

Week 3: Instrumental Variables and LATE

The Oregon Health Insurance Experiment

[Class slides](#) · [Download slides as PDF](#) · [After-class exercise](#) · [Download notes as PDF](#) · [Download Quarto source](#)

Prepare discussion items individually. The discussant list is provided separately on Moodle.

Learning Objectives

By the end of this topic, you should be able to:

- Explain why omitted variables can bias OLS and how a valid instrument supplies variation that identifies a causal effect.
- Derive the Wald and exactly identified IV estimators using the analogy principle and moment conditions, and explain their connection to two-stage least squares.
- State the assumptions for LATE and explain the steps showing why the IV estimate identifies the effect for compliers.
- Distinguish the effects of winning the Oregon lottery and receiving Medicaid, and interpret what the estimates imply for policy.

Before class

Read these notes and Finkelstein et al. (2012), “**The Oregon Health Insurance Experiment: Evidence from the First Year.**” Focus on the lottery and data in Sections II–III, the empirical framework in Section IV, and **Tables II, III, IV, V, VIII, and IX.** [Published paper](#) · [Paper from Amy Finkelstein’s website](#).

Textbook references

- **The Effect:** [Chapter 19, Instrumental Variables](#), particularly the explanation of the IV comparison and local average treatment effects.
- **Causal Inference: The Remix:** [Chapter 7, Instrumental Variables](#), particularly the Wald estimator, two-stage least squares, and LATE.
- **Mastering ’Metrics:** Chapter 3, *Instrumental Variables*. [Companion resources](#) · [Mastering Econometrics videos at Marginal Revolution University](#).
- **Mostly Harmless Econometrics — optional advanced reading:** Chapter 4, especially Sections 4.1 and 4.4 on IV and heterogeneous effects. [Book contents](#).

Use these books for complementary explanations. **Additional reading:** Angrist and Krueger (2001), “[Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments](#)”, *Journal of Economic Perspectives* 15(4): 69–85.

1. From selection on observables to instrumental variables

Last week, we constructed comparisons between treated and untreated individuals with similar observed characteristics. The identifying assumption was that those characteristics account for selection into treatment that is related to potential outcomes.

Suppose important determinants remain unobserved. For example, health needs can influence both whether someone obtains health insurance and how much health care they use. Comparing insured and uninsured people may therefore combine the effect of insurance with differences that would have existed without insurance.

An **instrument**, Z_i , supplies a source of variation in treatment whose relationship with the outcome can be given a causal interpretation. In Oregon, a lottery offered some people the opportunity to apply for Medicaid. Winning increased the probability of obtaining insurance. It did not guarantee that a person enrolled.

We distinguish three variables:

Symbol	Meaning in the Oregon application
Z_i	Household selected by the lottery
D_i	Individual obtains Medicaid coverage
Y_i	An outcome, such as having an outpatient visit

The effect of the **offer** and the effect of **receiving insurance** answer different policy questions. IV connects the two.

2. The Wald estimator

Recall: omitted variables bias

The true population relationship is

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 A_i + \nu_i,$$

where Y_i is earnings, D_i is **years of schooling**, and A_i is ability. The coefficient β_1 is the causal effect of an additional year of schooling. We assume the remaining disturbance ν_i is uncorrelated with schooling and ability.

Recall the population regression of ability on schooling:

$$A_i = \gamma_0 + \gamma_1 D_i + v_i.$$

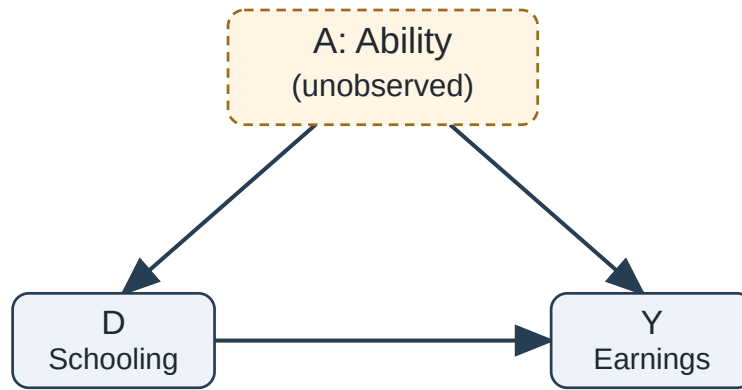
Substituting this relationship into the earnings equation gives

$$Y_i = (\beta_0 + \beta_2 \gamma_0) + (\beta_1 + \beta_2 \gamma_1) D_i + \beta_2 v_i + \nu_i.$$

Thus the population slope of the short regression, which omits ability, is

$$\beta_1^S = \beta_1 + \beta_2 \gamma_1.$$

If ability raises earnings and is positively related to schooling, the short regression overstates the causal effect of schooling.



Unobserved ability affects both schooling and earnings.

The backdoor path $D \leftarrow A \rightarrow Y$ connects schooling and earnings through ability. The short regression combines the causal effect along $D \rightarrow Y$ with this additional association.

Bias and unbiasedness of OLS

The estimated slope changes from sample to sample. Its **sampling distribution** describes the values we would obtain by repeatedly drawing samples of the same size from the same population. The **bias** of the estimator, relative to the causal effect β_1 , is

$$\text{Bias}(\hat{\beta}_1^S) = E[\hat{\beta}_1^S] - \beta_1.$$

Here the expectation is over repeated samples. An estimator is **unbiased** if its mean across those samples equals the parameter it estimates. Individual estimates can still be above or below that parameter.

To derive the expectation of the OLS slope, imagine repeatedly observing outcomes for the same schooling values. Write the collection of schooling values as $\mathbf{D} = (D_1, \dots, D_n)$, and suppose there is variation in schooling. Starting with $Y_i = \beta_0 + \beta_1 D_i + u_i$, the sample OLS formula gives

$$\begin{aligned} \hat{\beta}_1^S &= \frac{\sum_i (D_i - \bar{D}) Y_i}{\sum_i (D_i - \bar{D})^2} \\ &= \beta_1 + \frac{\sum_i (D_i - \bar{D}) u_i}{\sum_i (D_i - \bar{D})^2}. \end{aligned}$$

The intercept drops out because $\sum_i (D_i - \bar{D}) = 0$. Conditional on $\mathbf{D}$, both the denominator and each $D_i - \bar{D}$ are fixed, so we can move them outside the expectation:

$$E[\hat{\beta}_1^S | \mathbf{D}] = \beta_1 + \frac{\sum_i (D_i - \bar{D}) E[u_i | \mathbf{D}]}{\sum_i (D_i - \bar{D})^2}.$$

Under the **zero conditional mean assumption**, $E[u_i | \mathbf{D}] = 0$, the second term is zero. The estimator is therefore unbiased conditional on the schooling values. Averaging over possible sets of schooling values also gives $E[\hat{\beta}_1^S] = \beta_1$, provided the expectation exists. Conditioning allows us to do the calculation even when schooling is random in the

population.

When ability is omitted, $u_i = \beta_2 A_i + \nu_i$. If ability is related to schooling, the zero conditional mean assumption generally fails. The fraction in the conditional-expectation expression shows how the resulting conditional bias arises. The earlier expression $\beta_1^S = \beta_1 + \beta_2 \gamma_1$ describes the population short-regression slope; here we have examined the expectation of its sample estimator relative to the causal effect.

From the conditional expectation to the covariance formula

Substitute $u_i = \beta_2 A_i + \nu_i$ into the conditional expectation above and assume $E[\nu_i | \mathbf{D}] = 0$:

$$E[\hat{\beta}_1^S | \mathbf{D}] = \beta_1 + \beta_2 \frac{\sum_i (D_i - \bar{D}) E[A_i | \mathbf{D}]}{\sum_i (D_i - \bar{D})^2}.$$

The fraction is the conditional expectation of the sample slope from regressing ability on schooling, $E[\hat{\gamma}_1 | \mathbf{D}]$. This follows by applying the same OLS formula to that auxiliary regression:

$$\hat{\gamma}_1 = \frac{\sum_i (D_i - \bar{D}) A_i}{\sum_i (D_i - \bar{D})^2}.$$

To obtain the familiar population covariance expression for the finite-sample bias, now assume the **conditional mean of ability is linear in schooling**:

$$E[A_i | \mathbf{D}] = \gamma_0 + \gamma_1 D_i.$$

Equivalently, the residual in $A_i = \gamma_0 + \gamma_1 D_i + v_i$ has conditional mean zero given the schooling values. This is an additional assumption beyond merely writing the population regression.

