1 LATE

Consider the simple regression model:

$$Y = Y_0 + D(Y_1 - Y_0) = \beta_0 + \beta_1 D + U$$

where D is a binary treatment variable.

Problem 1.1. In general, ATT and the LATE will not be the same. This is because ATT is a weighted average of two effects: one on always-takers and one on compliers. Show that this is the case.

Solution. Note that $D = D_0 + Z(D_1 - D_0)$.

- ▷ Assuming monotonicity, $D = 1$ if and only if (1) $D_0 = 1$ (always takers) or (2) $Z = 1$ and $D_1 - D_0 = 1$. (compliers)
- ▷ Thus, we can rewrite the ATT using the law of total probability:

$$\begin{aligned}
&\mathbb{E}[Y_1 - Y_0 | D = 1] \\
=&\mathbb{E}[Y_1 - Y_0 | D = 1, D_0 = 1] P(D_0 = 1 | D = 1) \\
&+ \mathbb{E}[Y_1 - Y_0 | D_1 > D_0, Z = 1, D = 1] P(D_1 > D_0, Z = 1 | D = 1)
\end{aligned}$$

- ▷ Thus the ATT is equal to the ATT of the always takers ($\mathbb{E}[Y_1 - Y_0 | D = 1, D_0 = 1]$) multiplied by weight $P(D_0 = 1 | D = 1)$ plus the ATT of the compliers ($\mathbb{E}[Y_1 - Y_0 | D_1 > D_0, Z = 1, D = 1]$) multiplied by weight $P(D_1 > D_0, Z = 1 | D = 1)$.

Problem 1.2. When is LATE = ATE?

Solution. We consider two possibilities:

1. LATE is equal to ATE when the entire population consists of compliers.
   - ▷ With monotonicity, we can rule out the existence of defiers.
   - ▷ Thus, if there are no always-takers and no never-takers, LATE will be equal to ATE. This is because ATE can be written as a weighted average of the ATT on always takers, compliers, and never-takers by the law of total probability (as we did in the previous part).
2. LATE is equal to ATE when the MTE is constant.
   - ▷ If the MTE is constant, conditioning on D does not add any new information, so you have

$$\mathbb{E}[Y_1 - Y_0 | D_1 > D_0] = \mathbb{E}[Y_1 - Y_0]$$

2 More LATE

Consider the adjusted set of assumptions conditional on covariates:

- ▷ Exogeneity: $(Y_0, Y_1, D_0, D_1) \perp Z|X$
- ▷ Relevance: $P[D=1|X, Z=1] \neq P[D=1|X, Z=0]$ a.s.
- ▷ Monotonicity: $P[D_1 \geq D_0|X] = 1$ a.s.
- ▷ Overlap: $P[Z=1|X] \in (0,1)$ a.s.

Problem 2.1. Show that we can identify

$$LATE(x) \equiv \mathbb{E}\left[Y_1 - Y_0 | \underbrace{T=c}_{\text{compliers}}, X=x\right]$$

Solution. We can identify the $LATE(x)$ with the wald estimator:

$$\frac{\mathbb{E}[Y|Z=1, X=x] - \mathbb{E}[Y|Z=0, X=x]}{\mathbb{E}[D|Z=1, X=x] - \mathbb{E}[D|Z=0, X=0]}$$

We simplify the numerator and the denominator respectively:

- ▷ Numerator: $\mathbb{E}[Y|Z=1, X=x] - \mathbb{E}[Y|Z=0, X=x]$
  - * Note that

$$\mathbb{E}[Y|Z=1, X=x] = \mathbb{E}[Y_1 D_1 + Y_0(1-D_1)|X=x]$$
$$\mathbb{E}[Y|Z=0, X=x] = \mathbb{E}[Y_1 D_0 + Y_0(1-D_0)|X=x]$$

  - * Thus the numerator is equivalent to

$$\mathbb{E}[(Y_1 - Y_0)(D_1 - D_0)|X=x] = \mathbb{E}[Y_1 - Y_0|D_1 > D_0, X=x] P(D_1 > D_0|X=x)$$

- ▷ Denominator: $\mathbb{E}[D|Z=1, X=x] - \mathbb{E}[D|Z=0, X=0]$
  - * This is equivalent to

$$= \mathbb{E}[D_1|Z=1, X=x] - \mathbb{E}[D_0|Z=1, X=x]$$

  - * Assuming exogeneity:

$$= \mathbb{E}[D_1 - D_0|X=x]$$
$$= \mathbb{E}[1|D_1 - D_0 = 1, X=x] P(D_1 > D_0|X=x)$$
$$= P(D_1 > D_0|X=x)$$

Putting them together, we have

$$LATE\left(x\right)=\frac{\mathbb{E}\left[Y_1-Y_0|D_1>D_0|X=x\right]P\left(D_1>D_0|X=x\right)}{P\left(D_1>D_0|X=x\right)}=\mathbb{E}\left[Y_1-Y_0|\underbrace{D_1>D_0}_{\text{compliers}},X=x\right]$$

Problem 2.2. Show that

$$\mathbb{E}\left[Y_1-Y_0|T=c\right]=\mathbb{E}\left[LATE\left(X\right)\frac{P\left[T=c|X\right]}{P\left[T=c\right]}\right]$$

Solution. Note that $T=c$ corresponds to $D_1>D_0$.

