# Reading a regression table

| Outcome: Happiness | (1) | (2) |
|---|---|---|
| Gender $_i$ | 0.834 | 0.614 |
| | (0.032) | (0.045) |
| Education $_i$ | | −0.739 |
| | | (0.036) |

Notes: Estimations contain a constant term. Standard errors in parentheses.

- Gender is a dummy variable. Average happiness of individuals with gender=1 is 0.834 higher than of individuals with gender=0. This is significant at the 5% level, because the coefficient is more than twice the standard error
- Conditional on education (keeping education fixed/comparing people with the same level of education), gender=1 is associated with 0.614 higher happiness.
- Conditional on gender, each additional year of education is associated with 0.739 less points on the happiness scale.

# Omitted Variable Bias

Let $Y_i$ be the outcome variable, $X_i$ our regressor of interest, $W$ a series of control variables, and $Z_i$ the "omitted" variable.

$$\begin{aligned} \text{[Long regression]} \quad & Y_i = c_1 + \beta_L X_i + \lambda Z_i + W\pi + e_i \\ \text{[Short regression]} \quad & Y_i = c_2 + \beta_S X_i + W\pi + u_i \\ \text{[Auxiliary regression]} \quad & Z_i = c_3 + \pi_1 X_i + W\pi + v_i \end{aligned}$$

Then, the **Omitted variable bias formula** states that:

$$\underbrace{\beta_S}_{\text{Short =}} = \underbrace{\beta_L}_{\text{Long +}} + \underbrace{\lambda}_{\text{Omitted} \times} \cdot \underbrace{\pi_1}_{\text{Included}}$$

The OVB formula describes what happens to our coefficient of interest, $\beta$, as we include one additional variable $Z$ in the regression. We call $\lambda\pi_1$ the **omitted variable bias**. Direction of bias: multiply our guesses for the signs of $\delta$ and $\gamma$.

# Good and Bad controls

- Some controls are called ”bad controls”. These are:
    1. Variables that are themselves outcomes of a treatment: Treatment $\rightarrow$ Bad Control
    2. Variables that moderate the treatment effect: Treatment $\rightarrow$ Bad Control $\rightarrow$ Outcome
- **Rule of Thumb: Good controls are either pre-determined or immutable characteristics.**
- Another way to think about it: Controls help us make ”apples to apples” comparisons. We want to compare units that, in the absence of the treatment, would have the same outcomes, and differ only because they have different levels of the treatment.

# Logs: Cheatsheet

| Model | LHS | RHS | A change in $x$ by ... | is associated with a change in $y$ by ... |
|---|---|---|---|---|
| Level-Level | y | x | 1 unit | $\beta_1$ units |
| Level-Log | y | $\log(x)$ | 1% | $\beta_1/100$ units |
| Log-Level | $\log(y)$ | x | 1 unit | $100\beta_1$% |
| Log-Log | $\log(y)$ | $\log(x)$ | 1% | $\beta_1$% |

If you want to get a bonus star from me, write ”approximately” in log-interpretations.

# Hypothesis testing

$$\left|\frac{\hat{\beta}}{\mathrm{SE}(\hat{\beta})}\right| \geq 1.96$$

$$\Leftrightarrow |\,\text{t-stat}\,| \geq 1.96$$

$$\Leftrightarrow \text{p-value} \leq 0.05$$

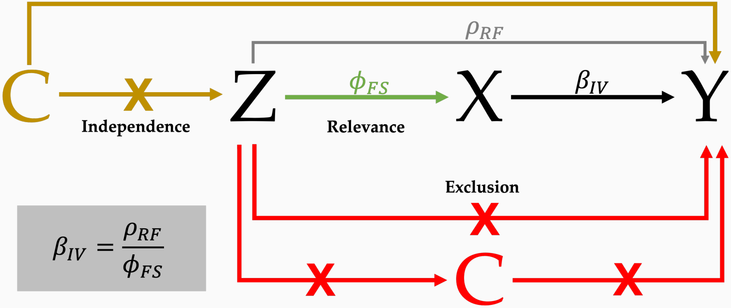
$$\Leftrightarrow 0 \notin \mathrm{CI}$$

When testing the null hypothesis $H_0$: $\beta = 0$ (against the alternative hypothesis $H_0$: $\beta \neq 0$), then if any of these conditions holds, we reject the null.