The numerator simplifies to

$$\begin{aligned} \sum_i (D_i - \bar{D}) E[A_i | \mathbf{D}] &= \sum_i (D_i - \bar{D}) (\gamma_0 + \gamma_1 D_i) \\ &= \gamma_1 \sum_i (D_i - \bar{D})^2. \end{aligned}$$

The intercept term vanishes because the deviations from the sample mean sum to zero. The remaining sum cancels the denominator, giving

$$E[\hat{\beta}_1^S | \mathbf{D}] = \beta_1 + \beta_2 \gamma_1.$$

The population slope in the ability-on-schooling regression is $\gamma_1 = \text{Cov}(A_i, D_i) / \text{Var}(D_i)$. Therefore,

$$E[\hat{\beta}_1^S | \mathbf{D}] = \beta_1 + \beta_2 \frac{\text{Cov}(A_i, D_i)}{\text{Var}(D_i)}.$$

Under these assumptions, the conditional bias relative to the causal effect is $\beta_2 \gamma_1$. Averaging over schooling samples yields the same unconditional bias, provided the expectation exists. If ability raises earnings and is positively related to schooling, this bias is positive.

What can we do if ability is unobserved?

A binary instrument

Write the earnings equation as $Y_i = \beta_0 + \beta_1 D_i + u_i$, with $u_i = \beta_2 A_i + \nu_i$.

Suppose we have a binary instrument, Z_i , that has **two properties at the same time**:

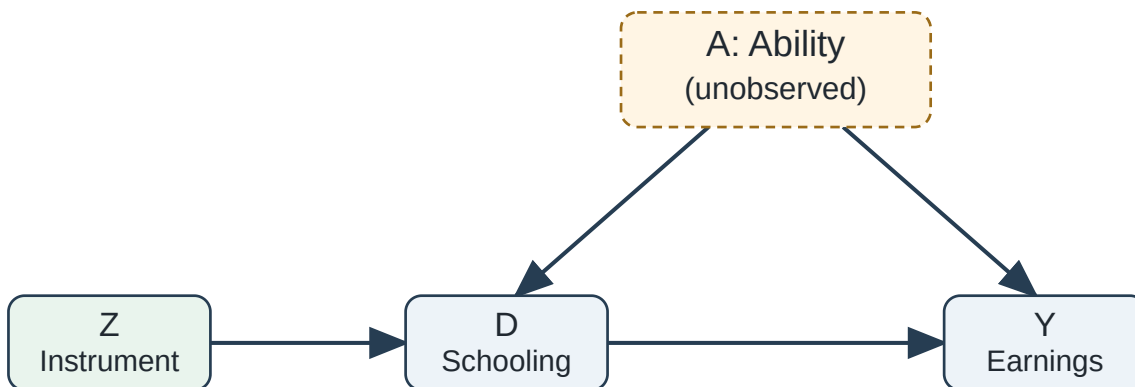
$$E[u_i | Z_i = 1] = E[u_i | Z_i = 0] = 0,$$

$$E[D_i | Z_i = 1] \neq E[D_i | Z_i = 0].$$

The instrument shifts schooling and affects earnings through schooling. We call the first relationship the **orthogonality assumption** and the second relationship the **relevance assumption**.

Finding a variable with either property is easy. The challenge is finding one that satisfies **both at the same time**: it predicts years of schooling while remaining uncorrelated with the unobserved determinants of earnings. For example, an independently generated random number can satisfy orthogonality, but will not predict schooling in the population. Ability predicts schooling, but is part of the unobserved determinants of earnings.

The **exclusion restriction** requires the instrument to affect earnings through schooling. Together with independence from the other determinants of earnings, it supports the orthogonality condition above, where we have normalised the mean of u_i to zero.



The instrument shifts schooling independently of unobserved ability and affects earnings through schooling.

The arrow $Z \rightarrow D$ represents relevance. Exclusion requires the effect of Z on earnings to operate through schooling, and independence requires Z to be independent of ability and the other unobserved determinants of earnings. Together, exclusion and independence support the orthogonality assumption.

Taking conditional expectations:

$$\begin{aligned} E[Y_i | Z_i = 1] &= \beta_0 + \beta_1 E[D_i | Z_i = 1], \\ E[Y_i | Z_i = 0] &= \beta_0 + \beta_1 E[D_i | Z_i = 0]. \end{aligned}$$

Subtract the second equation from the first:

$$E[Y_i | Z_i = 1] - E[Y_i | Z_i = 0] = \beta_1 [E[D_i | Z_i = 1] - E[D_i | Z_i = 0]].$$

Solving for the causal effect gives

$$\beta_1 = \frac{E[Y_i | Z_i = 1] - E[Y_i | Z_i = 0]}{E[D_i | Z_i = 1] - E[D_i | Z_i = 0]}.$$

Using the **analogy principle** to replace the population conditional means with the sample conditional means gives the **Wald estimator**:

$$\hat{\beta}_{1,\text{Wald}} = \frac{\bar{Y}_{Z=1} - \bar{Y}_{Z=0}}{\bar{D}_{Z=1} - \bar{D}_{Z=0}}.$$

The instrument-induced difference in outcomes equals the causal effect multiplied by the instrument-induced difference in treatment. Dividing by the treatment difference recovers the effect per unit of treatment.

An illustrative schooling and earnings model

For this illustration, the simulated relationships are

$$D_i = 11 + 5Z_i + 1.8A_i + v_i, \quad Y_i = 100 + 5D_i + 20A_i + \nu_i,$$

where Z_i is Bernoulli with probability one-half, and A_i , v_i , and ν_i are mutually independent, mean-zero normal variables, independent of Z_i . Their standard deviations are 1, 0.65, and 10 respectively. The figure uses 400 observations. These are simulated data used to illustrate the geometry.

The true effect of an additional year of schooling is 5. Independence of the variables used to generate schooling gives

$$\text{Cov}(A_i, D_i) = 1.8 \text{Var}(A_i) = 1.8,$$

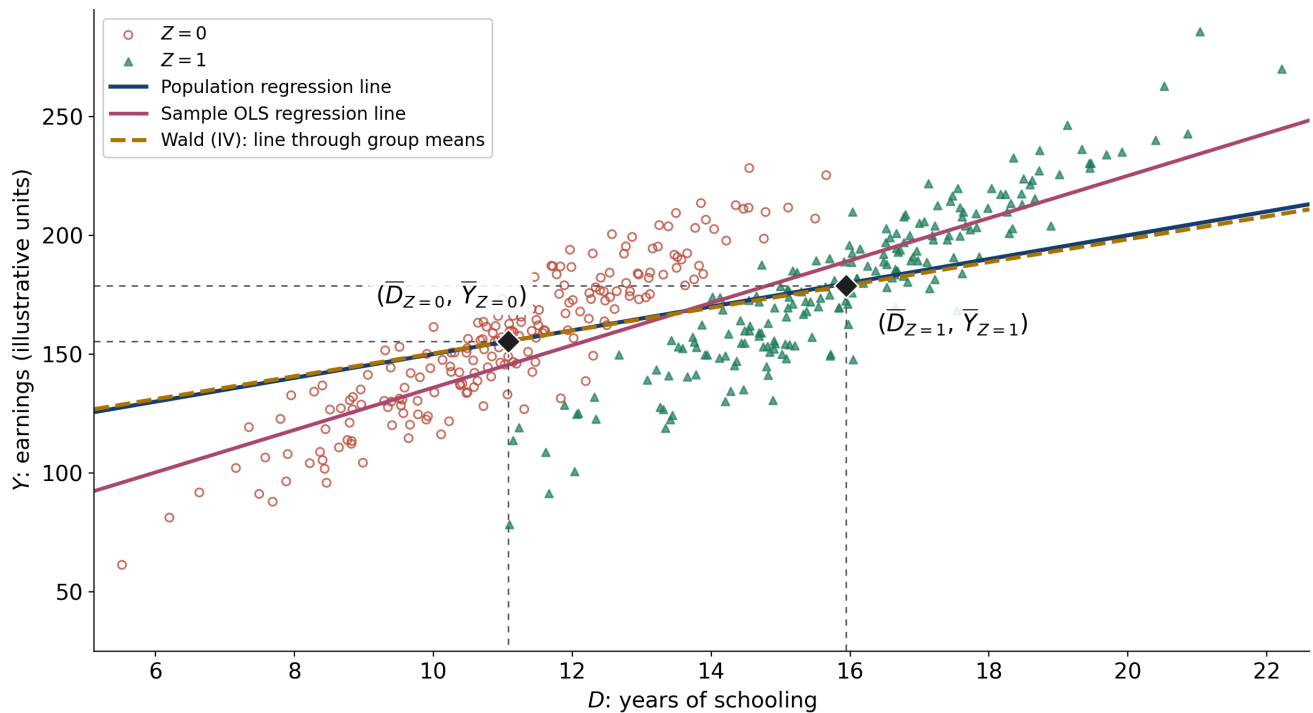
$$\text{Var}(D_i) = 25 \text{Var}(Z_i) + 1.8^2 \text{Var}(A_i) + \text{Var}(v_i) = 6.25 + 3.24 + 0.4225 = 9.9125.$$

Thus the population slope of earnings on schooling when ability is omitted is

$$\beta_1^S = 5 + 20 \frac{1.8}{9.9125} \approx 8.63.$$

The population omitted-variable bias is approximately 3.63. The sample OLS slope will vary around the population regression target; the covariance calculation here identifies that target without asserting exact finite-sample unbiasedness for this simulation.