▷ Since $LATE(x)$ is defined to be $\mathbb{E}\left[Y_1-Y_0|T=c,X=x\right]$, we have from LIE:

$$\begin{aligned}
\mathbb{E}\left[Y_1-Y_0|T=c\right]&=\mathbb{E}\left[\mathbb{E}\left[Y_1-Y_0|T=c,X=x\right]|T=c\right] && (\because)\ \text{LIE}\\
&=\mathbb{E}\left[LATE(x)|T=c\right]\\
&=\int_X LATE\left(x\right)f\left(x|T=c\right)dx\\
&=\int_X LATE\left(x\right)\frac{f\left(X=x,T=c\right)}{P\left(T=c\right)}dx && (\because)\ \text{Bayes Rule}\\
&=\int_X LATE\left(x\right)\frac{P\left(T=c|X=x\right)}{P\left(T=c\right)}f\left(x\right)dx && (\because)\ \text{Bayes Rule}\\
&=\mathbb{E}\left[LATE\left(x\right)\frac{P\left(T=c|X=x\right)}{P\left(T=c\right)}\right]
\end{aligned}$$

Problem 2.3. Suppose that X is discrete (a set of binary indicators). Excluded instruments are Z and XZ where $Z\in\{0,1\}$; the model is fully saturated. Show that

$$\beta_{TSLS}=\mathbb{E}\left[LATE\left(X\right)\frac{Var\left(p\left(X,Z\right)|X\right)}{\mathbb{E}\left[Var\left(p\left(X,Z\right)|X\right)\right]}\right]$$

where $p\left(X,Z\right)=Pr\left(D=1|X,Z\right)$.

Solution. Consider the following saturated regression:

$$\begin{aligned}
Y&=\beta_{TSLS}D+\alpha_X+\epsilon\\
D&=\pi_X Z+\gamma_X+u
\end{aligned}$$

where we include X-specific dummies in both stages and x-specific first-stage coefficients

▷ Note that:

$$\pi_X=\mathbb{E}\left[Z^2|X\right]^{-1}\mathbb{E}\left[ZD|X\right]$$

and define:

$$\begin{aligned}
\hat{D}&=\mathbb{E}\left[D|Z,X\right]=\pi_X Z+\gamma_X\\
\tilde{D}&=\hat{D}-\mathbb{E}\left[\hat{D}|X\right]
\end{aligned}$$

▷ We know that

$$\beta_{TSLS} = \frac{\text{Cov}\left[Y, \tilde{D}\right]}{\text{Var}\left[\tilde{D}\right]}$$

from the last problem set, since the second stage regression is of Y on $\hat{D}$ and X, and we can get the β on $\hat{D}$ by regressing Y on $\tilde{D}$, the residual of regressing $\hat{D}$ on X .

▷ The numerator can be written as:

$$\begin{aligned}
\text{Cov}\left[Y, \tilde{D}\right] &= \mathbb{E}\left[Y\tilde{D}\right] \\
&= \mathbb{E}\left[Y\left(\hat{D} - \mathbb{E}\left[\hat{D}|X\right]\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y\left(\hat{D} - \mathbb{E}\left[\hat{D}|X\right]\right)|Z, X\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y\hat{D}|Z, X\right] - \mathbb{E}\left[Y\mathbb{E}\left[\hat{D}|X\right]|Z, X\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y|Z, X\right]\left(\hat{D} - \mathbb{E}\left[\hat{D}|X\right]\right)\right] \qquad (\cdot.\cdot)\ \hat{D}, \mathbb{E}\left[\hat{D}|X\right] \text{ are functions of } Z \text{ and } X
\end{aligned}$$

Since $\mathbb{E}\left[Y|Z, X\right]$ can be simplifed as:

$$\begin{aligned}
\mathbb{E}\left[Y|Z, X\right] &= \mathbb{E}\left[Y|Z = 0, X\right] + Z\left(\mathbb{E}\left[Y|Z = 1, X\right] - \mathbb{E}\left[Y|Z = 0, X\right]\right) \\
&= \mathbb{E}\left[Y|Z = 0, X\right] + Z\frac{\mathbb{E}\left[Y|Z = 1, X\right] - \mathbb{E}\left[Y|Z = 0, X\right]}{\mathbb{E}\left[D|Z = 1, X\right] - \mathbb{E}\left[D|Z = 0, X\right]}\left(\mathbb{E}\left[D|Z = 1, X\right] - \mathbb{E}\left[D|Z = 0, X\right]\right) \\
&= \mathbb{E}\left[Y|Z = 0, X\right] + ZLATE\left(X\right)\left(\mathbb{E}\left[D|Z = 1, X\right] - \mathbb{E}\left[D|Z = 0, X\right]\right) \\
&= \mathbb{E}\left[Y|Z = 0, X\right] + LATE\left(X\right)\pi_X Z
\end{aligned}$$

Plugging in to the original expression yields:

$$\begin{aligned}
\text{Cov}\left[Y, \tilde{D}\right] &= \mathbb{E}\left[\left\{\mathbb{E}\left[Y|Z = 0, X\right] + LATE\left(X\right)\pi_X Z\right\}\left(\hat{D} - \mathbb{E}\left[\hat{D}|X\right]\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y|Z = 0, X\right]\left(\hat{D} - \mathbb{E}\left[\hat{D}|X\right]\right)\right] + \mathbb{E}\left[\mathbb{E}\left[LATE\left(X\right)\pi_X Z\left(\hat{D} - \mathbb{E}\left[\hat{D}|X\right]\right)\right.\right. \\
(\cdot.\cdot)\ \text{LIE} &= \mathbb{E}\left[\mathbb{E}\left[Y|Z = 0, X\right]\left(\mathbb{E}\left[\hat{D}|X\right] - \mathbb{E}\left[\hat{D}|X\right]\right) + \mathbb{E}\left[\mathbb{E}\left[LATE\left(X\right)\pi_X Z\left(\hat{D} - \mathbb{E}\left[\right.\right.\right.\right. \\
&= \mathbb{E}\left[LATE(X)\mathbb{E}\left[\pi_X Z\left(\hat{D} - \mathbb{E}\left[\hat{D}|X\right]\right)|X\right]\right] \\
(\cdot.\cdot)\ \mathbb{E}\left[\gamma_X\left(\hat{D} - \mathbb{E}\left[\hat{D}|X\right]\right)|X\right] = 0 &= \mathbb{E}\left[LATE(X)\mathbb{E}\left[\left(\pi_X Z + \gamma_X\right)\left(\hat{D} - \mathbb{E}\left[\hat{D}|X\right]\right)|X\right]\right] \\
&= \mathbb{E}\left[LATE(X)\mathbb{E}\left[\hat{D}\left(\hat{D} - \mathbb{E}\left[\hat{D}|X\right]\right)|X\right]\right] \\
&= \mathbb{E}\left[LATE(X)\left\{\mathbb{E}\left[\hat{D}^2|X\right] - \left(\mathbb{E}\left[\hat{D}|X\right]^2\right)\right\}\right] \\
&= \mathbb{E}\left[LATE\left(X\right)\text{Var}\left[\hat{D}|X\right]\right]
\end{aligned}$$

