Empirical Analysis IIIA - Problem Set 2

Manav Chaudhary, Sid Sah and George Vojta

Wednesday 1st July, 2020

Question 1

Part a

$$\begin{aligned} ATT =& \mathbb{E}[Y_1 - Y_0 | D = 1] \\ =& \underbrace{\mathbb{E}[Y_1 - Y_0 | D_1 = D_0 = 1]}_{\text{ATE of always takers}} P(D_1 = D_0 = 1) + \underbrace{\mathbb{E}[Y_1 - Y_0 | D_1 > D_0]}_{\text{ATE of compliers}} P(D_1 > D_0) \end{aligned}$$

Part b

When the population only consists of compliers i.e. $P(D_1 > D_0) = 1$, or if treatment effects are constant across agents.

Question 2

Part a

We can use a slightly modified version of the Wald estimator to estimate $LATE(x)$,

$$\begin{aligned} \frac{\mathbb{E}[Y|Z=1,X] - \mathbb{E}[Y|Z=0,X]}{\mathbb{E}[D|Z=1,X] - \mathbb{E}[D|Z=0,X]} =& \frac{\mathbb{E}[Y_0 + D_1(Y_1 - Y_0)|Z=1,X] - \mathbb{E}[Y_0 + D_0(Y_1 - Y_0)|Z=0,X]}{\mathbb{E}[D_1|Z=1,X] - \mathbb{E}[D_0|Z=0,X]} && \text{(Overlap)} \\ =& \frac{\mathbb{E}[(D_1 - D_0)(Y_1 - Y_0)|X]}{\mathbb{E}[D_1 - D_0|X]} && \text{(Exogeneity)} \\ =& \frac{\mathbb{E}[Y_1 - Y_0|T=c,X]P(T=c|X)}{P(T=c|X)} && \text{(Monotonicity and Relevance)} \\ =& \mathbb{E}[Y_1 - Y_0|T=c,X] \equiv LATE(x) \end{aligned}$$

Part b

$$\begin{aligned} \mathbb{E}[Y_1 - Y_0|T=c] =& \mathbb{E}[\mathbb{E}[Y_1 - Y_0|T=c,X=x]|T=c] && \text{(LIE)} \\ =& \mathbb{E}[LATE(x)|T=c] \\ =& \sum_{x \in X} LATE(x) P(X=x|T=c) \\ =& \sum_{x \in X} LATE(x) \frac{P(X=x,T=c)}{P(T=c)} \\ =& \sum_{x \in X} LATE(x) \frac{P(T=c|X=x)}{P(T=c)} P(X=x) \\ =& \mathbb{E}\left[\frac{LATE(x)P(T=c|x)}{P[T=c]}\right] \end{aligned}$$

Part c

Denote,[1]

$$\begin{aligned} Y =& \alpha_X + \beta_{TSLS} D + \varepsilon \\ D =& \pi_X + \pi_{1X} Z_i + \eta \end{aligned}$$

where π_X and α_X denote saturated models for covariates (full set of dummies for all values of X_i) and π_{1X} denotes a first-stage effect of Z_i for every value of X_i. Specifically,

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

Let,

$$\tilde{D} = \underbrace{\mathbb{E}[D|Z,X]}_{\equiv \hat{D}} - \mathbb{E}[\mathbb{E}[D|Z,X]|X] = \mathbb{E}[D|Z,X] - \mathbb{E}[D|X]$$

Then,

$$\begin{aligned} \beta_{TSLS} =& \frac{Cov(Y,\tilde{D})}{Var(\tilde{D})} \\ =& \frac{\mathbb{E}[Y\tilde{D}]}{\mathbb{E}[\tilde{D}^2]} \\ =& \frac{\mathbb{E}[\mathbb{E}[Y|Z,X]\tilde{D}]}{\mathbb{E}[\tilde{D}^2]} && \text{(LIE)} \end{aligned}$$

