaneous Causality

Introduction

- So far we assumed that X affects Y, but what if Y also affects X simultaneously ?
  - thus we have $Y_i = \beta_0 + \beta_1 X_1 + u_i$
  - we also have $X_i = \gamma_0 + \gamma_1 Y_1 + v_i$
- Assume that $Cov(v_i, u_i) = 0$, then

- Simultaneous causality leads to biased & inconsistent OLS estimate.

$$Cov(X_i, u_i) = \frac{\gamma_1}{1 - \gamma_1 \beta_1} Var(u_i)$$

Simultaneous causality bias

- Substituting $Cov(X_i, u_i)$ in the formula for the $\hat{\beta}_1$

- OLS estimate is **inconsistent** if simultaneous causality bias exits.

Solutions to simultaneous causality bias

- The most effective solution is to employ **Instrumental Variables** or other experimental designs.
- **Simultaneous Equations Models** offer a classical alternative, though they are somewhat outdated in modern practice.

Functional form misspecification

Functional form misspecification

- Functional form misspecification also makes the OLS estimator biased and inconsistent.
- It can be seen as an special case of **OVB**,in which the omitted variables are the terms that reflect the missing nonlinear aspects of the regression function.
- It often can be detected by plotting the data and the estimated regression functions, and it can be corrected by using different functional forms.
- More general way is to use **semi-parametric** or **nonparametric** methods.
    - **Matching and Propensity Scores Matching**(we will cover it in the next lecture).

Missing Data and Sample Selection

Introduction

- **Missing data** is a common characteristic of economic data sets. It can threaten internal validity if it violates the assumption that our data is a **random sample from the population** of interest.
- In Stata and R, normally values are denoted as "." or "NA" to indicate missing data.
- Whether it poses a threat to internal validity depends on the **reason** *why the data is missing*.

Three types of missing data

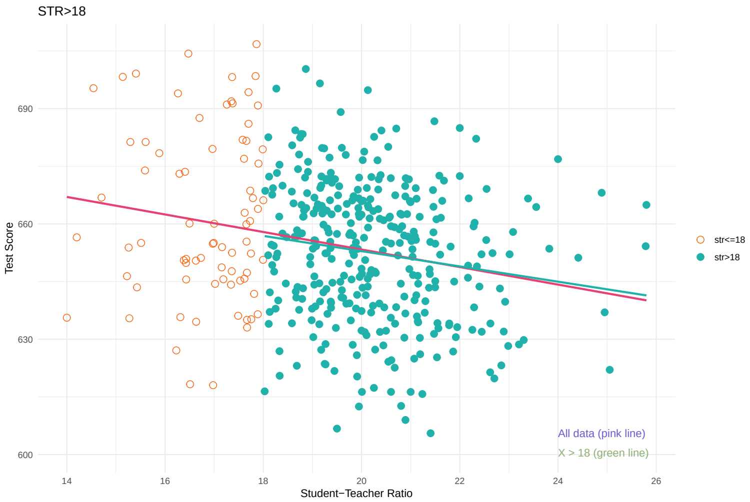
- We consider three types of missing data:
  1. **Missing completely at random**

2. **Missing based on X**: This shouldn't introduce significant bias into our analysis of the effect of X on Y, as long as **the number(or share) of missing data points is relatively small**.

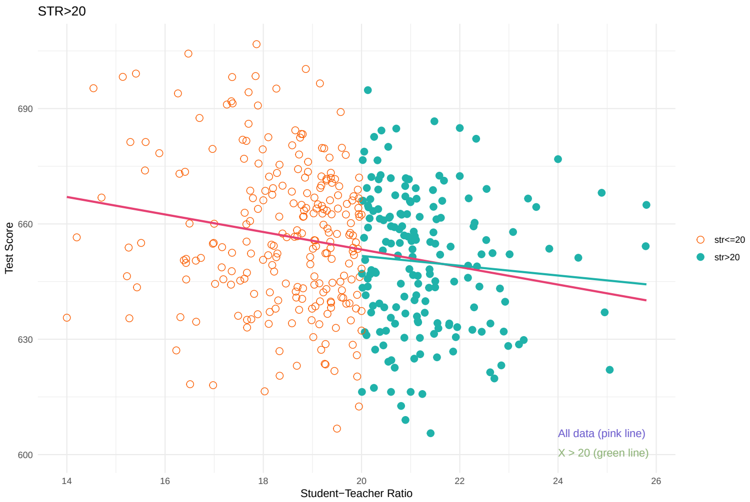
- And the conditional relationship between Y and X, thus the **causal effect** of X on Y, **remains unbiased** *within the observed data*.

Class Size and Test Score (STR>16)

Class Size and Test Score (STR>18)



Class Size and Test Score(STR>20)



Missing data based on Y

3. **Data is missing based on Y**: This is the most **problematic type of missing data**. It can introduce **significant bias** into our analysis of the effect of X on Y.

- Essentially, the key assumption for OLS regression is not hold any more in this case.
- Using OLS regression to analyze the effect of X on Y will introduce **bias**.

- Based on the missing data mechanism, we can classify three types of missing data in Y.

Missing data: Censored Data

(A) Data missing(only Y) because of a *selection process* that is related to the value of the dependent variable (Y), which is called **censored data**(删失数据）.

- An simple example: **the effect of education on income**.

- Two **improper** ways to deal with this problem:
  1. **Listwise deletion**: Drop the missing data points.
  2. **Imputation**: Use the top-coded income data (like 5000 RMB) to impute the missing data.
- Both methods will introduce bias into our analysis of the effect of education on income.

Missing data: Censored Data

- A special but useful case: **corner-solution models**.
  - The key feature of the behavior is that the decision can be divided into two parts:

- **Example**: Education on financial investment decision.
  - Many families does not have participation in financial investment. Then the investment data for these families is 0.
  - Other families have participation in financial investment. The investment data can be observed.

Missing data: Truncated Data

(B) Data are total missing(both X and Y) because of a *selection process* that is related to the value of the dependent variable (Y), which is called **truncated data**.