▷ The denominator can be written as:

$$\begin{aligned}
\text{Var}\left[\tilde{D}\right] &= \mathbb{E}\left[\tilde{D}^2\right] \\
&= \mathbb{E}\left[\left(\hat{D} - \mathbb{E}\left[\hat{D}|X\right]\right)^2\right] \\
&= \mathbb{E}\left[\hat{D}^2 - 2\hat{D}\mathbb{E}\left[\hat{D}|X\right] + \mathbb{E}\left[\hat{D}|X\right]^2\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\hat{D}^2 - 2\hat{D}\mathbb{E}\left[\hat{D}|X\right] + \mathbb{E}\left[\hat{D}|X\right]^2|X\right]\right] \qquad (\cdot.\cdot) \text{ LIE} \\
&= \mathbb{E}\left[\mathbb{E}\left[\hat{D}^2|X\right] - \mathbb{E}\left[\hat{D}|X\right]^2\right] \\
&= \mathbb{E}\left[\text{Var}\left[\hat{D}|X\right]\right]
\end{aligned}$$

▷ Collecting the numerator and the denominator, we have

$$\beta_{TSLS} = \frac{\text{Cov}\left[Y, \tilde{D}\right]}{\text{Var}\left[\tilde{D}\right]} = \frac{\mathbb{E}\left[LATE\left(X\right)\text{Var}\left[\hat{D}|X\right]\right]}{\mathbb{E}\left[\text{Var}\left[\hat{D}|X\right]\right]} = \mathbb{E}\left[LATE\left(X\right)\frac{\text{Var}\left[\hat{D}|X\right]}{\mathbb{E}\left[\text{Var}\left[\hat{D}|X\right]\right]}\right]$$

Problem 2.4. Show that for binary D and Z and any function $G = g\left(Y, X, D\right)$:

$$\mathbb{E}\left[G|T = c\right] = \frac{1}{P\left[T = c\right]}\mathbb{E}\left[\kappa G\right]$$

where

$$\kappa \equiv 1 - \frac{D\left(1 - Z\right)}{P\left(Z = 0|X\right)} - \frac{Z\left(1 - D\right)}{P\left[Z = 1|X\right]}$$

Solution. Note that

$$\mathbb{E}\left[\kappa G\right] = \mathbb{E}\left[G\right] - \mathbb{E}\left[\frac{D\left(1 - Z\right)}{P\left(Z = 0|X\right)}G\right] - \mathbb{E}\left[\frac{Z\left(1 - D\right)}{P\left[Z = 1|X\right]}G\right]$$

▷ The second term can be written as:

$$\begin{aligned}
\mathbb{E}\left[\frac{GD\left(1 - Z\right)}{P\left(Z = 0|X\right)}\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{GD\left(1 - Z\right)}{P\left(Z = 0|X\right)}|X\right]\right] \qquad (\cdot.\cdot) \text{ LIE} \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{GD\left(1 - Z\right)}{P\left(Z = 0|X\right)}|X, Z = 0\right]P\left(Z = 0|X\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[GD\left(1 - Z\right)|X, Z = 0\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[GD_0|X, Z = 0\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[G|X, Z = 0, D_0 = 1\right]P\left(D_0 = 1|X, Z = 0\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[G|X, Z = 0, D_0 = 1\right]P\left(D_0 = 1|X\right)\right] \qquad (\cdot.\cdot) \text{ exogeneity} \\
&= \mathbb{E}\left[\mathbb{E}\left[G|X, Z = 0, D_1 = D_0 = 1\right]P\left(D_0 = 1, D_1 = 1|X\right)\right] \qquad (\cdot.\cdot) \text{ monotonicity} \\
&= \mathbb{E}\left[\mathbb{E}\left[G|X, D_1 = D_0 = 1\right]P\left(D_0 = D_1 = 1|X\right)\right] \qquad (\cdot.\cdot) \text{ exogeneity}
\end{aligned}$$

▷ The third term can be analogously simplified as:

$$\mathbb{E}\left[\frac{GZ\,(1-D)}{P\,[Z=1|X]}\right] = \mathbb{E}\left[\mathbb{E}\left[G|X, D_1 = D_0 = 0\right] P\left(D_0 = D_1 = 0|X\right)\right]$$

▷ Therefore, since $\mathbb{E}\left[G\right] = \mathbb{E}\left[\mathbb{E}\left[G|X\right]\right]$, applying the law of total probability to $\mathbb{E}\left[G|X\right]$ and substituting in the above simplified values into the 2nd and the 3rd term yields:

$$\begin{aligned}
\mathbb{E}\left[\kappa G\right] &= \mathbb{E}\left[\mathbb{E}\left[G|X, D_1 = 1, D_0 = 0\right] P\left(D_1 = 1, D_0 = 0|X\right) + \mathbb{E}\left[G|X, D_1 = 0, D_0 = 1\right] P\left(D_1 = 0, D_0 = 1|X\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[G|X, D_1 = 1, D_0 = 0\right] P\left(D_1 = 1, D_0 = 0|X\right)\right] \qquad (\because) \text{ monotonicity} \\
&= \int_X \mathbb{E}\left[G|X = x,\ \text{compliers}\right] \frac{P\left(X = x,\ \text{compliers}\right)}{P\left(X = x\right)} P\left(X = x\right) dx \\
&= \int_X \mathbb{E}\left[G|X = x,\ \text{compliers}\right] \frac{P\left(X = x|\ \text{compliers}\right) P\left(\text{compliers}\right)}{P\left(X = x\right)} P\left(X = x\right) dx \\
&= \int_X \mathbb{E}\left[G|X = x,\ \text{compliers}\right] P\left(X = x|\ \text{compliers}\right) P\left(\text{compliers}\right) dx
\end{aligned}$$

Taking $P\left(\text{compliers}\right)$ outside the integral:

$$\begin{aligned}
\mathbb{E}\left[\kappa G\right] &= P\left(\text{compliers}\right) \int_X \mathbb{E}\left[G|X = x,\ \text{compliers}\right] P\left(X = x|\ \text{compliers}\right) dx \\
&= P\left(\text{compliers}\right) \mathbb{E}\left[G|\ \text{compliers}\right]
\end{aligned}$$

and thus we have shown the desired relationship.

