HE3021 Week 9 Tutorial 7 Attempt

Lui Yu Sen U1930037F

3/22/2021

Packages used:
haven, for reading dta files
tidyverse, for dataframe manipulation
stats, for GLM fitting
aod, for coefficient tests
prediction and margins, for evaluating probabilities and marginal effects for probit and logit models
AER, for 2-step regression for instrumental variables

1

a

```
mroz_logit <- glm(formula = inlf ~ nwifeinc + educ + kidslt6 + age + exper,
                  data = mroz, family = binomial(link = "logit"))
mroz_probit <- glm(formula = inlf ~ nwifeinc + educ + kidslt6 + age + exper,
                  data = mroz, family = binomial(link = "probit"))
summary(mroz_logit)
```

```
##
## Call:
## glm(formula = inlf ~ nwifeinc + educ + kidslt6 + age + exper,
##     family = binomial(link = "logit"), data = mroz)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5175  -0.9174   0.4441   0.8841   2.2974
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.153219   0.742068   1.554   0.1202
## nwifeinc    -0.019900   0.008268  -2.407   0.0161 *
## educ         0.223366   0.042969   5.198 2.01e-07 ***
## kidslt6     -1.463577   0.200353  -7.305 2.77e-13 ***
## age         -0.095141   0.013439  -7.080 1.45e-12 ***
## exper        0.117887   0.013386   8.807  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1029.75  on 752  degrees of freedom
## Residual deviance:  812.92  on 747  degrees of freedom
## AIC: 824.92
##
## Number of Fisher Scoring iterations: 4
```

There is a negative sign for net wife income. The more income the wife has, the less likely she is to seek a job. Additional income not as necessary.
There is a positive sign for education. The more educated she is, the more likely she is to seek a job, since there are greater expected wages from greater expected returns to education.
There is a negative sign for kids less than 6. The more young kids she has, the more time she has to expend on her kids' care and the less she has for labour force participation.
There is a negative sign for age, the older the woman, the less likely she is to seek employment, since she has declining health and abilities.
There is a positive sign for work experience. In general, a more experienced worker commands greater wages, increasing the expected wages for a prospective job seeker, thus higher chance for a woman to participate in the labour force.

b

```
test_individual <- data.frame()
test_individual <- rbind(test_individual, c(20, 10, 0, 30, 10))
colnames(test_individual) <- c("nwifeinc", "educ", "kidslt6", "age", "exper")

prediction(mroz_logit, at = test_individual)
```

```
## Data frame with 753 predictions from
##  glm(formula = inlf ~ nwifeinc + educ + kidslt6 + age + exper,
##     family = binomial(link = "logit"), data = mroz)
## with average prediction:

##  nwifeinc educ kidslt6 age exper      x
##        20   10       0  30    10 0.7881
```

```
margins(mroz_logit, at = test_individual)
```

```
## Average marginal effects at specified values

## glm(formula = inlf ~ nwifeinc + educ + kidslt6 + age + exper,     family = binomial(link = "logit"),

##  at(nwifeinc) at(educ) at(kidslt6) at(age) at(exper)  nwifeinc   educ kidslt6
##            20       10           0      30        10 -0.003323 0.0373 -0.2444
##       age   exper
##  -0.01589 0.01969
```

```
prediction(mroz_probit, at = test_individual)
```

```
## Data frame with 753 predictions from
##  glm(formula = inlf ~ nwifeinc + educ + kidslt6 + age + exper,
##     family = binomial(link = "probit"), data = mroz)
## with average prediction:

##  nwifeinc educ kidslt6 age exper    x
##        20   10       0  30    10 0.79
```

```
margins(mroz_probit, at = test_individual)
```

```
## Average marginal effects at specified values

## glm(formula = inlf ~ nwifeinc + educ + kidslt6 + age + exper,     family = binomial(link = "probit")

##  at(nwifeinc) at(educ) at(kidslt6) at(age) at(exper)  nwifeinc    educ kidslt6
##            20       10           0      30        10 -0.003277 0.03791 -0.2554
##       age   exper
##  -0.01669 0.01993
```

The logit value returned 0.788100707829806 compared to probit's 0.789979284631276. Logit is lower since the cumulative normal distribution approaches 1 faster than the logistic CDF.

c

The logit partial effect of experience is 0.0196869 compared to probit's 0.0199296. Since probit approaches 1 faster, the slope is steeper, thus a higher marginal effect of experience.

d

```
wald_test_age_exper0 <- wald.test(Sigma = vcov(mroz_logit), b = coef(mroz_logit), Terms = c(5, 6))
wald_test_age_exper0_pvalue <- wald_test_age_exper0$result$chi2[3] # below minimum printed value
wald_test_age_exper0
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 94.5, df = 2, P(> X2) = 0.0
```

$$H_0 : \beta_{age} = \beta_{exper} = 0, H_1 : otherwise \tag{1}$$

$$\alpha = 0.05,\ \chi^2 - test \tag{2}$$

The p-value was 0, which is less than 0.05. Thus we have sufficient evidence to reject the null hypothesis.