- **Example**: Innovation Investment on Totol Revenue of firms

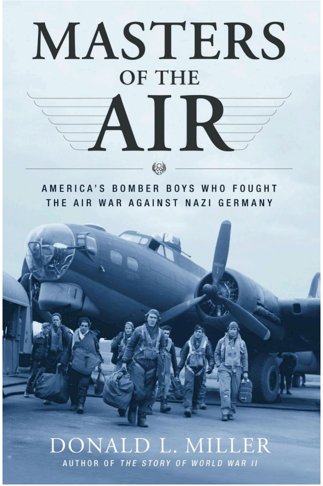
- Only use the firms in the sample to estimate the effect of innovation investment on total revenue will introduce **bias**.

Missing data: Sample Selection

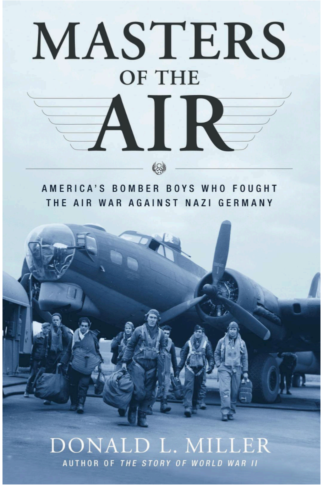
(C) Data are missing in Y because of a *selection process* that is related to another variable Z, which is called **sample selection data(样本选择数据)**.

- **Example**: Wage determination of married women(we will cover it in detail later on)
- NOTE: **sample selection or self-selection bias** v.s **selection bias.**

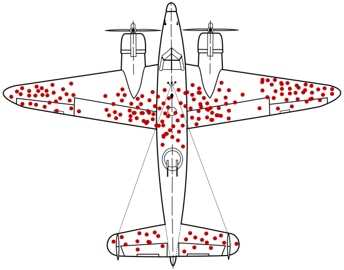
Survivorship Bias from WWII Aircrafts



Survivorship Bias from WWII Aircrafts



- The bullet holes of a bomber that, crucially, survived



- How to reinforce the armor to increase the survival of allied bombers?
- Which part of the bomber is more important?

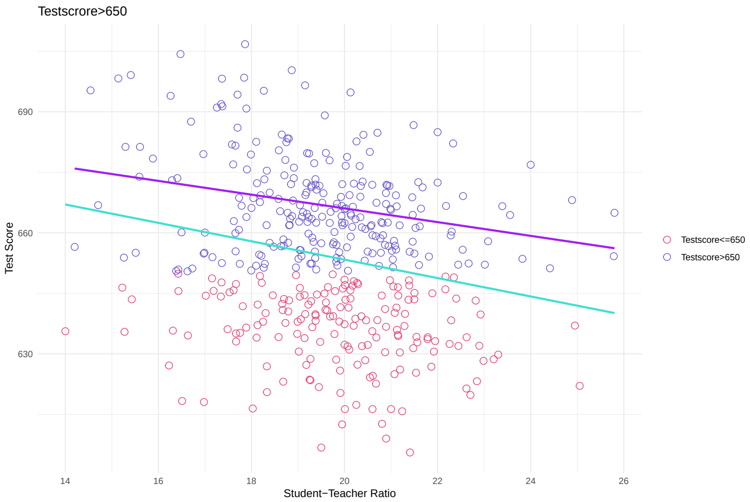
Missing data in Limited Dependent Variable Models

Type	Y is available	X is available	Selection Process
Censored	only $Y \in C^+$,other missing or zero	All samples in X are available.	Exogenous
Truncated	only $Y \in C^+$,other missing or zero	Only non-missing samples are available.	Exogenous
Sample Selection	only $Y \in C^+$,other missing or zero	All samples in X and Z are available.	Endogenous

Class Size and Test Score(Test Score>620)

Class Size and Test Score(Test Score>640)

Class Size and Test Score(Test Score>650)



Censored and Truncated Regression Models(I)

- Consider a **latent** variable regression model is

- Thus
  - Y_i^*: latent dependent variable
  - X_i': observed independent variables vector(only one variable we care most)
  - β: parameter vector(only one parameter we care most)
  - u_i: error term

Censored and Truncated Regression Models(II)

- The regression satisfies all the assumptions of the classical linear regression model.
- And we need additional assumptions to u_i:

$$u_i|X_i \sim N(0, \sigma^2)$$

  - where the expectation of u_i conditional on X_i is 0, because the regression model satisfies the 1st assumption of OLS regression, thus $E(u_i|X_i) = 0$.

Censored and Truncated Regression Models(III)

- **Censored Data**: The dependent variable is observed only if it exceeds(or less) a certain threshold.
  1. **Corner Solution Models**: Y is censored at 0.
  2. **Censored Regression Models**: Y is censored at a certain threshold.
- Suppose the latent variable is observed only if it exceeds a certain threshold 0, thus

The Expectation of Censored Data at Zero

- When the dependent variable is **censored at 0**, the expectation of Y_i^* is

- where Y_i is the **observed dependent variable**

Math Review: Truncated Density Function

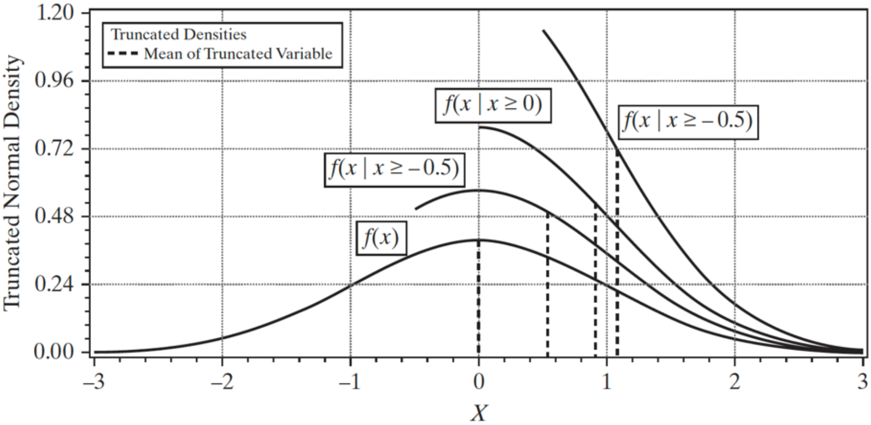
Truncated Density Function

If a continuous random variable X has p.d.f. $f(x)$ and c.d.f. $F(x)$ and a is a constant, then the conditional density function

$$f(x|x > a) = \begin{cases} \frac{f(x)}{1-F(a)} & if\ x > a \\ 0 & if\ x \leq a \end{cases}$$

- Please see the derivation in **Appendix**.