3 Dutch Medical Schools

Seats in Dutch medical schools are assigned through a lottery. Applicants to medical studies in the Netherlands are assigned to lottery categories based on their high school grades. The categories differ by the probability to be awarded a place (to win the lottery). If people lose a lottery, they can try again the following year.

The datset contains peoples' first lottery outcome for participants in 1988 and 1989, and whether they attended medical school, as well as earnings from a survey that was sent out in 2007. You plan to estimate the return to attending medical school (D) on earnings in 2007 $(\ln w)$ using the lottery outcome (Z) as your instrument.

Problem 3.1. Discuss instrument exogeneity, exclusion, and monotonicity.

Solution. We use the following notation:

- ▷ Y: log of earnings, $D = 1$ if attended medical school, $Z = 1$ if entered in the lottery
- ▷ For exogeneity, we need exclusion restriction and random assignment:
  - * **Exclusion** $(Y_{D=d,Z=1} = Y_{D=d,Z=0} = Y_{D=d})$
    - Conditional on $D = d$, Z should not be able to provide additional information about Y.
    - However, consider two students who both attend medical school are some point, but one won the lottery in his first application and the other did not. That means that the student who won the lottery is more likely to have had a higher GPA than the student who did not win.
    - In this case, potential earnings will be higher for the student who attended medical school and won the lottery in his first application than for the student who attended medical school and did not win the lottery the first time around. Thus, exclusion **likely will not hold**.
    - Conditional on GPA, however, we do have exclusion.
  - * **Random assignment:** $((Y_0, Y_1, D_0, D_1) \perp Z)$
    - GPA may be correlated with potential outcomes (Y_0, Y_1) and is definitely correlated with (Z). Thus, random assignment **does not hold**.
    - Conditional on GPA, however, we do have random assignment.
- ▷ Monotonicity:
  - * The defiers are people who would choose not to go to medical school if they won the lottery and choose to go to medical school if they didn't win the lottery. Monotonicity requires that we rule out this group of people. While this behavior seems unreasonable, it could be the case that the student entered lottery with intention of enrolling but couldn't due to extreme circumstances (e.g. illness) and later re-entered the lottery, this time all ready to accept.
  - * Since the above circumstances are very rare, monotonicity is **likely satisfied**.

Problem 3.2. Assess instrument relevance.

Solution. For an instrument to be relevant, we can run the first-stage regression of D on Z. We obtain the following results:

Source	SS	df	MS	Number of obs	=	1,476
				F(1, 1474)	=	712.96
Model	85.3087772	1	85.3087772	Prob > F	=	0.0000
Residual	176.37144	1,474	.119654979	R-squared	=	0.3260
				Adj R-squared	=	0.3255
Total	261.680217	1,475	.177410316	Root MSE	=	.34591

d	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
z	.5203044	.0194862	26.70	0.000	.4820809	.558528
_cons	.4100877	.0161988	25.32	0.000	.3783126	.4418629

- ▷ The coefficient attached to Z is statistically significant and has a reasonably R^2 of 32.6%. The F-statistic is 712.96 which is very large.
- ▷ We conclude that the instrument indeed appears to be relevant.

Problem 3.3. Estimate the return to attending medical school on earnings in 2007 using IV and interpret the results.

Solution. We obtain the following results when we don't include the covariates:

lnw	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
d	.1871175	.0504348	3.71	0.000	.0882671	.2859678
_cons	3.010613	.0406666	74.03	0.000	2.930908	3.090318

Instrumented: d
Instruments: z

- ▷ The coefficient $\beta_{2SLS} = 0.1871175$.
- ▷ On average, attending medical school leads to $e^{0.1871175} - 1 = 20.58\%$ increase in earnings for the complier population i.e. those who would attend medical school only if they won the lottery.
- ▷ As we argued earlier, the instrument is likely not valid without conditioning on the covariates, so the coefficient does not yield a reasonable interpretation.

Problem 3.4. Count the number of compliers, and compare them to the population of applicants in terms of gender.

Solution. To find the number of compliers, note that

$$\begin{aligned}
P(D = 0 \,\&\, Z = 0) &= P(\text{ never takers}) + P(\text{ compliers }) \\
P(D = 1 \,\&\, Z = 0) &= P(\text{ always takers}) \\
P(D = 0 \,\&\, Z = 1) &= P(\text{ never takers }) \\
P(D = 1 \,\&\, Z = 1) &= P(\text{ always takers }) + P(\text{ compliers })
\end{aligned}$$

and assuming Z is random, the distribution of types is the same for each value of Z and the population as a whole. We use the following tables and first find the propotion of compliers in the entire sample:

Population

Attended medical school	Admitted through the lottery 0	1
0	269	71
1	187	949

Males

Attended medical school	Admitted through the lottery 0	1
0	108	27
1	67	353

Females

Attended medical school	Admitted through the lottery 0	1
0	161	44
1	120	596

▷ The proportion of always takers:

$$p_a = P(D=1|Z=0) = \frac{P(D=1\&Z=0)}{P(Z=0)} = \frac{187}{187+269} = 0.4101$$

▷ The proportion of never takers:

$$p_n = P(D=0|Z=1) = \frac{P(D=0\&Z=1)}{P(Z=1)} = \frac{71}{71+949} = 0.0696$$

▷ The propotion of compliers: $1 - p_a - p_b = 0.5203$ which implies 768 compliers in the total population.