The graphical argument



Simulated illustration following the Wald figure in the lecture slides. Open circles and triangles distinguish the two instrument groups; diamonds mark their sample means.

The **population regression line** has the true causal slope, $\beta_1 = 5$. Here ability raises both years of schooling and earnings, so the **sample OLS line** is too steep. The **Wald (IV) line** passes through the two sample mean points, $(\bar{D}_{Z=0}, \bar{Y}_{Z=0})$ and $(\bar{D}_{Z=1}, \bar{Y}_{Z=1})$. Its slope is their vertical separation divided by their horizontal separation. The instrument shifts treatment independently of ability, allowing this comparison to recover the causal slope as the sample grows. In a finite sample, the IV and population lines can differ.

Reproducing the figure in Python

First generate 400 observations from the relationships above. Setting a random seed gives us the same data each time we run the code. Ability enters both equations, while the instrument is generated independently of ability and the disturbances.

```
import numpy as np
import pandas as pd
import statsmodels.formula.api as smf
import matplotlib.pyplot as plt

rng = np.random.default_rng(20260923)
n = 400

Z = rng.binomial(n=1, p=0.5, size=n)
A = rng.normal(loc=0, scale=1, size=n)
nu = rng.normal(loc=0, scale=10, size=n)
v = rng.normal(loc=0, scale=0.65, size=n)

D = 11 + 5 * Z + 1.8 * A + v
Y = 100 + 5 * D + 20 * A + nu

simulated = pd.DataFrame({"Y": Y, "D": D, "Z": Z, "A": A})
simulated.head()
```

	Y	D	Z	A
0	232.716597	18.348908	1	1.210493
1	152.226520	14.902881	1	-0.606686
2	199.842613	17.421453	1	0.238943
3	175.340173	11.811175	0	0.565863
4	126.716057	8.833038	0	-0.612828

Next estimate OLS and calculate the Wald estimator from the two instrument-group means. The intercept of the IV line is chosen so that it passes through the mean point for $Z = 0$; the Wald slope ensures it also passes through the mean point for $Z = 1$.

```
ols = smf.ols("Y ~ D", data=simulated).fit(cov_type="HC1")
means = simulated.groupby("Z")[["D", "Y"]].mean()

wald_slope = (
    (means.loc[1, "Y"] - means.loc[0, "Y"])
    / (means.loc[1, "D"] - means.loc[0, "D"])
)
wald_intercept = means.loc[0, "Y"] - wald_slope * means.loc[0, "D"]

pd.DataFrame({
    "Line": ["Population", "Sample OLS", "Wald (IV)"],
    "Slope": [5, ols.params["D"], wald_slope]
}).round(3)
```

	Line	Slope
0	Population	5.000
1	Sample OLS	8.917
2	Wald (IV)	4.807

Finally, plot the observations, the three lines, and the group means. The dashed horizontal and vertical guides locate each mean point on the axes.

```

plt.rcParams.update({"font.size": 14, "font.family": "DejaVu Sans"})
fig, ax = plt.subplots(figsize=(12, 6.7))

group0 = simulated.loc[simulated["Z"] == 0]
group1 = simulated.loc[simulated["Z"] == 1]
ax.scatter(group0["D"], group0["Y"], s=27, marker="o",
           facecolors="none", edgecolors="#c04b38", alpha=0.8,
           linewidths=1.1, label="$Z=0$")
ax.scatter(group1["D"], group1["Y"], s=25, marker="^",
           color="#20815d", alpha=0.7, label="$Z=1$")

# Evaluate each line over the range of observed treatment values.
d_grid = np.linspace(simulated["D"].min() - 0.4,
                     simulated["D"].max() + 0.4, 250)
ax.plot(d_grid, 100 + 5 * d_grid, color="#163e70", linewidth=2.6,
        label="Population regression line")
ax.plot(d_grid, ols.params["Intercept"] + ols.params["D"] * d_grid,
        color="#ab4a71", linewidth=2.4, label="Sample OLS regression line")
ax.plot(d_grid, wald_intercept + wald_slope * d_grid,
        color="#ac7300", linewidth=2.6, linestyle="--",
        label="Wald (IV): line through group means")

ax.set_xlim(d_grid[0], d_grid[-1])
ax.set_ylim(25, 295)
for z in [0, 1]:
    mean_d = means.loc[z, "D"]
    mean_y = means.loc[z, "Y"]
    ax.plot([d_grid[0], mean_d], [mean_y, mean_y, 25],
           color="#555555", linestyle=(0, (4, 4)), linewidth=1)
    ax.scatter(mean_d, mean_y, s=120, marker="D", color="#222222",
              edgecolors="white", zorder=8)
    label = (r"$\overline{D}_{Z=" + str(z)
            + r"}$, \, \overline{Y}_{Z=" + str(z) + r"}$")
    offset = (-85, 24) if z == 0 else (20, -30)
    ax.annotate(label, (mean_d, mean_y), xytext=offset,
               textcoords="offset points", fontsize=15, zorder=9,
               bbox={"facecolor": "white", "edgecolor": "none", "alpha": 0.9})

ax.set_xlabel("$D$: years of schooling")
ax.set_ylabel("$Y$: earnings (illustrative units)")
ax.spines[["top", "right"]].set_visible(False)
ax.legend(loc="upper left", fontsize=11.5, framealpha=0.95)
fig.tight_layout()
plt.close(fig)

```

Figure 1

3. From Wald to instrumental variables

First stage and reduced form

The **first stage** regresses the endogenous explanatory variable on **all the exogenous variables**. The **reduced form** regresses the outcome on **all the exogenous variables**. Both include the excluded instrument(s), any included exogenous covariates, and an intercept.

The terminology comes from **simultaneous-equations models**. We solve the system to express each endogenous variable as a function of all the exogenous variables and a disturbance. In IV applications, “the reduced form” usually refers to the equation for the outcome Y_i . The first-stage equation for D_i is also a reduced-form equation.

In our simplest example, substitute the first-stage relationship $D_i = \pi_0 + \pi_1 Z_i + v_i$ into the earnings equation $Y_i = \beta_0 + \beta_1 D_i + u_i$:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 (\pi_0 + \pi_1 Z_i + v_i) + u_i \\ &= (\beta_0 + \beta_1 \pi_0) + \beta_1 \pi_1 Z_i + (u_i + \beta_1 v_i) \\ &= \delta_0 + \delta_1 Z_i + e_i. \end{aligned}$$

Thus $\delta_0 = \beta_0 + \beta_1 \pi_0$, $\delta_1 = \beta_1 \pi_1$, and $e_i = u_i + \beta_1 v_i$. The reduced-form coefficient on the instrument is the causal effect multiplied by the first-stage coefficient. Relevance permits us to divide by π_1 , recovering $\beta_1 = \delta_1 / \pi_1$. When the model includes exogenous covariates, those covariates enter both the first stage and the reduced form as well.

Write the population regressions as

$$\begin{aligned} D_i &= \pi_0 + \pi_1 Z_i + v_i && \text{(first stage),} \\ Y_i &= \delta_0 + \delta_1 Z_i + e_i && \text{(reduced form).} \end{aligned}$$

With one endogenous explanatory variable and one excluded instrument, the model is **exactly identified**. The IV estimate equals the reduced-form coefficient on that instrument divided by its first-stage coefficient, using the same sample, weights, and included exogenous covariates. This single-coefficient ratio applies to the one-instrument, one-endogenous-variable case.

With a binary instrument, each slope is the difference between the corresponding group means. Therefore Wald is the estimated reduced-form slope divided by the estimated first-stage slope: $\hat{\beta}_{1,\text{Wald}} = \hat{\delta}_1 / \hat{\pi}_1$.

For $P(Z_i = 1) = p$,

$$\text{Cov}(Y_i, Z_i) = [E[Y_i | Z_i = 1] - E[Y_i | Z_i = 0]] p(1 - p),$$

$$\text{Cov}(D_i, Z_i) = [E[D_i | Z_i = 1] - E[D_i | Z_i = 0]] p(1 - p).$$

The common factor cancels in their ratio:

$$\beta_1 = \frac{\text{Cov}(Y_i, Z_i)}{\text{Cov}(D_i, Z_i)} = \frac{\text{Cov}(Y_i, Z_i) / \text{Var}(Z_i)}{\text{Cov}(D_i, Z_i) / \text{Var}(Z_i)}.$$

This expresses IV as the reduced-form coefficient divided by the first-stage coefficient. The covariance ratio also applies to a continuous instrument when $\text{Cov}(Z_i, u_i) = 0$ and $\text{Cov}(Z_i, D_i) \neq 0$. Replacing these population covariances with sample covariances gives the IV estimator.