2

a

$$Instrumental\ relevance:\ Cov(z,x) \neq 0 \tag{3}$$
$$Instrumental\ exogeneity:\ Cov(z,u) = 0 \tag{4}$$

b

$$\hat{Cov}(z,y) \tag{5}$$
$$= \hat{Cov}(z, \hat{\beta}_1 x_i) + \hat{Cov}(z, \hat{u}_i) \tag{6}$$
$$= \hat{\beta}_1 \hat{Cov}(z, x_i) \tag{7}$$
$$\hat{\beta}_1 = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})} \tag{8}$$

c

$$\hat{\beta}_{1,IV} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})} \tag{9}$$
$$= \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i)}{\sum_{i=1}^{n}(z_i - \bar{z})x_i} \tag{10}$$
$$\approx \frac{\sum_{i=1}^{n}(z_i - \bar{z})\beta_0 + \beta_1\sum_{i=1}^{n}(z_i - \bar{z})x_i + \sum_{i=1}^{n}(z_i - \bar{z})(u_i - \bar{u})}{\sum_{i=1}^{n}(z_i - \bar{z})x_i} \tag{11}$$
$$= \beta_1 + \frac{\frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})(u_i - \bar{u})}{\frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})x_i} \tag{12}$$

As the $n \to \infty$, and the sample gets larger, we take limits on both sides. We know that $lim_{n\to\infty}\frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})u_i = Cov(z,u)$ and $lim_{n\to\infty}\frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})x_i = Cov(z,x)$, so

$$lim_{n\to\infty}\hat{\beta}_1 = lim_{n\to\infty}(\beta_1 + \frac{\frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})(u_i - \bar{u})}{\frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})x_i}) \tag{13}$$
$$= \beta_1 + \frac{0}{Cov(z,x)} \tag{14}$$
$$= \beta_1 \tag{15}$$

So $\beta_{1,IV}$ is a consistent estimator.

d

We assume that the homoskedasticity assumption holds.

$$
\begin{aligned}
& Var(\hat{\beta}_{1,IV}|z,x) && (16)\\
&= Var(\frac{\sum_{i=1}^n (z_i-\bar{z})(u_i-\bar{u})}{\sum_{i=1}^n (z_i-\bar{z})(x_i-\bar{x})}|z,x) && (17)\\
&= \frac{Var(\sum_{i=1}^n (z_i-\bar{z})(u_i-\bar{u})|z,x)}{[\sum_{i=1}^n (z_i-\bar{z})(x_i-\bar{x})]^2} && (18)\\
&= \frac{\sum_{i=1}^n (z_i-\bar{z})^2 Var(u_i|z,x)}{[\sum_{i=1}^n (z_i-\bar{z})(x_i-\bar{x})]^2} && (19)\\
&= \frac{\sigma^2 \sum_{i=1}^n (z_i-\bar{z})^2}{[\sum_{i=1}^n (z_i-\bar{z})(x_i-\bar{x})]^2} \ \textit{since } z \textit{ and } u \textit{ are i.i.d.} && (20)\\
&= \frac{\sigma^2}{\sum_{i=1}^n (x_i-\bar{x})^2} \cdot \frac{\sum_{i=1}^n (x_i-\bar{x})^2 \cdot \sum_{i=1}^n (z_i-\bar{z})^2}{[\sum_{i=1}^n (z_i-\bar{z})(x_i-\bar{x})]^2} && (21)\\
&= \frac{\sigma^2}{\sum_{i=1}^n (x_i-\bar{x})^2} \cdot \frac{1}{[\frac{\sum_{i=1}^n (z_i-\bar{z})(x_i-\bar{x})}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}\cdot\sqrt{\sum_{i=1}^n (z_i-\bar{z})^2}}]^2} && (22)\\
&= \frac{\sigma^2}{\sum_{i=1}^n (x_i-\bar{x})^2} \cdot \frac{1}{\rho_{z,x}^2} && (23)
\end{aligned}
$$

Since $-1 \leq \frac{\sum_{i=1}^n (z_i-\bar{z})(x_i-\bar{x})}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}\cdot\sqrt{\sum_{i=1}^n (z_i-\bar{z})^2}} \leq 1$, then $0 \leq [\frac{\sum_{i=1}^n (z_i-\bar{z})(x_i-\bar{x})}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}\cdot\sqrt{\sum_{i=1}^n (z_i-\bar{z})^2}}]^2 \leq 1$, and $1 \leq \frac{1}{\rho_{z,x}^2}$, and

$$
Var(\hat{\beta}_{1,OLS}|z,x) = \frac{\sigma^2}{\sum_{i=1}^n (x_i-\bar{x})^2} \leq \frac{\sigma^2}{\sum_{i=1}^n (x_i-\bar{x})^2} \cdot \frac{1}{\rho_{z,x}^2} = Var(\hat{\beta}_{1,IV}|z,x)
$$