Repeating the process for males and females yields the complier proportions of

$$\text{males:} 1 - \frac{67}{67+108} - \frac{27}{27+353} = 1 - 0.3829 - 0.075 = 0.5421$$
$$\text{females:} 1 - \frac{120}{120+161} - \frac{44}{44+596} = 1 - 0.4270 - 0.06875 = 0.5043$$

The population proportion of compliers (0.5203) is greater than the proportion of compliers among females (0.5043) and but smaller than the proportion of compliers among males (0.5421).

Problem 3.5. Is the IV estimate an estimate of the ATT? Explain why or why not.

Solution. The IV estimate corresponds to LATE, the average treatment effect among compliers of the lottery.

▷ If the instrument is not valid (which is likely the case here without covariate conditioning), the wald estimator is not an estimate of LATE.

▷ If the instrument is valid, then the wald estimator (IV estimate) is an estimate of LATE. For it to be an estimate of the ATT, there should be no always takers in our sample. But we showed earlier that the proportion of always takers is around 41%, so it is not an estimate of the ATT.

Problem 3.6. Estimate the mean and distribution of Y_0 and Y_1 for compliers.

Solution. We can get the counterfactual distribution for the compliers through the following.

▷ Denote $f_{zd}(y)$ as the observed marginal distribution of Y conditional on D and Z :

$$f_{zd}(y) \equiv f(y|Z=z, D=d)$$

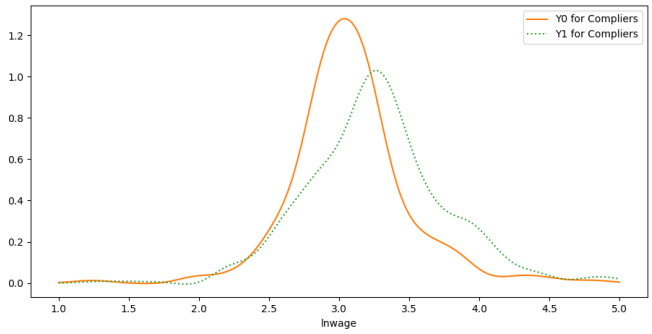
▷ This allows us to map $f_{zd}(y)$ to the marginal distribution of each type:

$$
\begin{aligned}
f_{10}(y) &= g_n(y) \\
f_{01}(y) &= g_a(y) \\
f_{00}(y) &= g_{c0}(y)\frac{p_c}{p_c+p_n} + g_n(y)\frac{p_n}{p_c+p_n} \\
f_{11}(y) &= g_{c1}(y)\frac{p_c}{p_c+p_a} + g_a(y)\frac{p_a}{p_c+p_a}
\end{aligned}
$$

and rearranging yields:

$$
g_{c0}(y) = f_{00}(y)\frac{p_c+p_n}{p_c} - f_{10}(y)\frac{p_n}{p_c} g_{c1}(y) = f_{11}(y)\frac{p_c+p_a}{p_c} - f_{01}(y)\frac{p_a}{p_c}
$$

▷ To obtain the $f_{00}, f_{01}, f_{10}, f_{11}$, we fit a kernel density estimate conditioned on each value of z and d. The resulting distribution looks as the following:



▷ To get the mean of Y_0 and Y_1 for compliers, we run a 2SLS regression of YD on Z and a 2SLS regression of $Y(1-D)$ on Z :

YD	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
d	3.264167	.0455034	71.73	0.000	3.174982	3.353352
_cons	-.0617161	.0366904	-1.68	0.093	-.133628	.0101958

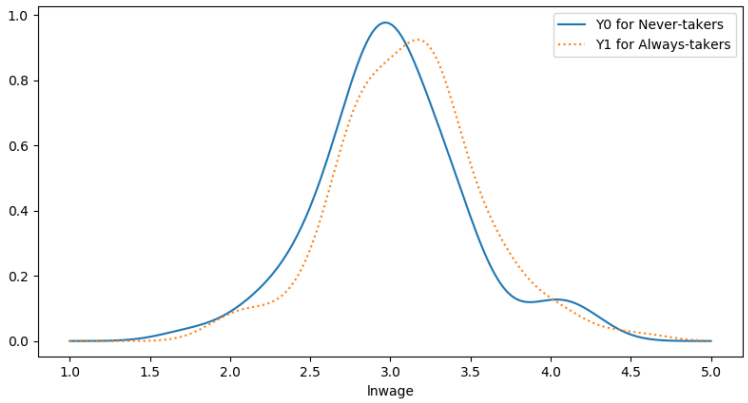
Y1_D	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
d1	3.077049	.0219074	140.46	0.000	3.034111	3.119987
_cons	-.0047203	.0072942	-0.65	0.518	-.0190167	.009576

* The regression results show that $\mathbb{E}[Y_1|\text{complier}] = 3.26417$ and $\mathbb{E}[Y_0|\text{complier}] = 3.0770$.

Problem 3.7. What can you say about Y_0 and Y_1 for always- and never-takers?

Solution. We already computed the distribution $f_{10}(y)$, which corresponds to the empirically distribution of never-takers and $f_{01}(y)$, which corresponds to the empirically distribution of always-takers.

▷ The plots of the distributions are shown below:



▷ The following means can be also computed from the sample:

```
. summarize lnw if z == 0 & d == 1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         lnw |        187    3.113672     .453354    1.88607   4.617243
```

```
. summarize lnw if z == 1 & d == 0        // Never Takers

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         lnw |         71    3.009236    .4592279   1.752539   4.194886
```

* $\mathbb{E}[Y_1|\text{ always takers}] = 3.113672$ and $\mathbb{E}[Y_0|\text{ never takers}] = 3.009236$
* Since we only observed treated always-takers and non-treated never takers, we can only estimate the above means.

Problem 3.8. The lottery is within lottery category and year, so your instrument is only exogenous within these groups. Estimate lottery category x year specific LATEs and combine these in one estimate. Compare this to the specification where you control for lottery category $\times$ year dummies and also interact the instrument with these dummies.

Solution. The specific LATEs can be obtained via:

$$\beta(x) = \frac{\mathbb{E}[Y|Z=1, X=x] - \mathbb{E}[Y|Z=0, X=x]}{\mathbb{E}[D|Z=1, X=x] - \mathbb{E}[D|Z=0, X=x]}$$

▷ First, we estimate category $\times$ year specific LATEs and combine them into one estimate.