IV as a method-of-moments estimator

The Wald estimator used the analogy principle: replace population conditional means with sample conditional means. We can construct IV in the same way from a population moment condition. This gives an estimator for either a binary or a continuous instrument.

Orthogonality concerns the covariance between the instrument and the **outcome disturbance**:

$$\begin{aligned}
0 &= \text{Cov}(Z_i, u_i) \\
&= E[(Z_i - E[Z_i])(u_i - E[u_i])] \\
&= E[(Z_i - E[Z_i])u_i].
\end{aligned}$$

The “useful result” lets us drop the mean from the second term because $E[Z_i - E[Z_i]] = 0$.

Now substitute $u_i = Y_i - \beta_0 - \beta_1 D_i$:

$$E[(Z_i - E[Z_i])(Y_i - \beta_0 - \beta_1 D_i)] = 0.$$

To see the cancellation explicitly, the term removed is

$$E[(Z_i - E[Z_i])E[u_i]] = E[u_i]E[Z_i - E[Z_i]] = 0.$$

Thus this simplification holds even without imposing $E[u_i] = 0$. It is the familiar useful result: centring one variable is sufficient when calculating a covariance. We retain u_i for the outcome disturbance and v_i for the first-stage residual.

Apply the **analogy principle**, replacing the population mean of Z_i with $\bar{Z}$ and the expectation with a sample average. Choose the estimated coefficients to satisfy

$$\frac{1}{n} \sum_i (Z_i - \bar{Z}) (Y_i - \hat{\beta}_0 - \hat{\beta}_1 D_i) = 0.$$

Multiply by n and expand:

$$\sum_i (Z_i - \bar{Z}) Y_i - \hat{\beta}_0 \sum_i (Z_i - \bar{Z}) - \hat{\beta}_1 \sum_i (Z_i - \bar{Z}) D_i = 0.$$

Because the centred instrument sums to zero, the intercept term vanishes. Rearranging gives

$$\hat{\beta}_{1,IV} = \frac{\sum_i (Z_i - \bar{Z}) Y_i}{\sum_i (Z_i - \bar{Z}) D_i}.$$

We can subtract $\bar{Y}$ in the numerator and $\bar{D}$ in the denominator without changing either sum, again because $\sum_i (Z_i - \bar{Z}) = 0$. Thus

$$\hat{\beta}_{1,IV} = \frac{\sum_i (Z_i - \bar{Z}) (Y_i - \bar{Y})}{\sum_i (Z_i - \bar{Z}) (D_i - \bar{D})} = \frac{\widehat{\text{Cov}}(Z_i, Y_i)}{\widehat{\text{Cov}}(Z_i, D_i)}.$$

The common sample-covariance divisor cancels. This expression requires a nonzero sample covariance between the instrument and schooling; the identifying relevance condition requires the corresponding population covariance to be nonzero.

To estimate the intercept as well, use the population condition $E[u_i] = 0$. Its sample counterpart is

$$\frac{1}{n} \sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_{1,IV} D_i) = 0,$$

which gives $\hat{\beta}_{0,IV} = \bar{Y} - \hat{\beta}_{1,IV} \bar{D}$.

This is the **method of moments**: choose the coefficients so that the sample counterparts of the population moment conditions hold. For a binary instrument, the covariance ratio simplifies to the Wald ratio of differences in group means. Two-stage least squares provides another way to compute the same estimate.

Using the instrument to predict treatment

Our goal is to isolate variation in treatment that is uncorrelated with the unobserved determinants of the outcome. The population first-stage regression gives the predicted component

$$D_i^* = \pi_0 + \pi_1 Z_i, \quad \pi_1 = \frac{\text{Cov}(D_i, Z_i)}{\text{Var}(Z_i)}.$$

Since this component is a linear function of the instrument,

$$\text{Cov}(D_i^*, u_i) = \pi_1 \text{Cov}(Z_i, u_i) = 0.$$

The first-stage residual is uncorrelated with Z_i and hence with D_i^* . Writing $D_i = D_i^* + v_i$ in the outcome equation gives

$$Y_i = \beta_0 + \beta_1 D_i^* + (\beta_1 v_i + u_i).$$

Thus D_i^* is uncorrelated with the disturbance in this equation. In a sample, we estimate this predicted component with $\widehat{D}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$. This leads directly to two-stage least squares.

Dividing schooling into exogenous and endogenous components

The fitted first-stage regression gives the exact sample decomposition

$$D_i = \widehat{D}_i + \hat{v}_i, \quad \widehat{D}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i, \quad \hat{v}_i = D_i - \widehat{D}_i.$$

The instrument divides variation in schooling into two components:

- The **exogenous part**, $\widehat{D}_i$, is the schooling predicted by the instrument. With a valid instrument, this estimates the population component D_i^* that is uncorrelated with the unobserved determinants of earnings.
- The **endogenous part**, $\hat{v}_i$, is the first-stage residual. It contains the remaining variation in schooling, including variation associated with unobserved ability. It can also contain variation unrelated to the earnings disturbance; the label identifies where the endogenous variation remains.

The population decomposition makes the identifying argument explicit:

$$D_i = D_i^* + v_i,$$

$$\text{Cov}(D_i, u_i) = \text{Cov}(D_i^*, u_i) + \text{Cov}(v_i, u_i) = \text{Cov}(v_i, u_i).$$

The first covariance on the right is zero under instrument validity. Thus the relationship between schooling and the outcome disturbance is carried by the residual component. IV uses the instrument-predicted component to estimate the causal effect of schooling.

The OLS first-stage residual is orthogonal to the fitted values in the sample. Exogeneity with respect to the outcome disturbance rests on the instrument assumptions; sampling variation can still produce a sample correlation between fitted schooling and that disturbance.

4. Two-stage least squares

The same estimate can be obtained in two stages:

1. Estimate the first-stage regression of D_i on an intercept and Z_i . Obtain $\widehat{D}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$.
2. Regress Y_i on an intercept and $\widehat{D}_i$.

The estimated second-stage regression is

$$Y_i = \hat{\beta}_{0,2SLS} + \hat{\beta}_{1,2SLS} \widehat{D}_i + \hat{r}_i.$$

Here $\hat{r}_i$ is the residual from regressing Y_i on an intercept and the predicted values $\widehat{D}_i$. The coefficient on $\widehat{D}_i$ is the 2SLS estimate of the causal slope under the IV assumptions.

With a binary instrument, the first stage assigns each observation its instrument group's mean treatment rate. The second stage uses the difference between those predicted treatment rates to explain the difference in outcomes.

To see the equivalence algebraically, the slope in the second regression is

$$\frac{\widehat{\text{Cov}}(\widehat{D}_i, Y_i)}{\widehat{\text{Var}}(\widehat{D}_i)} = \frac{\hat{\pi}_1 \widehat{\text{Cov}}(Z_i, Y_i)}{\hat{\pi}_1^2 \widehat{\text{Var}}(Z_i)} = \frac{\widehat{\text{Cov}}(Z_i, Y_i)}{\widehat{\text{Cov}}(Z_i, D_i)}.$$

Two-stage least squares (2SLS) generalises this procedure to multiple instruments and additional covariates.

Including covariates

Let X_{ik} denote covariate k for person i , with K exogenous covariates. The outcome equation and first stage are

$$Y_i = \beta_0 + \beta_1 D_i + \sum_{k=1}^K \beta_{k+1} X_{ik} + u_i,$$

$$D_i = \pi_0 + \pi_1 Z_i + \sum_{k=1}^K \pi_{k+1} X_{ik} + v_i.$$

The fitted first stage gives

$$\widehat{D}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i + \sum_{k=1}^K \hat{\pi}_{k+1} X_{ik}.$$

The estimated **second-stage regression** is

$$Y_i = \hat{\beta}_{0,2SLS} + \hat{\beta}_{1,2SLS} \widehat{D}_i + \sum_{k=1}^K \hat{\beta}_{k+1,2SLS} X_{ik} + \hat{r}_i.$$

The outcome is regressed on predicted treatment and the same exogenous controls. For IV standard errors, the outcome-equation residual uses observed treatment:

$$\hat{u}_i = Y_i - \hat{\beta}_{0,2SLS} - \hat{\beta}_{1,2SLS}D_i - \sum_{k=1}^K \hat{\beta}_{k+1,2SLS}X_{ik}.$$

An IV routine uses $\hat{u}_i$ and the appropriate IV variance calculation.

Include the same covariates in **both stages**. With one excluded instrument, the 2SLS coefficient remains the reduced-form coefficient divided by the first-stage coefficient when both regressions use the same observations, weights, and covariates.

The reduced form is $Y_i = \delta_0 + \delta_1 Z_i + \sum_{k=1}^K \delta_{k+1} X_{ik} + e_i$.