Notice we can simplify the $\mathbb{E}[Y|Z,X]$,

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

Substituting this into the numerator we get,

$$\begin{aligned} \beta_{TSLS} =& \frac{\mathbb{E}[\mathbb{E}[Y|Z=0,X]\tilde{D} + LATE(X)\pi_{1X}Z\tilde{D}]}{\mathbb{E}[\tilde{D}^2]} \\ =& \frac{\mathbb{E}[LATE(X)\pi_{1X}Z\tilde{D}]}{\mathbb{E}[\tilde{D}^2]} && \text{(Orthogonality of BLP)} \\ =& \frac{\mathbb{E}[LATE(X)(\pi_X + \pi_{1X}Z)\tilde{D}]}{\mathbb{E}[\tilde{D}^2]} && \text{(Orthogonality of BLP)} \\ =& \frac{\mathbb{E}[LATE(X)\hat{D}(\hat{D} - \mathbb{E}[\hat{D}|X])]}{\mathbb{E}[\tilde{D}^2]} && (\hat{D} = \mathbb{E}[D|Z,X] \equiv \pi_X + \pi_{1X}Z) \\ =& \frac{\mathbb{E}[LATE(X)\mathbb{E}[\hat{D}(\hat{D} - \mathbb{E}[\hat{D}|X])|X]]}{\mathbb{E}[\tilde{D}^2]} && \text{(LIE)} \\ =& \frac{\mathbb{E}[LATE(X)Var[\hat{D}|X]]}{\mathbb{E}[Var[\hat{D}|X]]} && \text{(LIE)} \\ =& \frac{\mathbb{E}[LATE(X)Var[p(X,Z)|X]]}{\mathbb{E}[Var[p(X,Z)|X]]} && \text{(Since } D \in \{0,1\}) \end{aligned}$$

[1] WLOG assume X contains 1

Part d

Using the definition of κ we can re-write $\mathbb{E}[\kappa G]$ as,

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

We can simplify the second term,

$$\begin{aligned}
\mathbb{E}\left[\frac{D(1-Z)}{P(Z=0|X)}G\right] =& \mathbb{E}\left[\mathbb{E}\left[\frac{GD(1-Z)}{P(Z=0|X)}|X\right]\right] \\
=& \mathbb{E}\left[\mathbb{E}\left[\frac{GD(1-Z)}{P(Z=0|X)}|X, Z=0\right] P(Z=0|X)\right] \\
=& \mathbb{E}\left[\mathbb{E}\left[GD_0|X, Z=0\right]\right] \\
=& \mathbb{E}\left[\mathbb{E}\left[G|X, Z=0, D_0=1\right] P(D_0=1|X, Z=0)\right] \\
=& \mathbb{E}\left[\mathbb{E}\left[G|X, D_0=D_1=1\right] P(D_0=D_1=1|X)\right] \qquad \text{(Monotonicity and Exogeneity)}
\end{aligned}$$

Similarly we can simplify the third term to,

$$\mathbb{E}\left[\mathbb{E}\left[G|X, D_0=D_1=0\right] P(D_0=D_1=0|X)\right]$$

We can re-write the first term as,

$$\begin{aligned}
\mathbb{E}[G] =& \mathbb{E}[\mathbb{E}[G|X]] && \text{(LIE)} \\
=& \mathbb{E}[\mathbb{E}[G|X, T=c]P(T=c|X) + \mathbb{E}[G|X, T=nt]P(T=nt|X) \\
&+ \mathbb{E}[G|X, T=at]P(T=at|X)] && \text{(monotonicity)}
\end{aligned}$$

Combining all three terms we get,

$$\begin{aligned}
\mathbb{E}[\kappa G] &= \mathbb{E}\left[\mathbb{E}[G|X, T=c]P(T=c|X)\right] \\
&= \sum_{x \in X} \mathbb{E}[G|X, T=c] \underbrace{P(T=c|X=x)P(X=x)}_{=P(X=x|T=c)P(T=c)} \\
&= P(T=c)\mathbb{E}[G|T=c]
\end{aligned}$$

Therefore,

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

Question 3

Part A: Instrument Exogeneity

Exclusion: There appears to be a clear violation of instrument exogeneity in this experiment. Exogeneity says that $Y_{d,0} = Y_{d,1}$: Clearly Z can affect Y through a different channel than actually attending medical school, for instance those with higher ability are more likely to win the lottery and have better outcomes. As a result we need some sort of other variable to instrument with or at least condition on so that we can make our potential outcomes of Y (wages) and Z independent. Our instrument exogeneity fails the exclusion. restriction.