* Using the population proportions as weights, we obtain the final estimate of 0.2121.
* Using the Angrist-Imbens weights which are defined to be

$$w(x) = \frac{\sigma^2_{\hat{D}}(x)}{\mathbb{E}\left[\sigma^2_{\hat{D}}(x)\right]} = \frac{\pi^2_x \sigma^2_Z(x)}{\mathbb{E}\left[\pi^2_x \sigma^2_Z(x)\right]}$$

  i.e. the numerator is the standard deviation of the fitted values when we regress D on Z conditional on X and the denominator is the population-weight-average of these weights for each corresponding X. We obtain the final estimate of 0.1915.
* The LATE estimates and the weights are shown in the table below:

	LATE(X)	w1	w2
3 x 1988	-0.803663	0.056911	0.002778
3 x 1989	1.564193	0.053523	0.005299
4 x 1988	0.119505	0.141599	0.052726
4 x 1989	0.011939	0.129404	0.026649
5 x 1988	0.159244	0.128726	0.197538
5 x 1989	0.518692	0.142276	0.058041
6 x 1988	0.209577	0.180217	0.339962
6 x 1989	0.145346	0.167344	0.317007

▷ Next, we run a 2SLS regression where we control for category $\times$ year dummies as well as interacting the instrument with these dummies in both stages.

* We obtain the final estimate of 0.1915 which corresponds to the Angrist-Imbens weights.
* The output is shown below:

IV-2SLS Estimation Summary

Dep. Variable:	lnw	R-squared:	0.0206
Estimator:	IV-2SLS	Adj. R-squared:	0.0152
No. Observations:	1476	F-statistic:	6.988e+04
Date:	Tue, Apr 23 2019	P-value (F-stat)	0.0000
Time:	05:44:38	Distribution:	chi2(9)
Cov. Estimator:	robust		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
1988 x 3	2.9484	0.0723	40.781	0.0000	2.8067	3.0901
1989 x 3	3.0202	0.0712	42.394	0.0000	2.8805	3.1598
1988 x 4	3.0783	0.0519	59.357	0.0000	2.9766	3.1799
1989 x 4	2.9459	0.0538	54.727	0.0000	2.8404	3.0514
1988 x 5	3.0629	0.0448	68.386	0.0000	2.9751	3.1506
1989 x 5	2.9795	0.0520	57.250	0.0000	2.8775	3.0815
1988 x 6	3.0331	0.0414	73.187	0.0000	2.9518	3.1143
1989 x 6	2.9633	0.0437	67.834	0.0000	2.8777	3.0489
d	0.1915	0.0501	3.8260	0.0001	0.0934	0.2896

4 Health Insurance

We are interested in how health insurance affects out-of-pocket expenditure on drugs, and have access to an extract from the Medical Expenditure Panel Survey of individuals over the age of 65 years. We want to estimate the following equation:

$$\text{ldrugexp} = \alpha + \gamma \text{hi\_empunion} + X\beta + U$$

where ldrugexp is log expenditure on prescribed medical drugs, hi_empunion is equal to one if the individual has supplemental health insurance and zero otherwise, and we control in X for age, gender, linc (log of household income), totchr (the no. of children) and blhisp (a dummy for being black or hispanic). Below is output from Stata with a number of results that may be useful in this exercise.

Problem 4.1. Explain why we may worry that having supplemental health insurance is endogenous in the equation above.

Solution. Our concern is that there may exist a variable that is correlated with both expenditure on prescribed medical drugs (Y) and whether the individual has health insurance (D).

▷ Person's health may be this endogenous variable. The worse the person's health is, (1) the more likely they incur expenditure on prescribed medical drugs (Y) and (2) the more likely they will have health insurance (D).

Problem 4.2. A suggested instrument is multlc, a dummy for whether the firm at which the individual is employed is a large operator with multiple locations. Why or why not may this be a good instrument (think about the conditions that need to hold to identify the LATE)? Using the output below, do you think that multlic is a weak instrument?

Solution. We consider each condition in detail:

▷ Exclusion: $Y_{D=d,Z=1}|X = Y_{D=d,Z=0}|X = Y_{D=d}|X$ – If people with higher potential drug spending self-select into the larger firms for reasons other than availability of health insurance, then exclusion is violated. This is likely as larger firms may offer greater salaries, and ths exclusion **is violated**.

▷ Random Assignment: $(Y_1, Y_0, D_0, D_1) \perp Z|X$ – Since health insurance, where you work, and medical costs are all a joint decision, random assignment will **likely not hold**.

▷ Relevance: $P[D = 1|X, Z = 1] \neq P[D = 1|X, Z = 0]$ – Individuals employed at large operator may be more likely to receive health insurance or be subsidized for health insurance. In the regression output, we see that regressing *hi_empunion* on *multlc* and controls yields significant eofficient with t-stat of 7.26 and F-stat of 120.25 (and similarly for each gender group). So relevance **holds**.

▷ Monotonicity: $P[D_1 \geq D_0|X] = 1$ – The defiers are the group of people who would only purchase health insurance if they worked at a smaller firm. This group may exist if insurance is easier to acquire or more appealing in smaller firms than larger firms, but this is unlikely in real life and thus monotonicity is **likely to hold**.

▷ Overlap: $P[Z = 1|X] \in (0, 1)$ – **This may be violated**. In the data, we see that very few people have $multlc = 1$ so after conditioning on covariates, it is possible that no individuals have $multlc = 1$ after conditioning on the covariates.

To assess whether the instrument is weak, we need to consider the first-stage regression of hi_empunion on multlc and other covariates:

hi_empunion	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
multlc	.1487593	.020504	7.26	0.000	.1085674	.1889513
totchr	.0109104	.0036859	2.96	0.003	.0036853	.0181354
age	-.0091799	.0007101	-12.93	0.000	-.0105717	-.007788
female	-.0792221	.0096843	-8.18	0.000	-.0982052	-.060239
blhisp	-.0741602	.0123788	-5.99	0.000	-.0984251	-.0498953
linc	.0720981	.0062189	11.59	0.000	.0599079	.0842883
_cons	.90169	.0589985	15.28	0.000	.7860412	1.017339

▷ We see that the coefficient is significant with a t-stat of 7.26 and F-stat of 120.25 which mitigates our concern about the instrument being potentially weak.