Python estimation and standard errors

Use an IV estimation routine to obtain the coefficient and its standard error. The formula in `linearmodels` places the endogenous variable and its excluded instrument inside square brackets:

```
import linearmodels.iv as iv

model = iv.IV2SLS.from_formula(
    "outcome ~ 1 + [insured ~ selected]", data=df
)
result = model.fit(cov_type="robust")
print(result.summary)
```

Here `1` requests an intercept, and `[insured ~ selected]` tells Python to instrument insurance with lottery selection. `cov_type="robust"` requests heteroskedasticity-consistent standard errors. Install the package in your Python environment with `python -m pip install linearmodels` if needed. [Package documentation](#).

Running the two regressions manually reproduces the coefficient. For inference, the IV routine uses the outcome-equation residual $Y_i - \hat{\beta}_0 - \hat{\beta}_1 D_i$ and the appropriate IV variance calculation. The second-stage OLS printout instead uses residuals based on $\hat{D}_i$ and gives the wrong standard errors for IV.

For binary outcomes this is a **linear probability model**. A coefficient of 0.20 is a 20-percentage-point effect. Disturbance variance can differ across observations, which motivates heteroskedasticity-consistent inference. In Oregon we additionally allow outcomes within a household to be correlated, using **household-clustered standard errors**.

Finite samples and consistency

Unbiasedness concerns the mean of an estimator across repeated samples of a given size: $E[\hat{\beta}_1] = \beta_1$.

Consistency concerns what happens as the sample size grows: the probability that the estimate is close to the true parameter approaches one.

For one instrument, the IV estimator is

$$\hat{\beta}_{1,IV} = \beta_1 + \frac{n^{-1} \sum_i (Z_i - \bar{Z}) u_i}{n^{-1} \sum_i (Z_i - \bar{Z}) D_i}.$$

With independent random sampling and finite second moments, the numerator converges to $\text{Cov}(Z_i, u_i) = 0$, and the denominator to $\text{Cov}(Z_i, D_i) \neq 0$. Hence

$$\hat{\beta}_{1,IV} \xrightarrow{p} \beta_1.$$

IV can be consistent without being unbiased in finite samples. We return to its finite-sample behaviour when studying weak instruments.

The notation $\xrightarrow{p}$ means **convergence in probability**. For any fixed positive tolerance, the probability that the estimate differs from the true parameter by more than that tolerance tends to zero as n grows. This is an asymptotic statement.

The OLS unbiasedness argument held the observed schooling values fixed and took expectations of the outcome disturbances. For IV, the denominator contains a sample relationship between the instrument and endogenous schooling. The expectation of that ratio generally does not equal the ratio of expectations. The probability-limit argument instead uses convergence of the two sample moments and a nonzero limiting denominator. In some IV models finite-sample moments do not exist, so consistency is particularly useful as a separate property.

5. Heterogeneous effects: whose effect does IV estimate?

Return to potential outcomes. Y_i^1 is the outcome with treatment and Y_i^0 the outcome without it. Now the effect $Y_i^1 - Y_i^0$ may vary across people.

We also need potential **treatment statuses**:

- D_i^1 : whether person i would receive treatment if $Z_i = 1$.
- D_i^0 : whether person i would receive treatment if $Z_i = 0$.

The superscript on Y denotes the **treatment state**; the superscript on D denotes the **instrument state**. Observed treatment is

$$D_i = Z_i D_i^1 + (1 - Z_i) D_i^0.$$

Four possible responses to the instrument

Type	D_i^0	D_i^1	Response to an offer
Never-taker	0	0	Receives no treatment under either assignment
Complier	0	1	Receives treatment when offered
Always-taker	1	1	Receives treatment under either assignment
Defier	1	0	Receives treatment only without the offer

These types describe behaviour relative to a **particular instrument**. We observe one treatment status for each person, so we generally cannot identify each person's type. For example, a treated lottery winner could be a complier or an always-taker.

The identifying assumptions

1. Independence. The instrument is independent of potential outcomes and potential treatment statuses:

$$Z_i \perp (Y_i^0, Y_i^1, D_i^0, D_i^1).$$

This condition may hold conditional on observed covariates. In Oregon, the relevant lottery comparison conditions on the number of household members on the list. The derivation below first considers an unconditionally random instrument; we return to conditional comparisons afterwards.

2. Exclusion. The instrument affects the outcome through treatment. To state this precisely, let $Y_i(d, z)$ denote the outcome under treatment d and instrument value z . Exclusion requires

$$Y_i(d, 1) = Y_i(d, 0) = Y_i^d.$$

We can then write the observed outcome as

$$Y_i = Y_i^0 + (Y_i^1 - Y_i^0) D_i.$$

3. Monotonicity. Orient the instrument so it encourages treatment. Assume

$$D_i^1 \geq D_i^0 \quad \text{for every } i.$$

The instrument may increase treatment or leave it unchanged. This rules out defiers.

4. Relevance. The instrument changes the probability of treatment:

$$E[D_i | Z_i = 1] - E[D_i | Z_i = 0] > 0.$$

As in Week 1, we assume that treatment is well defined and that one person's outcome does not depend on other people's treatment assignments (SUTVA).

6. Deriving the local average treatment effect

The LATE theorem

For binary treatment and a binary instrument, under **independence**, **exclusion**, **monotonicity** ($D_i^1 \geq D_i^0$), and **relevance**,

$$\frac{E[Y_i | Z_i = 1] - E[Y_i | Z_i = 0]}{E[D_i | Z_i = 1] - E[D_i | Z_i = 0]} = E[Y_i^1 - Y_i^0 | D_i^1 > D_i^0] = \tau_{\text{LATE}}.$$

The Wald ratio identifies the **average treatment effect for compliers**: people whose treatment status is changed by the instrument.

This is the LATE theorem of Imbens and Angrist (1994). The following steps show how each assumption enters the proof.

The reduced form

$$\begin{aligned} E[Y_i | Z_i = 1] &= E[Y_i^0 + (Y_i^1 - Y_i^0) D_i | Z_i = 1] \\ &= E[Y_i^0 + (Y_i^1 - Y_i^0) D_i^1]. \end{aligned}$$

The first equality uses **exclusion**: Z_i affects Y_i through treatment. The second uses $D_i = D_i^1$ when $Z_i = 1$ and **independence** to remove the conditioning on Z_i .

For clarity, the second step can be written in full as

$$E [Y_i^0 + (Y_i^1 - Y_i^0) D_i \mid Z_i = 1] = E [Y_i^0 + (Y_i^1 - Y_i^0) D_i^1 \mid Z_i = 1] = E [Y_i^0 + (Y_i^1 - Y_i^0) D_i^1] .$$

The first equality substitutes the treatment status that is observed when $Z_i = 1$; independence justifies the second equality. Similarly,

$$\begin{aligned} E [Y_i \mid Z_i = 0] &= E [Y_i^0 + (Y_i^1 - Y_i^0) D_i \mid Z_i = 0] \\ &= E [Y_i^0 + (Y_i^1 - Y_i^0) D_i^0] . \end{aligned}$$

Therefore the numerator of the Wald ratio is

$$E [Y_i \mid Z_i = 1] - E [Y_i \mid Z_i = 0] = E [(Y_i^1 - Y_i^0) (D_i^1 - D_i^0)] .$$

Using monotonicity

Monotonicity implies that $D_i^1 - D_i^0$ can equal **zero or one**. It excludes -1 as a possible value.

- For compliers, $D_i^1 - D_i^0 = 1$.
- For always-takers and never-takers, $D_i^1 - D_i^0 = 0$.

The product is zero for everyone except compliers. Its population mean is therefore the mean effect for compliers multiplied by their population share:

$$\begin{aligned} E [(Y_i^1 - Y_i^0) (D_i^1 - D_i^0)] \\ = E [Y_i^1 - Y_i^0 \mid D_i^1 > D_i^0] P (D_i^1 > D_i^0) . \end{aligned}$$

The outcome difference between the instrument groups equals the **average effect for compliers multiplied by their share of the population**.

The first stage

Now consider the denominator of the Wald ratio:

$$\begin{aligned} E [D_i \mid Z_i = 1] - E [D_i \mid Z_i = 0] &= E [D_i^1 - D_i^0] \\ &= P (D_i^1 > D_i^0) . \end{aligned}$$

The first equality follows from **independence**.

The second follows from **monotonicity**: the change in treatment equals one for compliers and zero for everyone else.

The first stage is therefore the **proportion of compliers**. **Relevance** ensures this proportion is positive, so we can divide by it.

Dividing the reduced form by the first stage

Putting the numerator and denominator together,

$$\begin{aligned}
& \frac{E[Y_i | Z_i = 1] - E[Y_i | Z_i = 0]}{E[D_i | Z_i = 1] - E[D_i | Z_i = 0]} \\
&= \frac{E[Y_i^1 - Y_i^0 | D_i^1 > D_i^0] P(D_i^1 > D_i^0)}{P(D_i^1 > D_i^0)} \\
&= E[Y_i^1 - Y_i^0 | D_i^1 > D_i^0].
\end{aligned}$$

The complier share cancels. The Wald ratio identifies the **local average treatment effect**.