Math Review: Truncated Density Function



- It amounts merely to **scaling the density** so that it integrates to **one** over the range above c.

Standard Normal Truncated Density Function

- If X is distributed as standard normal, thus $X \sim N(0,1)$, then the p.d.f and c.d.f are as follow

- And c is a scalar, then we can get the *Truncated Density Function* of a R.V. distributed in **Standard Normal**

- The **Expectation** of in a *standard normal truncated p.d.f*

The Expectation of Censored Data at Zero

- Recall we just obtain the expectation of Y_i(observed dependent variable)
  - Where the $\lambda(\frac{-X_i'\beta}{\sigma})$ is the **Inverse Mills Ratio.**
- Then the **population regression function** or the CEF of Y on X is

The Bias of OLS Estimator in Censored Regression Model

- The **unbiased** regression equation should be

- If you use the original regression model, which is

$$Y_i = X_i'\tilde{\beta} + v_i$$

  - which means that you put $\sigma\lambda(\frac{-X_i'\beta}{\sigma})$ into the error term v_i, it makes the error term v_i correlated with the independent variable X_i,
  - thus the OLS estimator $\tilde{\beta}$ will suffer the **OVB** bias.
  - The bias can not be corrected by controlling more independent variable X_i.
- We could use **MLE** method to estimate the parameter vector β and σ in the unbiased regression model, which is the **Tobit model**.

The Tobit Model Regression in Graphs

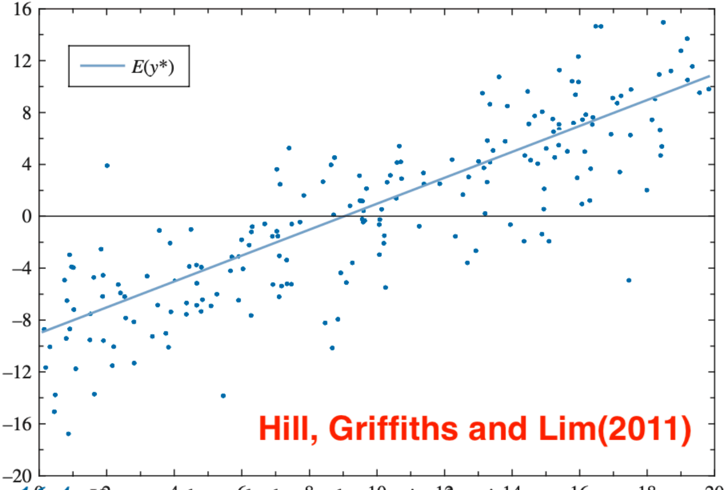


FIGURE 16.4 Uncensored sample data and regression function.

The Tobit Model Regression in Graphs

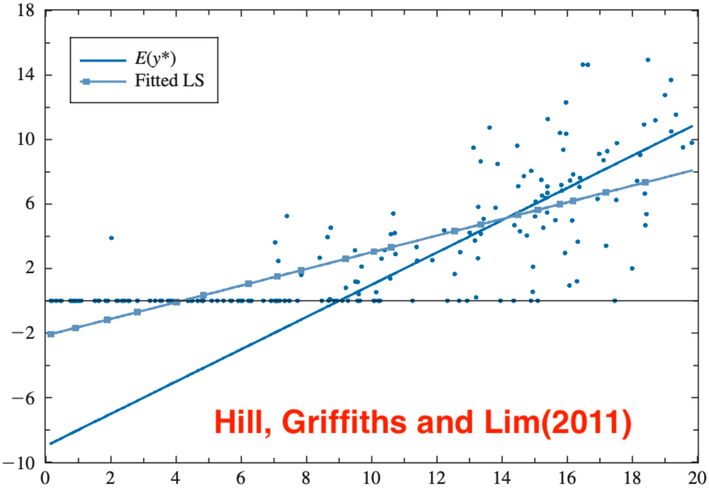


FIGURE 16.5 Censored sample data, and latent regression function and least squares fitted line.

The Tobit Model Regression in Graphs

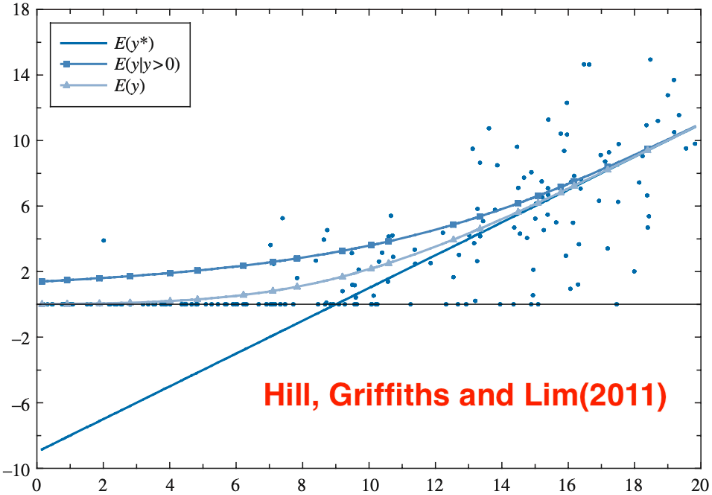


FIGURE 16.6 Censored sample data, and regression functions for observed and positive *y*-values.

Morz(1987):Labor Supply of Married Women

- Our regression equation is

$$Y_i = \beta_0 + \beta_1 X_i + Z_i' \beta_2 + u_i$$

- The dependent variable Y_i is the **hours worked per week**
- The treatment variables X_i are **education,**
- The control variables Z_i are **experience, age, and number of children under 6**.
- However, the working hours are not observed for those who do not work, thus we only have a **censored sample**.