Random Assignment: Random assignment says that $Y_{d,z}, D_Z \perp Z \forall d, z$. However again as mentioned above the potential outcomes Y_0, Y_1 are likely correlated with Z and thus the random assignment condition fails.

In a brutal twist of fate it seems that we fail instrument exogeneity in all possible ways. Rough start for the team here.

Monotonicity: Monotonicity is a bit more difficult to analyze. It seems very likely to hold because those who enter the lottery will be more likely to attend medical school should they actually win the lottery. Hypothetically there could be people who want to go more if they lost the lottery because that's the way they are wired. While saying monotonicity is definitely valid isn't bullet proof in this case it's probably a fair assumption to make that $D_1 > D_0$.

Part B: Are we at least relevant?

Running a regression of the treatment on the instrument has an incredibly statistically significant result which indicate that the our instrument is relevant for treatment:

	(1)
	Attended Med
Won Lottery	0.520***
	(0.0195)
Constant	0.410***
	(0.0162)
Observations	1476
R^2	0.326

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 1

Part C

Below we present a very basic instrumental variable regression of wages on attending medical school instrumented by whether one won the lottery: The table shows us that people who went to medical school log wages go up by .187. Note however there is not a great interpretation for these results in reality because our instrument is garbage so we aren't sure the underlying economic processes people are attending or now we are just saying something about our data.

VARIABLES	(1) Wages (Logged)
Attended Med	0.187***
	(0.0504)
Constant	3.011***
	(0.0407)
Observations	1,476
R^2	0.010

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 2

Part D: Breakdown By Gender

Below I present a table of instruments and treatments by gender and for the total population. Treated and Untreated at $D = 1, 0$ respectively with Won Lottery and Lost Lottery being $Z = 1, 0$ respectively: With

Table 3: Summary Stats By Gender

Gender	Treated_WonLottery	Treated_LostLottery	Untreated_WonLottery	Untreated_LostLottery	Total
Male	353	67	27	108	555
Female	596	120	44	161	921
Total	949	187	71	269	1476

Counts

Table 4

these we can then back out our compliers, always takers and never takers:

$$Prob(D = 0, Z = 0) = Prob(\text{Compliers (Lost Lot)}) + Prob(\text{Never Takers (Lost Lot)})$$
$$Prob(D = 1, Z = 0) = Prob(\text{Always Takers (Lost Lot)}$$
$$Prob(D = 1, Z = 1) = Prob(\text{Compliers (Won Lot)}) + Prob(\text{Always Takers (Won Lot)}$$
$$Prob(D = 0, Z = 1) = Prob(\text{Never Takers (Won Lot)}$$

Now we will impose the assumption that Z is random even though we had large problems with such an assumption earlier to back out the requisite populations. Here we know that all of these populations are a fixed type of the population whether or not they got treatment due to the randomness of Z. As such we can say the following about never takers:

$$Prob(\text{Never Takers (Won Lot)} = \frac{Prob(D = 0, Z = 1)}{Prob(Z = 1)}$$
$$= \frac{71}{71 + 949} = Prob(\text{Never Takers (Lost Lot)} = .07$$

Now with always takers:

$$Prob(\text{Always Takers (Lost Lot)} = \frac{Prob(D = 1, Z = 0)}{Prob(Z = 0)}$$
$$= \frac{187}{187 + 269} = Prob(\text{Always Takers (Won Lot)} = .41$$

And since we assumed no deniers we just calculate:

$$Prob(\text{Compliers}) = 1 - .41 - .07 = .52$$

Now, lets do the same thing with females:

$$Prob(\text{Never Takers (Won Lot)} = \frac{44}{44 + 596} = .07$$
$$Prob(\text{Always Takers (Lost Lot)} = \frac{120}{120 + 161} = .43$$
$$Prob(\text{Compliers}) = 1 - .43 - .07 = .50$$