The LATE theorem: proved

We have now established the result stated above.

For binary treatment and a binary instrument, under **independence, exclusion, monotonicity** ($D_i^1 \geq D_i^0$), and **relevance**,

$$\frac{E[Y_i | Z_i = 1] - E[Y_i | Z_i = 0]}{E[D_i | Z_i = 1] - E[D_i | Z_i = 0]} = E[Y_i^1 - Y_i^0 | D_i^1 > D_i^0] = \tau_{\text{LATE}}.$$

The Wald ratio identifies the **average treatment effect for compliers**: people whose treatment status is changed by the instrument.

“Local” refers to the compliers for this particular instrument.

Why monotonicity matters

If there are defiers as well as compliers, the reduced form becomes

$$\delta_1 = P(C) \tau_C - P(F) \tau_F,$$

and the first stage is $\pi_1 = P(C) - P(F)$. Here C and F denote compliers and defiers. The ratio subtracts the contribution of defiers. Even when treatment helps both groups, their responses to the instrument can offset each other in the reduced form.

LATE, ATT, and ATE

The **ATE** averages effects across everyone in the population. The **ATT** averages across those observed receiving treatment. Under monotonicity and random assignment, recipients include always-takers and the compliers assigned $Z_i = 1$. Their average effect can differ from the average effect among compliers.

If nobody can receive treatment without the offer, there are no always-takers. Under the assumptions above, treated individuals are then the randomly offered compliers, and LATE also identifies the ATT. In Oregon, some people in the lottery control group obtained Medicaid through other routes, so we retain the complier interpretation.

Your turn

Explain why two valid instruments for the same treatment could produce different IV estimates. What would you need to know about the people whose treatment choices each instrument changes?

Conditional random assignment

When independence holds within groups defined by X_i , the same proof applies **within each group**. With indicators for all such groups and a binary instrument, pooled 2SLS combines the group-specific LATEs. Its weights are proportional to the group's population share, the within-group variance of the instrument, and its first stage. With positive first stages these weights are nonnegative.

This matters in Oregon: the adjusted first stage and IV coefficient summarise comparisons within the relevant lottery and survey groups. Calling the adjusted first-stage coefficient the unconditional fraction of compliers would require additional qualifications about these weights.

7. The Oregon experiment

Policy and assignment

Your turn

Explain the policy problem and the lottery. What opportunity did winning provide? Distinguish the treatment and control groups defined by lottery assignment from the groups with and without Medicaid coverage.

In 2008, Oregon used a lottery to select people from a waiting list for the opportunity to apply for **OHP Standard**, a Medicaid programme for low-income adults. Applicants still had to complete the application and meet eligibility criteria. Some selected individuals did not obtain coverage; some unselected individuals obtained Medicaid through other routes.

The study's analysis population contains **74,922 individuals**, of whom **29,834** were selected. These are the analysis-sample counts after the authors' exclusions, rather than the initial lottery totals. See Sections II–III of Finkelstein et al. (2012).

Data

Your turn

Describe the administrative and survey data. What outcomes can each measure? Why do the analysis samples differ? What concerns arise when only some people respond to a survey?

The authors link lottery information to Medicaid enrolment, hospital discharge, credit report, and mortality records. A mail survey supplies information on health care use, financial strain, and self-reported health. The survey has an effective weighted response rate of approximately 50%. Its design includes intensive follow-up of a subsample of nonrespondents; the supplied survey weights account for that sampling design.

The exercise uses enrolment and survey outcomes from the [Oregon Health Insurance Experiment public-use data](#). Download the required data and documentation from the Week 3 section of Moodle. The exercise gives the file names, import code, and variables to use.

Balance: Table II

Your turn

Explain what Table II tests. Compare the match and response rates in Panel A and the joint balance tests in Panel B. How does the evidence bear on the validity of the comparisons, particularly among survey respondents?

TABLE II
TREATMENT: CONTROL BALANCE

	Control mean (std. dev.) for full sample	Difference between treatment and control		
		Full sample (2)	Credit report subsample (3)	Survey respondents subsample (4)
	(1)			
Panel A: Match/response rates				
Matched in September 2009 credit data	0.663 (0.473)		−0.0043 (0.0037) [0.247]	
Responded to survey	0.506 (0.500)			−0.016 (0.0066) [0.014]
Response time (in days)	53.0 (57.8)			1.638 (1.088) [0.132]
Panel B: Prerandomization characteristics				
Lottery list variables				
<i>F</i> -statistic		1.286	0.553	0.574
[<i>p</i> -value]		[0.239]	[0.836]	[0.820]
Pre-randomization outcomes				
<i>F</i> -statistic		0.543	0.921	1.266
[<i>p</i> -value]		[0.844]	[0.518]	[0.281]
Both of the above				
<i>F</i> -statistic		0.915	0.793	0.782
[<i>p</i> -value]		[0.56]	[0.726]	[0.680]
<i>N</i>		74,922	49,980	23,741

Notes. Standard errors in parentheses; per comparison *p*-values in square brackets. In Panel A, we analyze match and response rates. The first column reports the mean and standard deviation for the control sample of the outcome shown in the left column. Columns (3) and (4) report estimated differences between treatments and controls for the dependent variable shown in the left column and the sample indicated in the column heading. Specifically they report the coefficient on *LOTTERY* based on estimating equation (1). All regressions include household fixed effects and cluster on household. In addition, in column (4) we include survey wave fixed effects and the interaction of survey wave fixed effects and household fixed effects and use survey weights. The full sample (i.e., the sample used in the hospital discharge and mortality data) is used in column (3), and the entire survey sample is used in column (4). In Panel B we report the *F*-statistic and *p*-value from regressing multiple prerandomization characteristics on *LOTTERY* in equation (1). “Lottery list variables” are common across all three samples and consist of nine demographic variables derived from information provided at the time of lottery sign up: year of birth; sex; whether English is the preferred language for receiving materials; whether the individual signed him- or herself up for the lottery or were signed up by a household member; whether the individual gave his or her address as a PO box; whether he or she signed up the first day the lottery list was open; the median household income in the 2000 census from their ZIP code; whether the ZIP code given is within a census-defined MSA; and whether he or she provided a phone number on sign up. “Prerandomization outcomes” are specific to the sample (we look at the hospital outcomes that we subsequently analyze for column (2), the credit report outcomes we subsequently analyze for column (3), and a few measures from each that approximate survey questions we subsequently analyze for column (4)). More detail on the prerandomization outcomes, the exact regression specifications, and the results for each variable analyzed in Panel B are presented in Online Appendix 3, Table A13.

Finkelstein et al. (2012), Table II.

Read Panel A separately from Panel B. Similar measured baseline characteristics are reassuring about the observed samples. Selection into responding could still depend on unobserved outcomes. The paper discusses sensitivity to this concern.

First stage: Table III

Your turn

Explain the first row of Table III, including the control mean, estimated first stage, units, and standard error. Why is the first stage less than one? Why might it differ across samples? Also consider the food-stamp results when assessing exclusion.

TABLE III
FIRST-STAGE ESTIMATES

	Full sample		Credit report subsample		Survey respondents	
	Control mean (1)	Estimated FS (2)	Control mean (3)	Estimated FS (4)	Control mean (5)	Estimated FS (6)
(1) Ever on Medicaid	0.141	0.256 (0.0035)	0.135	0.255 (0.0042)	0.135	0.290 (0.0066)
(2) Ever on OHP Standard	0.027	0.264 (0.0029)	0.028	0.264 (0.0036)	0.026	0.302 (0.0055)
(3) # of months on Medicaid	1.408	3.355 (0.045)	1.352	3.366 (0.055)	1.509	3.943 (0.090)
(4) On Medicaid, end of study period	0.106	0.148 (0.0031)	0.101	0.151 (0.0038)	0.105	0.189 (0.0061)
(5) Currently have any insurance (self-report)					0.325	0.179 (0.0077)
(6) Currently have private insurance (self-report)					0.128	−0.0076 (0.0053)
(7) Currently on Medicaid (self-report)					0.117	0.197 (0.0063)
(8) Currently on Medicaid					0.105	0.191 (0.0060)
(9) Ever on TANF	0.031	0.0011 (0.0013)	0.028	0.0021 (0.0016)	0.023	0.0019 (0.0025)
(10) TANF benefits (\$)	124	−1.659 (5.813)	111	1.543 (6.571)	100	−4.991 (10.884)
(11) Ever on food stamps	0.606	0.017 (0.0029)	0.594	0.018 (0.0035)	0.622	0.023 (0.0054)
(12) Food stamp benefits (\$)	1776	61.3 (15.0)	1787	60.0 (18.8)	2202	122.4 (33.4)
<i>N</i>		74,922		49,980		23,741