Morz(1987):Labor Supply of Married Women

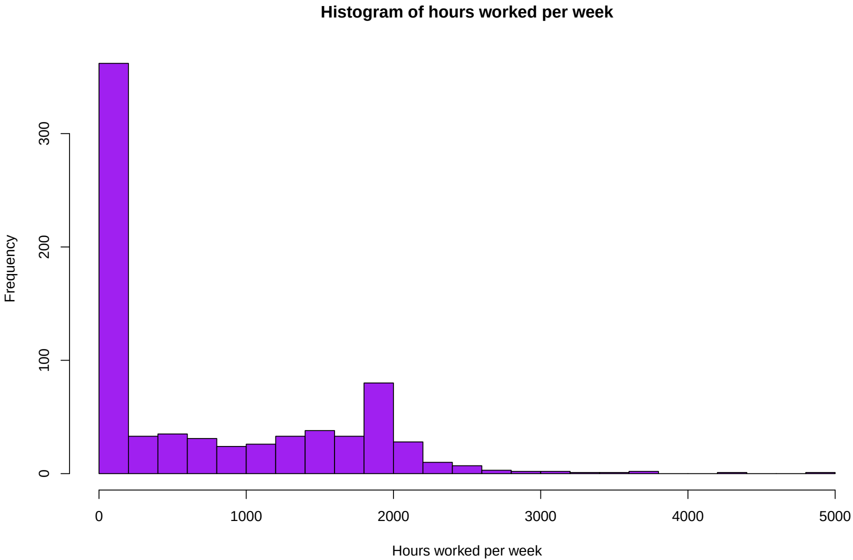


Table 4: Labor Supply of Married Women

	Dependent variable:		
	hours	hours2	hours
	OLS	*OLS*	*Tobit*
	(1)	(2)	(3)
educ	27.086^{**}	-16.462	73.291^{***}
	(12.240)	(15.581)	(20.475)
exper	48.040^{***}	33.936^{***}	80.535^{***}
	(3.642)	(5.009)	(6.288)
age	-31.308^{***}	-17.108^{***}	-60.768^{***}
	(3.961)	(5.458)	(6.888)
kidsl6	-447.855^{***}	-305.309^{***}	-918.918^{***}
	(58.413)	(96.449)	(111.661)
Constant	$1,335.306^{***}$	$1,829.746^{***}$	$1,349.876^{***}$
	(235.649)	(292.536)	(386.299)
Observations	753	428	753
Adjusted R^2	0.253	0.117	
Log Likelihood			-3,827.143
F Statistic	64.711^{***} (df = 4; 748)	15.123^{***} (df = 4; 423)	

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

hours are the observed hours worked per week for all observations, and hours2 is the observed hours worked per week only for those who work.