Finally lets do the same thing with Males:

$$Prob(\text{Never Takers (Won Lot)} = \frac{27}{27 + 353} = .07$$
$$Prob(\text{Always Takers (Lost Lot)} = \frac{67}{67 + 108} = .38$$
$$Prob(\text{Compliers}) = 1 - .38 - .07 = .55$$

We confirm these estimates with the first stages below (more than one way to skin a cat I guess):

	(1)	(2)	(3)
	d	d	d
z	0.504***	0.546***	0.520***
	(0.0247)	(0.0317)	(0.0195)
_cons	0.427***	0.383***	0.410***
	(0.0206)	(0.0262)	(0.0162)
N	921	555	1476
R^2	0.311	0.350	0.326
gender	Female	Male	Full Population

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 5: Using Wald First Stage To Get Prop of Each Gender is Compier

We can use these numbers in conjunction with the population ratio of men and women (women = 921/1476 = .624 $\Rightarrow$ men = .376) to get to the proportion of compliers who are men and women. Following slide 14:

$$P(X = x|D_1 > D_0) = P(X = x)\frac{E[D|Z = 1, X = x] - E[D|Z = 0, X = x]}{E[D|Z = 1] - E[D|Z = 0}$$
$$\text{Inputting our numbers for women: } = (.624)\frac{.50}{.52} = .6$$

So women make up 60% of the compliers meaning men make up the other 40%.

Part E

Is the IV estimate an estimate of the ATT? Well the IV regressor gives us a LATE under the right circumstances, but since instrument exogeneity clearly fails here there's no way we get the right thing. But let's suppose for a second that we had all the right conditions. If so LATE only equals the ATT if there are no always takers, which they clearly are: The proof follows as such:

Part F

Here we follow the procedure from the slides laid out on slides 20-30: First lets show some notation:

$$\begin{aligned} f_{zd}(y) &= f(y|Z=z, D=d) \\ f_{01}(y) &= g_{\text{never}}(y) \\ f_{10}(y) &= g_{\text{always}}(y) \\ f_{00}(y) &= g_{\text{never}}(y)Prob(\text{Never Takers}) + g_{\text{Compliers, Z=0}}(y)Prob(\text{Compliers}) \\ f_{11}(y) &= g_{\text{always}}(y)Prob(\text{Always Takers}) + g_{\text{Compliers, Z=1}}(y)Prob(\text{Compliers}) \end{aligned}$$

Where again we are invoking randomization of Z so that we can back out the actual compliers, never takers and always takers. we can also input our results from part D:

$$\begin{aligned} f_{01}(y) &= g_{\text{never}}(y) \\ f_{10}(y) &= g_{\text{always}}(y) \\ f_{00}(y) &= g_{\text{never}}(y)\frac{.07}{.07+.52} + g_{\text{Compliers,Z=0}}(y)\frac{.52}{.07+.52} \\ f_{11}(y) &= g_{\text{always}}(y)\frac{.41}{.41+.52} + g_{\text{Compliers,Z=1}}(y)\frac{.52}{.41+.52} \end{aligned}$$

Which rearranges to:

$$\begin{aligned} g_{\text{Compliers,Z=0}}(y) &= f_{00}(y)\frac{.52+.07}{.52} - f_{10}(y)\frac{.07}{.52} \\ g_{\text{Compliers,Z=1}}(y) &= f_{11}(y)\frac{.52+.41}{.52} - f_{01}(y)\frac{.41}{.52} \end{aligned}$$

Below I present distributions for various combinations of treatments and instruments:

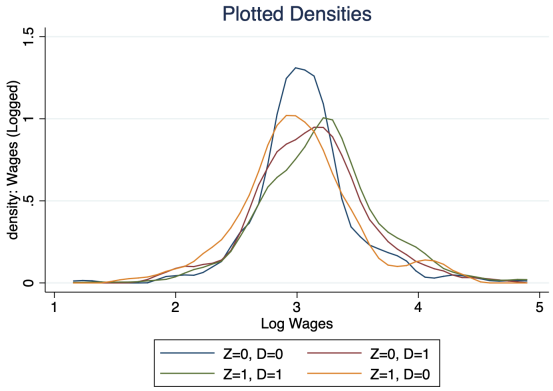

Figure 1

We then use these density functions, along with our results from Part F to find the distributions for Y_0 and Y_1 for the compliers:

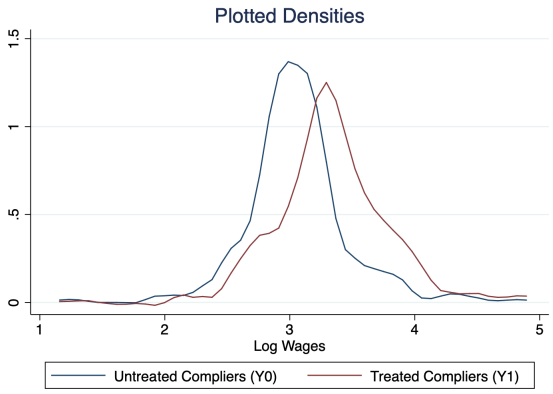

Figure 2

Then, using the code from topic 3 slide 30 we will calculate the mean values for Y_0 and Y_1 for the compliers: Which gives us values $E[Y_1|C = c] = 3.264$ and $E[Y_0|C = c] = 3.077$.

	(1)
VARIABLES	y1
Attended Med	3.264***
	(0.0388)
Constant	-0.0617**
	(0.0275)
Observations	1,476
R^2	0.910

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 6: Mean of Y_1 for the compliers

	(1)
VARIABLES	y0
md	3.077***
	(0.0293)
Constant	-0.00472
	(0.00475)
Observations	1,476
R^2	0.976

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 7: Mean of Y_0 for the compliers

Part G

We now present the distributions for the Always and Never Takers where again we assume that Z is assigned randomly:

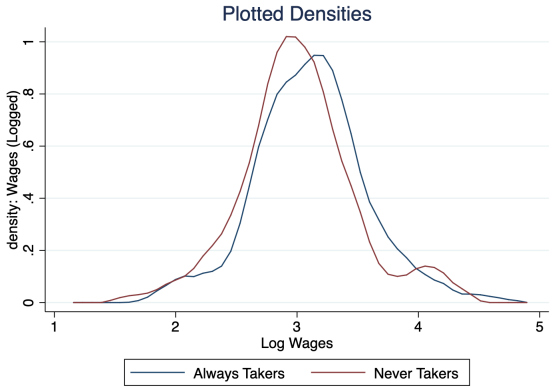

Figure 3: Distributions for Always and Never Takers (Y_1 and Y_0 Respectively)

And then we present a table of means below: Here $E[Y_1|C = a] = 3.11$ and $E[Y_0|C = n] = 3.01$. Note that

Table 8: Summary Stats By Takers

Type	mean	min	max	count
Always Taker	3.113672	1.88607	4.617243	187
Never Taker	3.009236	1.752539	4.194886	71

Table 9: Summary Statistics for the Always and Never Takers

because we can not create counter factuals for these two populations means are the best we can do.

Part H

Below I present the LATE estimates: The first 8 lines Estimate Category and Year Specific LATEs, 2SLS Line does 2SLS controlling for interactions in both stages, consistent with Angrist Impends Weights nad I conclude with Population weights. Great.

Table 10: LATE Calculations

Year	lotcateg	LATE	pi_x	variance_z	weight_final	observations
1988	3	-.8036624	.1688889	.0968158	.0488212	84
1988	4	.119505	.3393711	.1828763	.3723632	209
1988	5	.1592434	.6011594	.2401838	1.534555	190
1988	6	.209577	.6574712	.2468435	1.886405	266
1989	3	1.564191	.3054054	.0600454	.0990132	79
1989	4	.0119386	.2723763	.157013	.2059366	191
1989	5	.5186929	.3251226	.2182957	.4079428	210
1989	6	.1453456	.6689347	.2394589	1.89434	247
2SLS		.1915246				1476
AI Weights		.1915612				
Pop Weighted		.2128387				

First 8 lines Estimate Category and Year Specific LATEs, 2SLS Line does 2SLS controlling for interactions in both stages, consistent with Angrist Impends Weights

Table 11: Late Estimates

Question 4

Part A

We would worry about having supplemental health insurance as being endogenous if we believed that the potential drug expenditures (that is the drug expenditures for a given individual when that person does vs does not have supplemental health insurance) are not independent of having supplemental health insurance, even when conditioning on the included covariates. In this case, there are many reasons to suspect this. We might think that both having supplemental insurance and drug expenditures are associated with unobserved, underlying health state: less healthy individuals are more likely to seek out health insurance and more likely to spend more on drugs. We might also think that both are associated with the type of employment a given person has or had: certain types of jobs offer health insurance and the type of people who seek out/are qualified for these jobs might tend to have different health levels resulting in differing drug purchasing.