▷ Similar conclusions can be drawn from the gender-specific regressions.

Problem 4.3. Derive the indirect least squares representation of the IV-estimator using multlc as an instrument, and calculate it using the output below. Interpret the estimate.

Solution. To get the indirect least squares representation of the IV-estimator, write the regressions in a simultaeous equation format:

$$\begin{aligned} Y &= \alpha + \gamma D + X'\beta + u \\ D &= \alpha^* + \gamma^* Z + X'\beta^* + u^* \end{aligned}$$

▷ Combining these two equations, we have

$$\begin{aligned} Y &= (\alpha + \gamma\alpha^*) + \gamma\gamma^* Z + X'(\beta + \gamma\beta^*) + (u + \gamma u^*) \\ &= \tilde{\alpha} + \tilde{\gamma} Z + X'\tilde{\beta} + \tilde{u} \end{aligned}$$

▷ The regession of D on Z yields the following estimate:

hi_empunion	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
multlc	.1487593	.020504	7.26	0.000	.1085674	.1889513
totchr	.0109104	.0036859	2.96	0.003	.0036853	.0181354
age	-.0091799	.0007101	-12.93	0.000	-.0105717	-.007788
female	-.0792221	.0096843	-8.18	0.000	-.0982052	-.060239
blhisp	-.0741602	.0123788	-5.99	0.000	-.0984251	-.0498953
linc	.0720981	.0062189	11.59	0.000	.0599079	.0842883
_cons	.90169	.0589985	15.28	0.000	.7860412	1.017339

▷ The regression of Y on Z yields the following estimate:

ldrugexp	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
multlc	-.2002194	.0540601	-3.70	0.000	-.3061878	-.0942509
totchr	.4401428	.0093589	47.03	0.000	.4217975	.4584882
age	-.0053332	.0019369	-2.75	0.006	-.0091299	-.0015366
female	.0501264	.0252882	1.98	0.047	.0005566	.0996962
blhisp	-.1481236	.0341141	-4.34	0.000	-.2149941	-.081253
linc	.0252773	.0137866	1.83	0.067	-.0017472	.0523018
_cons	6.000931	.1559161	38.49	0.000	5.695304	6.306557

▷ We are interested in γ which can be obtained from

$$\gamma = \frac{\tilde{\gamma}}{\gamma^*} = \frac{-0.2002194}{0.1487593} = -1.34593$$

▷ This estimate implies that having health insurance reduces drug expenditure by $e^{-1.34593} - 1 = -73.96\%$ among compliers i.e. those who would get health insurance only if they work at a larger operator with multiple locations.

Problem 4.4. Assuming $\beta = 0$, derive the IV-estimator using the moments (covariances), and calculate it using the output below.

Solution. If we assume $\beta = 0$, then we have

$$\begin{aligned} Y &= \alpha + \gamma D + u \\ D &= \alpha^* + \gamma^* Z + u^* \end{aligned}$$

Combining these equations yields:

$$\begin{aligned} Y &= (\alpha + \gamma\alpha^*) + (\gamma\gamma^*) Z + (u + \gamma u^*) \\ &= \tilde{\alpha} + \tilde{\gamma} Z + \tilde{u} \end{aligned}$$

▷ The OLS estimate of Y on Z:

$$\tilde{\gamma} = \frac{\text{Cov}[Y, Z]}{\text{Var}[Z]} = \frac{-0.016529}{0.58204}$$

▷ The OLS estimate of D on Z:

$$\gamma^* = \frac{\text{Cov}[D, Z]}{\text{Var}[Z]} = \frac{0.014051}{0.58204}$$

▷ Therefore:

$$\gamma = \frac{\tilde{\gamma}}{\gamma^*} = \frac{-0.016529}{0.014051} = -1.1764$$

▷ This estimate implies that having health insurance reduces drug expenditure by $e^{-1.1764} - 1 = -69.15\%$ among compliers i.e. those who would get health insurance only if they work at a larger operator with multiple locations.

Problem 4.5. What is the share of females in the complier group? How does this compare to the overall population? How does this affect your interpretation of the estimates?

Solution. To get the share of the females in the complier group, we need to compute the number of compliers in the female group and the male group.

▷ First, consider the following tabulation for females:

```
. table hi_empunion multlc if female == 1

------------------------
Insured   |   Multiple
thro      |  locations
emp/union |     0      1
----------+-------------
        0 | 3,721    125
        1 | 1,792    184
------------------------
```

* We know that the number of always takers must be

$$\begin{aligned}&(3,721+125+1,792+184)\times\frac{P(D=1\&Z=0)}{P(Z=0)}\\=&5,822\times\left(\frac{1,792}{5,513}\right)=1,892\end{aligned}$$

and the number of never takers must be

$$\begin{aligned}&(3,721+125+1,792+184)\times\frac{P(D=0\&Z=1)}{P(Z=1)}\\=&5,822\times\left(\frac{125}{309}\right)=2,355\end{aligned}$$

which implies that the number of compliers is

$$5,822-1,892-2,355=1,575$$

▷ Second, consider the following tabulation for males:

```
. table hi_empunion multlc if female == 0

------------------------
Insured   |   Multiple
thro      |  locations
emp/union |      0      1
----------+-------------
        0 | 2,267    120
        1 | 1,683    197
------------------------
```

▷ We know that the number of always takers must be

$$\begin{aligned}&(2,267+120+1,683+197)\times\frac{P(D=1\&Z=0)}{P(Z=0)}\\=&4,267\times\left(\frac{1,683}{3,950}\right)=1,818\end{aligned}$$

and the number of never takers must be

$$\begin{aligned}&(2,267+120+1,683+197)\times\frac{P(D=0\&Z=1)}{P(Z=1)}\\=&4,267\times\left(\frac{120}{317}\right)=1,615\end{aligned}$$

which implies that the number of compliers is

$$4,267-1,818-1,615=834$$

▷ Thus, the total number of compliers is $1,575+834=2,409$ and the share of females is 65.38%.