Tobit Model in R

- Following the general principle of the nonlinear models, the estimate coefficients are not meaningful.
- We need to use the **marginal effects** to interpret the results.

```
#>           Marg. Eff. Std. Error t value  Pr(>|t|)
#> educ        44.3724    12.3299  3.5988 0.0003408 ***
#> exper       48.7583     3.7685 12.9383 < 2.2e-16 ***
#> age        -36.7905     4.1097 -8.9522 < 2.2e-16 ***
#> kidsl6    -556.3382    66.4736 -8.3693 4.441e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Sample Selection Models

Example: Wage determination of married women

- A Classical Example: wage determination for Married Women

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

  - Y_i is logwage
  - X_i is schooling years
- The sample selection problem arises in that the sample consists only of women who chose to work.
  - If the selection to work is random,then OK.
  - But in reality, married women choose to work probably they are *smarter,more ambitious* and more risk-preferent which normally can not observed or measured in the data(Z).

Heckman Sample Selection Model(I)

- A **two-equation behavioral model**

1. *selection equation*

where Z_i is a latent variable which indicates **the propensity of working** for a married woman

- the error term e_i satisfies

$$E[e_i|W_i] = 0$$

- Then Z_i is a dummy variable to represent whether a woman to work or not actually,thus

Heckman Sample Selection Model(II)

2. *outcome equation*

- **where the outcome(Y_i^*) can be observed only when Z_i=1 or $Z_i^* > 0$**

- **The error term u_i satisfies $E[u_i|X_i] = 0$**

Heckman Sample Selection Model(III)

- The conditional expectation of wages on X_i is

$$E[Y_i^*|X_i] = X_i'\beta$$

- The conditional expectation of wages on X_i is only for women who work($Z^* > 0$)

Heckman Sample Selection Model(IV)

- If u_i and e_i is independent, then $E[u_i|e_i > -W_i'\gamma] = 0$, then

$$E[Y_i^*|X_i, Z_i^* > 0] = E[Y_i^*|X_i] = X_i'\beta$$

  - It means using sample-selected data does not make the estimation of β biased.
- But in reality, unobservables in the two equations, thus u_i and e_i, are likely to be **correlated**
  - eg. innate ability,ambitions,
- Instead assume that u_i and e_i are **jointly normal distributed**, which can be standardized easily, thus

- where we let $\sigma_e^2 = 1$, and ρ is the correlation coefficient between u_i and e_i

Math Review: Two Normal Distributed R.V.s

Two Normal Distributed R.V.s

For any two normal variables (n_0, n_1) with zero mean, we can write $n_1 = \alpha_0 n_0 + \eta$, where $\eta \sim N(0, \sigma_\eta)$ and $E(\eta | n_0) = 0$.Then we have

Math Review: Two Normal Distributed R.V.s

Two Normal Distributed R.V.s

For any two normal variables (n_0, n_1) with zero mean, we can write $n_1 = \alpha_0 n_0 + \eta$, where $\eta \sim N(0, \sigma_\eta)$ and $E(\eta | n_0) = 0$.Then we have

$$\alpha_0 = \frac{Cov(n_0, n_1)}{Var(n_0)}$$

Math Review: Two Normal Distributed R.V.s

Two Normal Distributed R.V.s

For any two normal variables (n_0, n_1) with zero mean, we can write $n_1 = \alpha_0 n_0 + \eta$, where $\eta \sim N(0, \sigma_\eta)$ and $E(\eta | n_0) = 0$.Then we have

$$\alpha_0 = \frac{Cov(n_0, n_1)}{Var(n_0)}$$

or

$$E(n_1 \mid n_0) = \frac{Cov(n_0, n_1)}{Var(n_0)} n_0$$

Math Review: Two Normal Distributed R.V.s

Two Normal Distributed R.V.s

For any two normal variables (n_0, n_1) with zero mean, we can write $n_1 = \alpha_0 n_0 + \eta$, where $\eta \sim N(0, \sigma_\eta)$ and $E(\eta | n_0) = 0$.Then we have

$$\alpha_0 = \frac{Cov(n_0, n_1)}{Var(n_0)}$$

or

$$E(n_1 \mid n_0) = \frac{Cov(n_0, n_1)}{Var(n_0)} n_0$$

Then

$$n_1 = E(n_1 \mid n_0) + \eta = \frac{Cov(n_0, n_1)}{Var(n_0)} n_0 + \eta$$

Heckman Sample Selection Model(V)

- For two normal variables u_i and e_i with zero means, we have

- Then

where $\eta \sim N(0, \sigma_\eta)$ and $E(\eta | e_i) = 0$

Heckman Sample Selection Model(VI)

- Then the conditional expectation of u_i

Heckman Sample Selection Model(VII)

- Then the conditional expectation of u_i

Heckman Sample Selection Model(VIII)

- Then the conditional expectation of wages on X_i is only for women who work($Z^* > 0$)

- Turning it into a regression form

$$Y_i = X_i'\beta + \sigma_\lambda \lambda(W_i'\gamma) + u_i$$

- Recall our original wage determination equation

$$Y_i = X_i'\tilde{\beta} + v_i$$

- Likewise, the error term v_i is correlated with the independent variable X_i, thus the OLS estimator $\tilde{\beta}$ will suffer the **OVB** bias.

Heckman Sample Selection Model(IX)

- It means that if we could include $\lambda(W_i'\gamma)$ as **an additional regressor** into the outcome equation, thus we run

obtaining the **unbiased** and **consistent** estimate β using a self-selected sample.

- The coefficient before $\lambda(\cdot)$ can be testing significance to indicate whether the term should be included in the regression, in other words, *whether the selection should be corrected*.

Heckit Model Estimation: a two-step method

1. Estimate selection equation using **all observations**,thus

- obtain estimates of parameters $\hat{\gamma}$
- computer the **Inverse Mills Ratio(IMR)** $\frac{\phi(W_i'\hat{\gamma})}{\Phi(W_i'\hat{\gamma})} = \hat{\lambda}(W_i'\hat{\gamma})$

2. Estimate the outcome equation using **only the selected observations**.

- **Note**: standard error is not right, have to be adjusted because we use $\hat{\lambda}(W_i'\hat{\gamma})$ instead of $\lambda(W_i'\gamma)$ in the estimation.

An Example: Wage Equation for Married Women

TABLE 17.7 Wage Offer Equation for Married Women

Dependent Variable: log(*wage*)		
Independent Variables	OLS	Heckit
educ	.108 (.014)	.109 (.016)
exper	.042 (.012)	.044 (.016)
$exper^2$	−.00081 (.00039)	−.00086 (.00044)
constant	−.522 (.199)	−.578 (.307)
$\hat{\lambda}$	—	.032 (.134)
Sample size *R*-squared	428 .157	428 .157

Wrap Up

- Missing data is a common problem in practice for empirical researchers.
- Missing data can be caused by many reasons, such as **non-response**, **attrition**, **non-sampling error**, etc.
- Missing data can be **missing completely at random (MCAR)**, **missing at random (MAR)**, or **missing not at random (MNAR)**.