Part B

There are four required assumptions regarding an IV for the Wald estimand to estimate the LATE:

1) Random Assignment: $Y_{d,z}, D_z \perp Z|X$

This assumption appears unlikely to hold. Individuals have at least some ability to decide on their values of Z, the kind of firm they work at, and can presumably do so while considering expected values of $Y_{d,z}$ and D_z (the latter is presumably known rather than "expected" in most cases: people tend to know about the benefits of a job they're going to accept or reject). One story of violation could be that if large operators are more likely to offer health insurance, then people who value health insurance, and are thus more likely to end up with it either way, might self-select into such jobs.

2) Exclusion Restriction: $Y_{d,1} = Y_{d,0} = Y_d$

This restriction requires that the instrument only affects the outcome through the treatment. This assumption again seems unlikely to hold. Suppose that large operators tend to have more educated workers, or

workers that live in "better" neighborhoods, and such workers are healthier. Then, outside of any effect that goes through the treatment, outcomes would differ directly based on the instrument.

3) First Stage: $E[D_1 - D_0] \neq 0$ **& 4) Monotonicity:** $D_1 \geq D_0$

These two assumptions seem relatively likely to hold. We might expect that larger firms may have more resources and thus be more able to provide health insurance to employees. This would suggest that the treatment varies with the instrument, and, when it varies for a specific individual, it does so in a given direction (upwards). However, it would not be impossible to think of stories in which these assumptions are violated. In retail and food services, some larger chains tend to keep workers at certain numbers of hours in order to avoid having to provide benefits, such as health insurance. We might think that local businesses are less willing or able to do such things, in which case we would see instances of $D_1 < D_0$. We might also think that variation in offered health insurance is primarily explained by industry, rather than size of company, in which case the presence of a first-stage effect might be violated or close to violated. However, the data shows that this instrument is unlikely to have no first-stage effect or be a weak instrument. The attached Stata output shows that, in the sample, workers at large operators are nearly 15% more likely to have supplemental insurance, when conditioning on covariates.

Part C

Denoting LD as log drug expenditure, SHI as supplemental health insurance, and ML as working for a large operator, we can represent our IV setup as:

$$LD = \alpha_1 + \gamma_1 SHI + \beta_1' X + u_1$$

$$SHI = \alpha_2 + \gamma_2 ML + \beta_2' X + u_2$$

Then, subbing the latter into the former, we have:

$$LD = \alpha_1 + \gamma_1\alpha_2 + \gamma_1\gamma_2 ML + \gamma_1\beta_2' X + \gamma_1 u_2 + \beta_1' X + u_1$$

Performing an OLS regression of ML and X on SHI will produce an estimate of $\frac{Cov(\tilde{SHI},\tilde{ML})}{Var(\tilde{ML})}$ where $\tilde{SHI}$ and $\tilde{ML}$ are the residuals of those values after subtracting their BLP's based on X. We can then see that:

$$Cov(\tilde{SHI}, \tilde{ML}) = Cov(\gamma_2\tilde{ML} + \varepsilon_2 - BLP(\varepsilon_2|X), \tilde{ML}) = \gamma_2 Var(\tilde{ML})$$

where the second equality uses the independence of the potential outcomes of SHI and ML conditional on X. And so the OLS estimate of ML and X on SHI will produce an estimate of γ_2.
Similarly, an OLS regression of ML and X on LD will produce an estimate of $\frac{Cov(\tilde{LD},\tilde{ML})}{Var(\tilde{ML})}$. We now see that:

$$Cov(\tilde{LD}, \tilde{ML}) = Cov(\gamma_1\gamma_2\tilde{ML} + \gamma_1(\varepsilon_2 - BLP(\varepsilon_2|X)) + \varepsilon_1 - BLP(\varepsilon_1|X), \tilde{ML}) = \gamma_1\gamma_2 Var(\tilde{ML})$$

where the second equality again uses the randomization assumption as well as the exclusion restriction. Thus, we can produce the indirect least-squares estimator as and calculate it from the attached Stata output as:

$$\hat{\gamma}_{1,ILS} = \frac{\widehat{Cov(\tilde{LD}, \tilde{ML})}}{Var(\tilde{ML})} \Big/ \frac{\widehat{Cov(\tilde{SHI}, \tilde{ML})}}{Var(\tilde{ML})} = \frac{-0.2002194}{0.1487593} = -1.3459$$

This estimate suggests that, if all of our necessary assumptions hold, the mean member of the complier group, those who only get supplemental health insurance when working for a large operator, have a -1.3459 lower log expenditure on drugs when they have health insurance.

Part D

This assumption makes our framework equations a bit simpler:

$$SHI = \alpha_2 + \gamma_2 ML + u_2$$

$$LD = \alpha_1 + \gamma_1 SHI + u_1 = \alpha_1 + \gamma_1\alpha_2 + \gamma_1\gamma_2 ML + \gamma_1 u_2 + u_1$$

Then, taking some covariances, we get:

$$Cov(SHI, ML) = \gamma_2 Var(ML) + Cov(u_2, ML) = \gamma_2 Var(ML)$$

$$Cov(LD, ML) = \gamma_1\gamma_2 Var(ML) + Cov(\gamma_1 u_2 + u_1, ML) = \gamma_1\gamma_2 Var(ML)$$

where the second equalities uses the the, this time unconditional, assumptions of random assignment and the exclusion restriction. From the data we get:

$$\hat{\gamma}_{1,IV} = \frac{\widehat{Cov(LD, ML)}}{\widehat{Cov(SHI, ML)}} = \frac{-0.016529}{0.014051} = -1.1764$$

Part E

Generally speaking we can estimate the proportion of compliers:

$$P(C) = 1 - P(A) - P(N) = 1 - P(SHI = 1|ML = 0) - P(SHI = 0|ML = 1)$$

Estimating this from the subset of the data comprised of females, we get:

$$\widehat{P(C|Female)} = 1 - \frac{1,792}{1,792 + 3,721} - \frac{125}{125 + 184} = 0.27$$

Out of 5,822 total females, this suggests 1,572 are compliers. Repeating the exercise for males, we have:

$$\widehat{P(C|Male)} = 1 - \frac{1,683}{1,683 + 2,267} - \frac{120}{120 + 197} = 0.195$$

Out of 4,267 males, this suggests that 832 are compliers. Thus, the proportion of compliers who are female is 0.654 compared to a proportion of 0.577 for the total population. We see that females are more prevalent in the complier group than in the total population. The interpretation of this will vary with our assumptions about the distribution of treatment effects. For instance, if we assumed that treatment effect is constant by gender, this would suggest that our LATE estimates place more weight on the female treatment effect. If we assume that treatment effects are constant, on the other hand, this doesn't tell us anything at all.

Part F

In the previous section we established that the estimated proportion of females in the complier group is 0.654. We can now do similar things for the always-take and never taker groups.

Always-Takers:

$$P(A) = P(SHI = 1|ML = 0)$$

$$\widehat{P(A|Female)} = \frac{1,792}{1,792 + 3,721} = 0.325$$

$$\widehat{P(A|Male)} = \frac{1,683}{1,683 + 2,267} = 0.426$$

This would suggest that there are 1,892 female always-takers and 1,818 male always-takers, implying that the proportion of females among always-takers is 0.510.

Never-Takers:

$$P(N) = P(SHI = 0|ML = 1)$$
$$\widehat{P(N|Female)} = \frac{125}{125 + 184} = 0.405$$
$$\widehat{P(N|Male)} = \frac{120}{120 + 197} = 0.379$$

This would suggest that there are 2,358 female never-takers and 1,617 male never-takers, implying that the proportion of females among always-takers is 0.593.