▷ The total population is $5,822+4,267=10,089$ and the share of females is 57.71%.

▷ Thus, the female are 13.64% over-represented in the compliers relative to the total population.

Problem 4.6. Assuming $\beta = 0$, what is the share of females in the three groups of compliers, always-takers, and never-takers?

Solution. We proceed as the following:

▷ We showed earlier that for females, the share of always-takers and never-takers are:

$$p_a = \left(\frac{1,792}{5,513}\right) = 0.3250, \quad p_n = \left(\frac{125}{309}\right) = 0.4045$$

and equivalently for males:

$$p_a = \left(\frac{1,683}{3,950}\right) = 0.4261, \quad p_n = \left(\frac{120}{317}\right) = 0.3785$$

▷ In the always-takers group:

* # of Males: $0.4261 \times ((1 - 0.5771) \times 10,089)$ =1,818
* # of Females: $0.3250 \times (0.5771 \times 10,089)$ =1,892
* Therefore, the share of females is .5100.

▷ In the never-takers group:

* # of Males: $0.3785 \times ((1 - 0.5771) \times 10,089) = 1,615$
* # of Females: $0.4045 \times (0.5771 \times 10,089) = 2,355$
* Therefore, the share of females is .5932

▷ In the compliers group, we already showed that the share of females is 0.5771.

Problem 4.7. Using the means and counts of ldrugexp from the output below, estimate $\mathbb{E}[Y_0|$ never taker$]$, $\mathbb{E}[Y_1|$ always taker $]$, $\mathbb{E}[Y_0|$ complier $]$ and $\mathbb{E}[Y_1|$ complier $]$ where Ys are the potential outcomes for $ldrugexp$ with and without supplemental health insurance. How do the complieres compare to the other groups, and what do you conclude about external validity?

Solution. We proceed as the following.

▷ The means for always-takers and never-takers can be found from this tabulation:

```
-----------------------------
Insured   |
thro      | Multiple locations
emp/union |        0          1
----------+------------------
        0 | 6.464303  6.029153
          |    5,988       245
          |
        1 | 6.558737    6.3345
          |    3,475       381
-----------------------------
```

* Never-takers: $\mathbb{E}[Y_0|$ never taker $] = \mathbb{E}[Y_0|D = 0, Z = 1] = 6.029153$
* Always-takers: $\mathbb{E}[Y_1|$ always taker $] = \mathbb{E}[Y_0|D = 1, Z = 0] = 6.558737$

▷ To find the means for compliers, compute the counterfactual means through this formula:

$$\mathbb{E}\left[Y_1|\text{ compliers }\right] = \frac{\mathbb{E}\left[YD|Z=1\right] - \mathbb{E}\left[YD|Z=0\right]}{\mathbb{E}\left[D|Z=1\right] - \mathbb{E}\left[D|Z=0\right]}$$

* Let's first show why this is true.
  - Consider the terms in the numerator:

$$\begin{aligned}\mathbb{E}\left[YD|Z=1\right] &= \mathbb{E}\left[Y_1D_1 + Y_0\left(1-D_1\right)D_1|Z=1\right] \\ &= \mathbb{E}\left[Y_1D_1 + Y_0\left(1-D_1\right)D_1\right] \\ \mathbb{E}\left[YD|Z=0\right] &= \mathbb{E}\left[Y_1D_0 + Y_0\left(1-D_0\right)D_0|Z=0\right] \\ &= \mathbb{E}\left[Y_1D_0 + Y_0\left(1-D_0\right)D_0\right]\end{aligned}$$

    which yields:

$$\mathbb{E}\left[YD|Z=1\right] - \mathbb{E}\left[YD|Z=0\right] = \mathbb{E}\left[Y_1|D_1 > D_0\right]P\left(D_1 > D_0\right)$$

  - The denominator:

$$\mathbb{E}\left[D|Z=1\right] - \mathbb{E}\left[D|Z=0\right] = \mathbb{E}\left[D_1 - D_0\right] = P\left(D_1 > D_0\right)$$

  - Combining the two terms, we thus have:

$$\mathbb{E}\left[Y_1|\text{ compliers }\right] = \mathbb{E}\left[Y_1|D_1 > D_0\right]$$

* Since we don't have a regression of YD on Z, rewrite

$$\mathbb{E}\left[Y_1|\text{ compliers }\right] = \frac{\mathbb{E}\left[Y_1|D=1,Z=1\right]P\left(D=1|Z=1\right) - \mathbb{E}\left[Y_1|D=1,Z=0\right]P\left(D=0|Z=0\right)}{\mathbb{E}\left[D|Z=1\right] - \mathbb{E}\left[D|Z=0\right]}$$

  Plugging in the values:

$$\mathbb{E}\left[Y_1|\text{ compliers }\right] = \frac{6.3345\frac{381}{381+245} - 6.558737\frac{3475}{3475+5988}}{\frac{381}{381+245} - \frac{3475}{3475+5988}} = 5.9934$$

▷ Repeating the process similarly:

$$\begin{aligned}\mathbb{E}\left[Y_1|\text{ compliers }\right] &= \frac{\mathbb{E}\left[Y_0|D=0,Z=1\right]P\left(D=0|Z=1\right) - \mathbb{E}\left[Y_0|D=0,Z=0\right]P\left(D=0|Z=0\right)}{\mathbb{E}\left[\left(1-D\right)|Z=1\right] - \mathbb{E}\left[\left(1-D\right)|Z=0\right]} \\ &= \frac{6.29153\frac{381}{381+245} - 6.464303\frac{5988}{3475+5988}}{\frac{245}{381+245} - \frac{5988}{3475+5988}} = 6.1830\end{aligned}$$

Since the means for always-takers and never-takers are quite different from the means obtained for compliers, the LATE results (which pertain to the compliers) are difficult to generalize to the whole population, so external validity may be an issue here.