  - Normally, missing values in X may not be a serious problem.
  - Missing values in Y are problematic.
- Limited Dependent Variable Models are using to deal with missing data in Y.
  - Tobit model
  - Heckman Sample Selection Model

Sources of Inconsistency of OLS Standard Errors

Introduction

- A different threat to internal validity.Even if the OLS estimator is **consistent** and the sample is large, **inconsistent standard errors** will let you make a **bad judgment** about the effect of the interest in the population.
- There are two main reasons for inconsistent standard errors:

1. Heteroskedasticity: The solution to this problem is to use **heteroskedasticity-robust standard errors** and to construct **F-statistics** using a heteroskedasticity-robust variance estimator.

Sources of Inconsistency of OLS Standard Errors

2. **Correlation of the error term across observations**:
    - This will not happen if the data are obtained by sampling at random from the population.(**i.i.d**)
    - Sometimes, however, sampling is only **partially random**.
        - When the data are repeated observations on the same entity over time.(**time series**)
        - Another situation in which the error term can be correlated across observations is when sampling is based on a geographical or other group unit.(**cluster**)

- Both situation means that the assumptions

- the second key assumption in OLS is partially violated.
- In this case, the OLS estimator is still **unbiased** and**consistent**, but the standard errors are **inconsistent**.

Clustering Standard Error: A Simple Example

- Suppose we still focus on the topic of class size and student performance, but now the data are collecting on **students** rather than school district.
- Our regression model is

- $TestScore_{ig}$ is the *dependent variable* for student i in class g, with G groups.
- $ClassSize_g$ the *independent variable*(or treatment variable), **varies only at the group level**(class).
- Intuitively,the test score of students in the same class(g) tend to be correlated.

Clustering Standard Error(I)

- Recall the variance of the OLS estimator:

Clustering Standard Error(II)

- When the sample is clustered, which means that the observations are only randomly sampled across clusters, g and G is the number of clusters.
- Then the numerator of the variance of the OLS estimator is:

Clustering Standard Error(III)

- **Substituting** $Cov(u_{ig}, u_{kg}) = \begin{cases} \sigma_u^2 & \text{if } i = k \\ \rho\sigma_u^2 & \text{if } i \neq k \end{cases}$**:**

- **This final expression shows how intraclass correlation** ρ **inflates the variance through the additional cross-product terms.**

Clustering Standard Error(IV)

- **Stata: use option** `vce(cluster clustvar)`**. Where** `clustvar` **is a variable that identifies the groups in which on observables are allowed to correlate.**
- **R: the** `vcovHC()` **function from** `plm` **package**

Magnitude of β_1

Introduction

- The criteria for determining the magnitude of β_1 are as follows:
  - **large enough** to make sense.
  - **Question**: *How large is considered large enough?*
- The magnitude of β_1 is not only determined by the actual relationship between X and Y, but also by the units in which X and Y are measured.
- Recall the class size and student performance example, the coefficient β_1 is -2.38, which means that if class size increases by 1, then student performance decreases by 2.38 points.
  - Whether the -2.38 is large or small depends on the scale of the variables and distribution of the data.
- Normally, we compare the magnitude of β_1 to the **mean value of** Y or the **standard deviation of** Y.

Standardized Variables

- Assume Xs and Y are all continuous variables, then we run a multiple regression model

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_k X_{ik} + \hat{u}_i$$

- Because $\Sigma \hat{u}_i = 0$ and $\overline{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}_1 + \cdots + \hat{\beta}_k \bar{X}_k$,then

$$Y_i - \bar{Y} = \hat{\beta}_1 (X_{i1} - \bar{X}_1) + \hat{\beta}_2 (X_{i2} - \bar{X}_2) + \cdots + \hat{\beta}_k (X_{ik} - \bar{X}_k) + \hat{u}_i$$

- Then, we obtain following expressions

$$\begin{aligned} \frac{Y_i - \bar{Y}}{\sigma_y} = & \hat{\beta}_1 \frac{\sigma_{x1}}{\sigma_y} \frac{(X_{i1} - \bar{X}_1)}{\sigma_{x1}} + \hat{\beta}_2 \frac{\sigma_{x1}}{\sigma_y} \frac{(X_{i2} - \bar{X}_2)}{\sigma_{x2}} + \cdots + \\ & \hat{\beta}_k \frac{\sigma_{x1}}{\sigma_y} \frac{(X_{i2} - \bar{X}_k)}{\sigma_{x_k}} + \frac{\hat{u}_i}{\sigma_y} \end{aligned}$$

Standardized Variables

- Then we have a standardized regression model

$$Z_y = \hat{\phi}_1 Z_1 + \hat{\phi}_2 Z_2 + \cdots + \hat{\phi}_k Z_k + v_i$$

  where Z_y denotes the **Z**-score of Y, Z_1 denotes the **Z-score** of X_1,and so on.

- The estimate coefficients

$$\hat{\phi}_j = (\hat{\sigma}_j / \hat{\sigma}_y)\, \hat{\beta}_j \text{ for } j = 1, \ldots, k$$

- $\hat{\phi}_j$ are traditionally called **standardized coefficients** or **beta coefficients** , which can be explained as if X_j increases by **1 standard deviation**, then Y changes by ϕ **standard deviations**.

Standardized Only One X

- Consider a linear regression model as usual

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + u_i$$

- Or

$$Y_i = \beta_0 + \beta_1 X_{i1} + C\Gamma' + u_i$$

- Where $\Gamma = (\beta_2, ..., \beta_k), C_i = (X_{2i}, ..., X_{ki})$
- If we only standardize X_1 and leave other variables as they are, then the standardized version of X_1 is defined as

$$Z_1 = \frac{X_1 - \bar{X}_1}{\sigma_{x_1}}$$

- Then we have the standardized regression model

Standardized Only One X

- Substitute Z_1 back into the original regression equation in place of X_1, we have

$$\begin{aligned} Y_i =& \beta_0 + \beta_1 \left(\frac{X_1 - \bar{X}_1}{\sigma_{x_1}} \right) + C\Gamma' + u_i \\ =& \beta_0 - \frac{\bar{X}_1}{\sigma_{x_1}} + \beta_1 \frac{X_1}{\sigma_{x_1}} + C\Gamma' + u_i \end{aligned}$$

- Then we have the marginal effect of X_1 on Y as

$$\begin{aligned} \frac{\partial Y}{\partial X_1} &= \beta_1 \frac{1}{\sigma_{x_1}} \\ \Rightarrow \beta_1 &= \frac{\partial Y}{\frac{\partial X_1}{\sigma_{x_1}}} \end{aligned}$$