Part G

Maintaining the assumption of monotonicity throughout this section, we can directly estimate:

$$E[Y_1|N] = E[Y * ML|SHI = 0, ML = 1] = E[Y|SHI = 0, ML = 1] \approx 6.029153$$

$$E[Y_0|A] = E[Y * (1 - ML)|SHI = 1, ML = 0] = E[Y|SHI = 1, ML = 0] \approx 6.558737$$

The first equalities in the statements above depend on exclusion restrictions: the always-takers/never-takers in those particular cells do not have different outcomes than if they had the opposite instrument value (their treatment value would be constant regardless). The estimates of potential outcomes require a bit more work:

$$\begin{aligned}
E[Y_1|SHI = 1] =& E[Y|SHI = 1] \\
E[Y_1|SHI = 1, ML = 1] =& E[Y_1|SHI = 1, ML = 1, C]P(C|SHI = 1, ML = 1) \\
&+ E[Y_1|SHI = 1, ML = 1, A]P(A|SHI = 1, ML = 1) && \text{(LIE)} \\
E[Y_1|SHI = 1, ML = 1, C] =& \frac{E[Y_1|SHI = 1, ML = 1] - E[Y_1|SHI = 1, ML = 1, A]P(A|SHI = 1, ML = 1)}{P(C|SHI = 1, ML = 1)} \\
E[Y_1|C] =& \frac{E[Y_1|SHI = 1, ML = 1] - E[Y_1|A]P(A|SHI = 1, ML = 1)}{P(C|SHI = 1, ML = 1)} && \text{(Excl. restr.)} \\
E[Y_1|C] =& \frac{E[Y_1|SHI = 1, ML = 1] - E[Y_1|SHI = 1, ML = 0]\frac{P(A)|ML=1|}{|ML=1,SHI=1|}}{\frac{P(C)|ML=1|}{|ML=1,SHI=1|}} \\
& && \text{(Excl. restr.)}
\end{aligned}$$

The above expression can all be estimated from the data:

$$\widehat{E[Y_1|C]} = \frac{6.3345 - 6.558737\frac{0.367*626}{381}}{\frac{0.242*626}{381}} = 5.9849$$

We can do a similar thing for the potential outcome without treatment:

$$\begin{aligned}
E[Y_0|SHI = 0] =& E[Y|SHI = 0] \\
E[Y_0|SHI = 0, ML = 0] =& E[Y_0|SHI = 0, ML = 0, C]P(C|SHI = 0, ML = 0) \\
&+ E[Y_0|SHI = 0, ML = 0, N]P(N|SHI = 0, ML = 0) && \text{(LIE)} \\
E[Y_0|SHI = 0, ML = 0, C] =& \frac{E[Y_0|SHI = 0, ML = 0] - E[Y_0|SHI = 0, ML = 0, N]P(N|SHI = 0, ML = 0)}{P(C|SHI = 0, ML = 0)} \\
E[Y_0|C] =& \frac{E[Y_0|SHI = 0, ML = 0] - E[Y_0|N]P(N|SHI = 0, ML = 0)}{P(C|SHI = 0, ML = 0)} && \text{(Excl. restr.)} \\
E[Y_0|C] =& \frac{E[Y_0|SHI = 0, ML = 0] - E[Y_0|SHI = 0, ML = 1]\frac{P(N)|ML=0|}{|ML=0,SHI=0|}}{\frac{P(C)|ML=0|}{|ML=0,SHI=0|}} \\
& && \text{(Excl. restr.)}
\end{aligned}$$

And again, all of this can be estimated from the data:

$$\widehat{E[Y_0|C]} = \frac{6.464303 - 6.029153\frac{0.391*9,463}{5,988}}{\frac{0.242*9,463}{5,988}} = 7.161516$$

We see that these estimates suggest that the complier group have quite different potential outcome drug spending than those of their always-taker and never-taker peers. This would suggest that there is relatively low external validity to the analysis. If the effects that we have estimated are for a subpopulation who are distinct from the rest of the population in their spending habits, then it it seems likely that treatment effects would differ as well.