# IV summary

We need the following three assumptions for IV to work:

1. **Relevance**: $Z$ must truly affect $X$
2. **Independence/Exogeneity**: $Z$ is as good as randomly assigned
3. **Exclusion Restriction**: The **only** way that $Z$ affects $Y$ is via $X$.




$$\beta_{IV} = \frac{\rho_{RF}}{\phi_{FS}}$$

# IV: The LATE is the treatment effect for the compliers

| Potential outcomes! (unobserved) | | ***Does not get voucher (Z=0)*** | |
|---|---|---|---|
| | | *Eats chocolate (D=1)* | *Does not eat chocolate (D=0)* |
| ***Gets voucher (Z=1)*** | *Eats chocolate (D=1)* | Always-takers:<br>$E(D \mid Z=1)=E(D \mid Z=0)=1$<br>➔ $E(Y \mid Z=1)=E(Y \mid Z=0)$ | Compliers |
| | *Does not eat chocolate (D=0)* | Defiers | Never-takers:<br>$E(D \mid Z=1)=E(D \mid Z=0)=0$<br>➔ $E(Y \mid Z=1)=E(Y \mid Z=0)$ |

| | Before | After |
|---|---|---|
| Treated | $\alpha+\beta$ | $\alpha+\beta+\gamma+\delta$ |
| Untreated | $\alpha$ | $\alpha+\gamma$ |

We can calculate the DiD estimate in two ways:

$$DiD = E[\underbrace{\left(Y_{i1}^T - Y_{i0}^T\right)}_{\text{Change for treated}} - \underbrace{\left(Y_{i1}^C - Y_{i0}^C\right)}_{\text{Change for untreated}}] = [(\alpha+\beta+\gamma+\delta)-(\alpha+\beta)] - [(\alpha+\gamma)-(\alpha)]$$

$$= E[\underbrace{\left(Y_{i1}^T - Y_{i1}^C\right)}_{\text{After-difference}} - \underbrace{\left(Y_{i0}^T - Y_{i0}^C\right)}_{\text{Before-difference}}] = [(\alpha+\beta+\gamma+\delta)-(\alpha+\gamma)] - [(\alpha+\beta)-(\alpha)]$$

We need to assume parallel trends **after the treatment** for causality. Verify using data **before the treatment**. Estimate DiD with regression:

$$Y_{it} = \alpha + \beta \text{ Treated }_i + \gamma \text{ Post }_t + \delta \text{ Treated }_i \cdot \text{ Post }_t + u_{it}$$

To generalize, estimate model with unit fixed effects $\alpha_i$ and time FE $\delta_t$:

$$Y_{it} = \alpha_i + \delta_t + \beta^{FE} \text{ X }_{it} + u_{it}$$

# Regression discontinuity designs

**Setup**: We need a running variable $X$, impacting a treatment $D$ discontinuously at a threshold, and an outcome $Y$.

**Assumption**: The effect of $X$ and any other variable $C$ on outcome $Y$ is smooth around the discontinuity. In particular: No strategic behavior around cutoff. Can do a ”balance check” around threshold.

**Sharp RD**: Probability of $D$ jumps from 0 to 1 at threshold. To estimate:

$$Y_i = \alpha + \beta_1 X_i + \beta_2 X_i \cdot D_i + \beta^{RD} D_i + \varepsilon_i$$

**Fuzzy RD**: Probability jumps at threshold. Then, we simply use IV and instrument for $D$ with being just above the threshold.

# How to approach an exam question

1. Think: About the question, about the real world
2. Start with the numbers you see
    - One-sentence summary
    - Direction (positive or negative?)
    - Statistical significance (significant or insignificant, level?)
    - (Economic) magnitude (big or small?)
3. Then: Establish whether estimated relationship is causal or not
    - What do the results mean? Correlation (interesting) or causality (policy-relevant)
    - Is X-variable randomized? Do we have valid counterfactuals?
    - If not: Do you expect bias? Of which sort (OVB, reverse causality, bad controls, . . . )?
    - Find a plausible story for bias (using the OVB formula)

There are no traps!