- The estimate coefficients $\hat{\beta}_1$ is can be interpreted as follows:
  - if X_1 increases by **1 standard deviation**, then Y changes by β_1 units.

Standardized Only Y

- If we only standardize Y and leave other variables as they are, then the standardized version of Y is defined as

$$Z_Y = \frac{Y - \bar{Y}}{\sigma_Y}$$

- Then the regression model becomes

$$\begin{aligned} Z_Y &= \beta_0 + \beta_1 X_1 + C\Gamma' + u_i \\ \Rightarrow \frac{Y - \bar{Y}}{\sigma_Y} &= \beta_0 + \beta_1 X_1 + C\Gamma' + u_i \\ \Rightarrow \frac{Y}{\sigma_Y} &= \beta_0 + \frac{\bar{Y}}{\sigma_Y} + \beta_1 X_1 + C\Gamma' + u_i \end{aligned}$$

Standardized Only Y

- Then we have the marginal effect of X_1 on Y as

$$\frac{\partial Y}{\partial X_1} = \beta_1 \sigma_y$$
$$\Rightarrow \beta_1 = \frac{\frac{\partial Y}{\sigma_y}}{\partial X_1}$$

- The estimate coefficients $\hat{\beta}_1$ is can be interpreted as follows:
  - if X_1 increases by **1 unit**, then Y changes by β_1 *standard deviation*.

Wrap Up

- There are five primary threats to the internal validity of a multiple regression study:
  1. Omitted variables
  2. Functional form misspecification
  3. Errors in variables (measurement error in the regressors)
  4. Missing Data and Sample selection
  5. Simultaneous causality
- Besides, the data structure may violate the 2th OLS regression assumption, thus random sampling.
  1. Times series
  2. Cluster data
  3. Spatial data
- Last but not least, the **magnitude of** β_1 matters.

Wrap Up

- Each of these, if present, results in failure of the first least squares assumption,which in turn means that the OLS estimator is biased and inconsistent.
- Incorrect calculation of the standard errors also poses a threat to internal validity.
- Applying this list of threats to a multiple regression study provides a systematic way to assess the internal validity of that study.

External validity

Definition

- Suppose we estimate a regression model that is internally valid.
- Can the statistical inferences be generalized from the population and setting studied to other populations and settings?

Threats to external validity

1. Differences in populations
    - The population from which the sample is drawn might differ from the population of interest
    - For example, if you estimate the returns to education for *men*, these results might not be informative if you want to know the returns to education for *women*.
2. Differences in settings
    - The setting studied might differ from the setting of interest due to differences in laws, institutional environment and physical environment.
    - For example, the estimated returns to education using data from the U.S might not be informative for China.
    - Because the educational system is different and different institutions of the labor market.

Application to the case of class size and test score

- This analysis was based on test results for California school districts.
- Suppose for the moment that these results are internally valid. To what other populations and settings of interest could this finding be generalized?
  - generalize to colleges: it is implausible
  - generalize to other U.S. elementary school districts: it is plausible

Wrap up

- It is not easy to make your studies valid internally.
- Even harder when you consider generalize your findings.
- Then common way to generalize the findings actually is to repeat to make the studies internal valid.
- Then we make a generalizing conclusions based on a bunch of internal valid studies.

Example: Test Scores and Class Size

External Validity

- Whether the California analysis can be generalized—that is, whether it is externally valid—depends on the population and setting to which the generalization is made.
- we consider whether the results can be generalized to other elementary public school districts in the United States.
  - more specifically, 220 public school districts in *Massachusetts* in 1998.
  - if we find similar results in the California and Massachusetts, it would be evidence of external validity of the findings in California.
  - Conversely, finding different results in the two states would raise questions about the internal or external validity of at least one of the studies.

Comparison of the California and Massachusetts data.

TABLE 9.1 Summary Statistics for California and Massachusetts Test Score Data Sets

	California		Massachusetts	
	Average	Standard Deviation	Average	Standard Deviation
Test scores	654.1	19.1	709.8	15.1
Student–teacher ratio	19.6	1.9	17.3	2.3
% English learners	15.8%	18.3%	1.1%	2.9%
% Receiving lunch subsidy	44.7%	27.1%	15.3%	15.1%
Average district income ($)	$15,317	$7226	$18,747	$5808
Number of observations	420		220	
Year	1999		1998	

Regressor	(1)	(2)	(3)	(4)	(5)	(6)
Student–teacher ratio (*STR*)	−1.72** (0.50)	−0.69* (0.27)	−0.64* (0.27)	12.4 (14.0)	−1.02** (0.37)	−0.67* (0.27)
STR^2				−0.680 (0.737)		
STR^3				0.011 (0.013)		
% English learners		−0.411 (0.306)	−0.437 (0.303)	−0.434 (0.300)		
% English learners > median? (Binary, *HiEL*)					−12.6 (9.8)	
$HiEL \times STR$					0.80 (0.56)	
% Eligible for free lunch		−0.521** (0.077)	−0.582** (0.097)	−0.587** (0.104)	−0.709** (0.091)	−0.653** (0.72)
District income (logarithm)		16.53** (3.15)				
District income			−3.07 (2.35)	−3.38 (2.49)	−3.87* (2.49)	−3.22 (2.31)
District income^2			0.164 (0.085)	0.174 (0.089)	0.184* (0.090)	0.165 (0.085)
District income^3			−0.0022* (0.0010)	−0.0023* (0.0010)	−0.0023* (0.0010)	−0.0022* (0.0010)
Intercept	739.6** (8.6)	682.4** (11.5)	744.0** (21.3)	665.5** (81.3)	759.9** (23.2)	747.4** (20.3)

Test scores and class size in MA

F-Statistics and p-Values Testing Exclusion of Groups of Variables

	(1)	(2)	(3)	(4)	(5)	(6)
All STR variables and interactions $= 0$				2.86 (0.038)	4.01 (0.020)	
$STR^2, STR^3 = 0$				0.45 (0.641)		
$Income^2, Income^3$			7.74 (< 0.001)	7.75 (< 0.001)	5.85 (0.003)	6.55 (0.002)
$HiEL, HiEL \times STR$					1.58 (0.208)	
SER	14.64	8.69	8.61	8.63	8.62	8.64
$\overline{R}^2$	0.063	0.670	0.676	0.675	0.675	0.674