Notes. Standard errors in parentheses. Even-numbered columns report the coefficient and standard error on *LOTTERY* from estimating the first-stage equation (4) with the dependent variable *INSURANCE* defined in the left column; odd-numbered columns report the control mean for that measure of *INSURANCE*. “Full sample” is the sample analyzed in the hospital discharge and mortality data. All regressions include dummies for household size and adjust standard errors for household clusters. The regressions in columns (2) and (4) also include lottery draw dummies; the regressions in column (6) also include dummies for survey wave and survey wave interacted with household size dummies, and use survey weights. The insurance measures are taken from the Medicaid enrollment administrative data except for those labeled “self-report” (rows (5) through (7)) which are taken from the survey. In the survey, respondents could report various types of insurance; we define “private insurance” as employer or private insurance and “any insurance” as Medicaid, Medicare, employer, private, or other insurance. In rows (1) and (2), “ever” refers to enrollment ever during the study period, as defined in the text. In row (3), “# of months” refers to number of months enrolled during the study period. In row (8), insurance is measured as being on Medicaid according to the state Medicaid enrollment data on the day the survey was returned. All outcomes are measured for the individual except that in row (10) (12) where TANF (food stamp) benefits measure total household benefits received over the study period. For purposes of measuring the first stage, the study period is defined as ending on September 30, 2009. For the second-stage outcomes that we analyze in the administrative data, the study period in the first stage begins with the notification date (which varies by lottery draw). For the second stage outcomes that we analyze in the survey data, the study period in the first stage begins on the first notification date (March 10, 2008).

Finkelstein et al. (2012), Table III.

In the full sample, lottery selection raises the probability of ever having Medicaid during the study period by approximately **25.6 percentage points**. The comparable estimate among survey respondents is about **29 percentage points**. These are increases in enrolment associated with selection, rather than the enrolment rate of all lottery winners.

The empirical specification

For the study’s outcome equation, write

$$Y_i = \beta_0 + \beta_1 D_i + \sum_{k=1}^K \beta_{k+1} X_{ik} + u_i.$$

Lottery selection Z_i instruments Medicaid coverage D_i . The authors estimate the **ITT** by regressing Y_i on lottery selection and the relevant controls. They estimate the **first stage** by regressing Medicaid coverage on the same instrument and controls.

Three design details matter:

- **Household-list size:** when any listed household member won, the household could apply. Include indicators for the number of household members on the lottery list.
- **Survey design:** for survey outcomes, also include survey-wave indicators and their interactions with household-list-size indicators, and use the supplied survey weights. The exercise represents each observed combination with one categorical variable.
- **Inference:** cluster standard errors by household. This permits heteroskedasticity and correlation among people in the same household.

For administrative outcomes, the paper also includes lottery-draw indicators and, where available, the corresponding pre-randomisation outcome. Keep each reported outcome paired with the first stage for its own estimation sample and specification.

Hospital use: Table IV

Your turn

Interpret Panel A of Table IV. Distinguish the control mean, ITT, and LATE. Convert the first row's effects into percentage points. What does the uncertainty permit you to conclude about admissions through the emergency room?

TABLE IV
HOSPITAL UTILIZATION

	Control mean (1)	ITT (2)	LATE (3)	<i>p</i> -values (4)
Panel A: Extensive margin				
All hospital admissions	0.067 (0.250)	0.0054 (0.0019)	0.021 (0.0074)	[0.004]
Admissions through ER	0.048 (0.214)	0.0018 (0.0016)	0.0070 (0.0062)	[0.265]
Admissions not through ER	0.029 (0.167)	0.0041 (0.0013)	0.016 (0.0051)	[0.002]
Panel B: All hospital admissions				
Days	0.498 (3.795)	0.026 (0.027)	0.101 (0.104)	[0.329] {0.328}
List charges	2,613 (19,942)	258 (146)	1,009 (569)	[0.077] {0.106}
Procedures	0.155 (1.08)	0.018 (0.0083)	0.070 (0.032)	[0.031] {0.059}
Standardized treatment effect		0.012 (0.0067)	0.047 (0.026)	[0.073]
Panel C: Admissions through ER				
Days	0.299 (2.326)	0.023 (0.017)	0.089 (0.067)	[0.183] {0.187}
List charges	1,502 (12,749)	163 (96)	636 (376)	[0.091] {0.171}
Procedures	0.081 (0.694)	0.0080 (0.0054)	0.031 (0.021)	[0.135] {0.187}
Standardized treatment effect		0.011 (0.0069)	0.044 (0.027)	[0.100]
Panel D: Admissions not through ER				
Days	0.199 (2.38)	0.0033 (0.017)	0.013 (0.065)	[0.841] {0.842}
List charges	1,110 (12,422)	98 (91)	384 (356)	[0.281] {0.383}
Procedures	0.075 (0.708)	0.010 (0.0056)	0.038 (0.022)	[0.080] {0.162}
Standardized treatment effect		0.0077 (0.0068)	0.030 (0.026)	[0.254]

Notes. Standard errors in parentheses; per comparison *p*-values in square brackets; family-wise *p*-values in curly brackets. Table investigates non-childbirth-related hospitalizations during the time period from notification date to August 31, 2009. All outcomes are measured unconditionally (i.e., are not conditional on admission). Column (2) reports the coefficient and standard error on *LOTTERY* from estimating equation (1) by OLS. Column (3) reports the coefficient and standard error on *INSURANCE* from estimating equation (3) by IV; for the IV estimates in column (3), the endogenous variable *INSURANCE* is defined as “ever on Medicaid” during our study period and the first stage is given in the first row of Table III. Column (4) reports the per comparison *p*-value and (where applicable) the family-wise *p*-value across the three different measures of utilization used to create the standardized treatment effect. Standardized treatment effect reports results based on equation (2). All regressions include household size fixed effect, lottery draw fixed effects, and the analogous outcome measure for the time period from January 1, 2008, through the notification date. All standard errors are clustered on the household. Sample consists of entire sample universe ($N = 74,922$).

Finkelstein et al. (2012), Table IV.

For any hospital admission, the reported ITT is 0.0054 and the LATE is approximately 0.021. The latter corresponds to a **2.1-percentage-point increase** in admission probability for the relevant compliers. The simple calculation $0.0054/0.256 \approx 0.021$ illustrates the rescaling. Exact calculations use unrounded estimates and the same covariates and sample. Admissions through an emergency room are hospital admissions; they do not include every emergency-department visit.

Survey health care use: Table V

Your turn

Interpret the outpatient-visit row of Table V. Explain the difference between any visit and number of visits. Which first stage from Table III belongs with the survey results? What is the estimated effect of Medicaid on the probability of any outpatient visit?

TABLE V
HEALTH CARE UTILIZATION (SURVEY DATA)

	Extensive margin (any)				Total utilization (number)			
	Control mean (1)	ITT (2)	LATE (3)	<i>p</i> -values (4)	Control mean (5)	ITT (6)	LATE (7)	<i>p</i> -values (8)
Prescription drugs currently	0.637 (0.481)	0.025 (0.0083)	0.088 (0.029)	[0.002] {0.005}	2.318 (2.878)	0.100 (0.051)	0.347 (0.176)	[0.049] {0.137}
Outpatient visits last six months	0.574 (0.494)	0.062 (0.0074)	0.212 (0.025)	[<0.0001] {<0.0001}	1.914 (3.087)	0.314 (0.054)	1.083 (0.182)	[<0.0001] {<0.0001}
ER visits last six months	0.261 (0.439)	0.0065 (0.0067)	0.022 (0.023)	[0.335] {0.547}	0.47 (1.037)	0.0074 (0.016)	0.026 (0.056)	[0.645] {0.643}
Inpatient hospital admissions last six months	0.072 (0.259)	0.0022 (0.0040)	0.0077 (0.014)	[0.572] {0.570}	0.097 (0.4)	0.0062 (0.0062)	0.021 (0.021)	[0.311] {0.510}
Standardized treatment effect		0.050 (0.011)	0.173 (0.036)	[<0.0001]		0.040 (0.011)	0.137 (0.038)	[0.0003]
Annual spending ^a					3,156 (108)	226 (108)	778 (371)	[0.037]

Notes. Standard errors in parentheses; per comparison *p*-values in square brackets; family-wise *p*-values in curly brackets. Hospital admissions exclude childbirth. Columns (2) and (6) report the coefficient and standard error on *LOTTERY* from estimating equation (1) by OLS. Columns (3) and (7) report the coefficient and standard error on *INSURANCE* from estimating equation (3) by IV; for the IV estimates in column (3), the endogenous variable *INSURANCE* is defined as “ever on Medicaid” during our study period and the first stage is given in the first row of Table III. Columns (4) and (8) report the per-comparison *p*-value and the family-wise *p*-value across the four different measures of utilization used to create the standardized treatment effect. Standardized treatment effect reports results based on equation (2). All regressions include household size fixed effects, survey wave fixed effects, and the interaction between the two. All standard errors are clustered on the household and all regressions are weighted using survey weights. Sample consists of survey responders ($N=23,741$).