These regressions were estimated using the data on Massachusetts elementary school districts described in Appendix 9.1. Standard errors are given in parentheses under the coefficients, and p-values are given in parentheses under the F-statistics. Individual coefficients are statistically significant at the *5% level or **1% level.

Test scores and class size in MA

TABLE 9.3 Student–Teacher Ratios and Test Scores: Comparing the Estimates from California and Massachusetts

	OLS Estimate $\hat{\beta}_{STR}$	Standard Deviation of Test Scores Across Districts	Estimated Effect of Two Fewer Students per Teacher, In Units of: Points on the Test	Standard Deviations
California				
Linear: Table 9.3(2)	−0.73 (0.26)	19.1	1.46 (0.52)	0.076 (0.027)
Cubic: Table 9.3(7) *Reduce STR from 20 to 18*	—	19.1	2.93 (0.70)	0.153 (0.037)
Cubic: Table 9.3(7) *Reduce STR from 22 to 20*	—	19.1	1.90 (0.69)	0.099 (0.036)
Massachusetts				
Linear: Table 9.2(3)	−0.64 (0.27)	15.1	1.28 (0.54)	0.085 (0.036)

Standard errors are given in parentheses.

Internal Validity

- The similarity of the results for California and Massachusetts does not ensure their internal validity.
- **Omitted variables**: teacher quality or a low student-teacher ratio might have families that are more committed to enhancing their children's learning at home or migrating to a better district.
- **Functional form**: Although further functional form analysis could be carried out, this suggests that the main findings of these studies are unlikely to be sensitive to using different nonlinear regression specifications.
- **Errors in variables**: The average student-teacher ratio in the district is a broad and potentially inaccurate measure of class size.
  - Because students' mobility, the STR might not accurately represent the actual class sizes, which in turn could lead to the estimated class size effect being biased toward zero.

Internal Validity

- **Selection**: data cover all the public elementary school districts in the state that satisfy minimum size restrictions, so there is no reason to believe that sample selection is a problem here.
- **Simultaneous causality**: it would arise if the performance on tests affected the student–teacher ratio.
- **Heteroskedasticity** and **correlation of the error term** across observations.
    - It does not threaten internal validity.
    - Correlation of the error term across observations, however, could threaten the consistency of the standard errors because the assumption of simple random sampling is violated.

Appendix

Math Review: Truncated Density Function

Truncated Density Function

The proof follows from the definition of a conditional probability is

$$Pr(A|B) = \frac{Pr(AB)}{Pr(B)}$$

then,

$$F(x|X > c) =$$

Math Review: Truncated Density Function

Truncated Density Function

The proof follows from the definition of a conditional probability is

$$Pr(A|B) = \frac{Pr(AB)}{Pr(B)}$$

then,

$$F(x|X > c) = \frac{Pr(X < x, X > c)}{Pr(X > c)}$$

Math Review: Truncated Density Function

Truncated Density Function

The proof follows from the definition of a conditional probability is

$$Pr(A|B) = \frac{Pr(AB)}{Pr(B)}$$

then,

$$F(x|X > c) = \frac{Pr(X < x, X > c)}{Pr(X > c)} = \frac{Pr(c < X < x)}{1 - F(c)}$$

Math Review: Truncated Density Function

Truncated Density Function

The proof follows from the definition of a conditional probability is

$$Pr(A|B) = \frac{Pr(AB)}{Pr(B)}$$

then,

$$F(x|X > c) = \frac{Pr(X < x, X > c)}{Pr(X > c)} = \frac{Pr(c < X < x)}{1 - F(c)}$$

$$= \frac{F(x) - F(c)}{1 - F(c)}$$

then,

$$f(x|x > c) = \frac{d}{dx} F(x|X > c) = \frac{\frac{d}{dx}[F(x)] - 0}{1 - F(c)} = \frac{f(x)}{1 - F(c)}$$

The Expectation in a Standard Normal Truncated

Proof

$$E(x|x > c) =$$

The Expectation in a Standard Normal Truncated

Proof

$$E(x|x>c) = \int_{c}^{+\infty} x f(x|x>c) dx =$$

The Expectation in a Standard Normal Truncated

Proof

$$E(x|x>c)=\int_{c}^{+\infty}xf(x|x>c)dx=\int_{c}^{+\infty}x\frac{\phi\left(x\right)}{1-\Phi\left(c\right)}dx$$

The Expectation in a Standard Normal Truncated

Proof

$$E(x|x>c)=\int_c^{+\infty} x f(x|x>c)dx=\int_c^{+\infty} x\frac{\phi(x)}{1-\Phi(c)}dx$$

$$=\frac{1}{1-\Phi(c)}\int_c^{+\infty} x\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx$$

The Expectation in a Standard Normal Truncated

Proof

$$E(x|x>c)=\int_c^{+\infty} xf(x|x>c)dx=\int_c^{+\infty} x\frac{\phi(x)}{1-\Phi(c)}dx$$

$$=\frac{1}{1-\Phi(c)}\int_c^{+\infty} x\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx$$

$$=\frac{1}{1-\Phi(c)}\int_c^{+\infty} \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}d(\frac{x^2}{2})$$

The Expectation in a Standard Normal Truncated

Proof

$$
\begin{aligned}
E(x|x>c) &= \int_c^{+\infty} x f(x|x>c)dx = \int_c^{+\infty} x \frac{\phi(x)}{1-\Phi(c)} dx \\
&= \frac{1}{1-\Phi(c)} \int_c^{+\infty} x \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \\
&= \frac{1}{1-\Phi(c)} \int_c^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} d(\frac{x^2}{2}) \\
&= \frac{1}{1-\Phi(c)} \int_{\frac{c^2}{2}}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-t} d(t)
\end{aligned}
$$

The Expectation in a Standard Normal Truncated

Proof

$$E(x|x>c) = \int_c^{+\infty} xf(x|x>c)dx = \int_c^{+\infty} x\frac{\phi(x)}{1-\Phi(c)}dx$$

$$= \frac{1}{1-\Phi(c)}\int_c^{+\infty} x\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx$$

$$= \frac{1}{1-\Phi(c)}\int_c^{+\infty} \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}d(\frac{x^2}{2})$$

$$= \frac{1}{1-\Phi(c)}\int_{\frac{c^2}{2}}^{+\infty} \frac{1}{\sqrt{2\pi}}e^{-t}d(t)$$

$$= \frac{1}{1-\Phi(c)} \times \frac{1}{\sqrt{2\pi}} - e^{-t} \Big|_{\frac{c^2}{2}}^{+\infty}$$

The Expectation in a Standard Normal Truncated

Proof

$$E(x|x>c) = \int_c^{+\infty} xf(x|x>c)dx = \int_c^{+\infty} x\frac{\phi(x)}{1-\Phi(c)}dx$$

$$= \frac{1}{1-\Phi(c)}\int_c^{+\infty} x\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx$$

$$= \frac{1}{1-\Phi(c)}\int_c^{+\infty} \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}d(\frac{x^2}{2})$$

$$= \frac{1}{1-\Phi(c)}\int_{\frac{c^2}{2}}^{+\infty} \frac{1}{\sqrt{2\pi}}e^{-t}d(t)$$

$$= \frac{1}{1-\Phi(c)}\times\frac{1}{\sqrt{2\pi}} - e^{-t}\ |_{\frac{c^2}{2}}^{+\infty}$$

$$= \frac{1}{1-\Phi(c)}\times\frac{1}{\sqrt{2\pi}}e^{-\frac{c^2}{2}} = \frac{\phi(c)}{1-\Phi(c)}$$