^aTo calculate the implied spending effects associated with the estimated utilization effects, we use data from the 2002–2007 (pooled) Medical Expenditure Panel Survey (MEPS) on expenditures of all nonelderly (age 19–64) adults below 100% of poverty who are publicly insured. This gives us a total sample of over 7,500 individuals. We use their expenditures (all inflated with the CPI-U to 2007 dollars) to calculate average expenditures per outpatient visit, average expenditures per ER visit, average expenditures per inpatient visit (for visits not related to childbirth), and average semi-annual (six-month) spending on prescription drug. All spending is total expenditures (i.e., not just insured) expenditures. The underlying costs are \$150 per outpatient visit, \$435 per ER visit, \$7,523 per inpatient visit, and \$156 six-month expenditure per current prescription drug; we multiply these all by two to get annual costs.

Finkelstein et al. (2012), Table V.

The reported ITT for any outpatient visit is about **6.2 percentage points**, and the LATE is about **21.2 percentage points**. The after-class exercise estimates this relationship from the public survey files.

Financial strain: Table VIII

Your turn

Interpret the effects in Table VIII. Use at least two outcomes, discussing magnitude and uncertainty. Why might a change in out-of-pocket spending or medical debt matter for evaluating health insurance?

TABLE VIII
FINANCIAL STRAIN (SURVEY DATA)

	Control mean (1)	ITT (2)	LATE (3)	<i>p</i> -values (4)
Any out of pocket medical expenses, last six months	0.555 (0.497)	−0.058 (0.0077)	−0.200 (0.026)	[<0.0001] {<0.0001}
Owe money for medical expenses currently	0.597 (0.491)	−0.052 (0.0076)	−0.180 (0.026)	[<0.0001] {<0.0001}
Borrowed money or skipped other bills to pay medical bills, last six months	0.364 (0.481)	−0.045 (0.0073)	−0.154 (0.025)	[<0.0001] {<0.0001}
Refused treatment because of med- ical debt, last six months	0.081 (0.273)	−0.011 (0.0041)	−0.036 (0.014)	[0.01] {0.01}
Standardized treatment effect		−0.089 (0.010)	−0.305 (0.035)	[<0.0001]

Notes. Standard errors in parentheses; per comparison *p*-values in square brackets; family-wise *p*-values in curly brackets. Column (2) reports the coefficient and standard error on *LOTTERY* from estimating equation (1) by OLS. Column (3) reports the coefficient and standard error on *INSURANCE* from estimating equation (3) by IV; for the IV estimates in column (3), the endogenous variable *INSURANCE* is defined as “ever on Medicaid” during our study period and the first stage is given in the first row of Table III. Column (4) reports the per comparison *p*-value and the family-wise *p*-value across the four different measures of financial strain used to create the standardized treatment effect. Standardized treatment effect reports results based on equation (2). All regressions include household size fixed effects, survey wave fixed effects, and the interaction between the two. All standard errors are clustered on the household and all regressions are weighted using survey weights. Sample consists of survey responders ($N=23,741$).

Finkelstein et al. (2012), Table VIII.

Health and external validity: Table IX

Your turn

Explain the evidence on health in Table IX. Distinguish the administrative outcome from self-reported outcomes, and distinguish imprecise estimates from evidence of a small effect. Whose treatment effect is estimated, and how far would you generalise the findings to another insurance expansion?

TABLE IX
HEALTH

	Control mean (1)	ITT (2)	LATE (3)	p-values (4)
Panel A: Administrative data				
Alive	0.992 (0.092)	0.00032 (0.00068)	0.0013 (0.0027)	[0.638]
Panel B: Survey data				
Self-reported health good/very good/excellent (not fair or poor)	0.548 (0.498)	0.039 (0.0076)	0.133 (0.026)	[<0.0001] [<0.0001]
Self-reported health not poor (fair, good, very good, or excellent)	0.86 (0.347)	0.029 (0.0051)	0.099 (0.018)	[<0.0001] [<0.0001]
Health about the same or gotten better over last six months	0.714 (0.452)	0.033 (0.0067)	0.113 (0.023)	[<0.0001] [<0.0001]
# of days physical health good, past 30 days*	21.862 (10.384)	0.381 (0.162)	1.317 (0.563)	[0.019] [0.018]
# days poor physical or mental health did not impair usual activity, past 30 days*	20.329 (10.939)	0.459 (0.175)	1.585 (0.606)	[0.009] [0.015]
# of days mental health good, past 30 days*	18.738 (11.445)	0.603 (0.184)	2.082 (0.64)	[0.001] [0.003]
Did not screen positive for depression, last two weeks	0.671 (0.470)	0.023 (0.0071)	0.078 (0.025)	[0.001] [0.003]
Standardized treatment effect		0.059 (0.011)	0.203 (0.039)	[<0.0001]

Notes. Standard errors in parentheses; per comparison *p*-values in square brackets; family-wise *p*-values in curly brackets. Column (2) reports the coefficient and standard error on *LOTTERY* from estimating equation (1) by OLS. Column (3) reports the coefficient and standard error on *INSURANCE* from estimating equation (3) by IV; for the IV estimates in column (3), the endogenous variable *INSURANCE* is defined as “ever on Medicaid” during our study period and the first stage is given in the first row of Table III. Column (4) reports the per comparison *p*-value and the family-wise *p*-value across the different measures used to create the standardized treatment effect. Standardized treatment effect reports results based on equation (2). All regressions include household size fixed effects and standard errors are clustered on the household. The regressions in Panel A include lottery draw fixed effects, and the dependent variable “alive” is measured from the notification date through September 2009 ($N=74,922$). The regressions in Panel B include survey wave fixed effects and the interaction of survey wave fixed effects with household size fixed effects, and are weighted using the survey weights ($N=23,741$).

*These questions were worded to ask about number of days health “not good” or “impaired”; we switched the sign for consistency with the other measures. See Online Appendix Figure A4 for the exact survey wording.

Finkelstein et al. (2012), Table IX.

The population entered a waiting list for a particular programme in a particular setting. The IV estimate applies to people whose Medicaid coverage was changed by the lottery. Generalising requires considering eligibility, existing access to care, the insurance package, provider capacity, and the follow-up period.

Assessing the IV assumptions

The lottery provides the basis for independence conditional on household-list size. The first-stage estimates establish relevance. Monotonicity asks whether winning could cause someone to lose Medicaid coverage that they would otherwise have obtained. Exclusion asks whether selection affects outcomes through any route other than the Medicaid coverage being instrumented.

Table III also reports changes in food-stamp receipt. Consider whether assistance with applications or information could influence other programmes directly, or whether other-programme use changes as a consequence of obtaining Medicaid. A downstream consequence of Medicaid can be part of its total effect; a separate pathway from the lottery to an outcome raises an exclusion concern. Similarly, effects on a person’s outcome through another household member’s coverage need attention when interpreting an IV for the person’s own coverage.

Internal validity concerns the credibility of the causal interpretation for the study’s comparison. **External validity** concerns the applicability of that effect to other populations, policies, and settings.

8. After class

The [ungraded exercise](#) asks you to estimate the first stage, the ITT, and a 2SLS effect for outpatient use; verify the ratio relationship; and interpret the result for compliers. Use Quarto to keep the code, output, and explanation together.

References

- Angrist, J. D., and A. B. Krueger (2001). “Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments.” *Journal of Economic Perspectives* 15(4): 69–85. [DOI](#).
- Angrist, J. D., and J.-S. Pischke (2009). *Mostly Harmless Econometrics: An Empiricist’s Companion*. Princeton University Press, Chapter 4. [Book](#).
- Angrist, J. D., and J.-S. Pischke (2015). *Mastering ’Metrics: The Path from Cause to Effect*. Princeton University Press, Chapter 3. [Companion resources](#).
- Cunningham, S. *Causal Inference: The Remix*. Online edition, Chapter 7. [Chapter](#).
- Finkelstein, A., S. Taubman, B. Wright, M. Bernstein, J. Gruber, J. P. Newhouse, H. Allen, K. Baicker, and the Oregon Health Study Group (2012). “The Oregon Health Insurance Experiment: Evidence from the First Year.” *Quarterly Journal of Economics* 127(3): 1057–1106. [DOI](#).
- Huntington-Klein, N. *The Effect: An Introduction to Research Design and Causality*. Online edition, Chapter 19. [Chapter](#).
- Imbens, G. W., and J. D. Angrist (1994). “Identification and Estimation of Local Average Treatment Effects.” *Econometrica* 62(2): 467–475. [DOI](#).