3

a

```
wage_model_IV <- ivreg(formula = lwage ~ educ| sibs, data = wage2)
wage_model_sibs <- lm(formula = lwage ~ sibs, data = wage2)
summary(wage_model_sibs)
```

```
##
## Call:
## lm(formula = lwage ~ sibs, data = wage2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.97662 -0.25857  0.02503  0.28572  1.22677
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.861076   0.022078 310.771  < 2e-16 ***
## sibs        -0.027904   0.005908  -4.723 2.68e-06 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4164 on 933 degrees of freedom
## Multiple R-squared:  0.02335,    Adjusted R-squared:  0.0223
## F-statistic: 22.31 on 1 and 933 DF,  p-value: 2.68e-06
```

On average, an additional sibling changes wages by -2.7904412% approximately.
The formulae used to calculate them are different. With IV, the model uses the fitted values from regressing education on sibs instead of just the educ values.
With more siblings, a family has less income to distribute to each child, holding income constant. So it is likely that they are unable to afford the same education than if they were to have less children. With less education, the likelihood of securing employment falls, so wages are lower on average.

b

```
educ_brthord_model <- lm(formula = educ ~ brthord, data = wage2)
summary(educ_brthord_model)
```

```
##
## Call:
## lm(formula = educ ~ brthord, data = wage2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8668 -1.5842 -0.7362  2.1332  6.1117
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.14945    0.12868 109.962  < 2e-16 ***
## brthord     -0.28264    0.04629  -6.106 1.55e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.155 on 850 degrees of freedom
##   (83 observations deleted due to missingness)
## Multiple R-squared:  0.04202,    Adjusted R-squared:  0.04089
## F-statistic: 37.29 on 1 and 850 DF,  p-value: 1.551e-09
```

```
# statistically significant, fulfills instrumental relevancy condition
```

$$H_0 : \beta_{brthord} = 0, H_1 : otherwise \tag{24}$$

$$\alpha = 0.05,\ \chi^2 - test \tag{25}$$

The p-value is 1.0204424×10^{-9}, which is less than 0.05. Thus, there is sufficient evidence to reject the null hypothesis. brthord likely fulfills the instrument relevancy condition.

c

```
reduced_form_educ <- lm(educ ~ brthord + sibs, data = wage2)
summary(reduced_form_educ)
```

```
##
## Call:
## lm(formula = educ ~ brthord + sibs, data = wage2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1438 -1.6854 -0.6852  2.0090  5.9950
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.29650    0.13329 107.260  < 2e-16 ***
## brthord     -0.15267    0.05708  -2.675 0.007619 **
## sibs        -0.15287    0.03987  -3.834 0.000135 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.137 on 849 degrees of freedom
##   (83 observations deleted due to missingness)
## Multiple R-squared:  0.05833,    Adjusted R-squared:  0.05611
## F-statistic: 26.29 on 2 and 849 DF,  p-value: 8.33e-12
```

```
wald.test(Sigma = vcov(reduced_form_educ), b = coef(reduced_form_educ), Terms = 2)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 7.2, df = 1, P(> X2) = 0.0075
```

The identification assumption means that the coefficient for brthord cannot be 0. Since we are using sibs as an exogenous variable and brthord as the instrumental variable, then we need to make sure that after partialling out the effect of sibs on education, brthord is still instrumentally relevant with education. To test for it, an LM-test was conducted, where the null hypothesis is that the coefficient is zero under $\alpha = 0.05$. The p-value returned was $0.007475 < 0.05$, so we have sufficient evidence to reject the null hypothesis. Thus, the indentification assumption can be held.

d

```
wage_model_IV_sibs_brthord <- ivreg(formula = lwage ~ educ + sibs | brthord + sibs, data = wage2)
display <- summary(wage_model_IV_sibs_brthord)
display
```

```
##
## Call:
```

```
## ivreg(formula = lwage ~ educ + sibs | brthord + sibs, data = wage2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.84808 -0.26227  0.03841  0.29901  1.30836
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.938527   1.055690   4.678 3.37e-06 ***
## educ        0.136994   0.074681   1.834   0.0669 .
## sibs        0.002111   0.017372   0.122   0.9033
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.427 on 849 degrees of freedom
## Multiple R-Squared: -0.05428,    Adjusted R-squared: -0.05676
## Wald test:  10.9 on 2 and 849 DF,  p-value: 2.124e-05
```

The standard error for educ/brthord is 0.0746812, which is large.
The standard error for sibs is 1.8343859, which is large.
Using sibs on its own when it is highly correlated with education increases the variance of the parameter estimators, since

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (1 - R^2)} \cdot \frac{1}{\hat{\rho}^2_{z,x}} \tag{26}$$

where R^2 is from the regression of education on sibs. Since sibs is significantly correlated with education, then this is closer to 1 (p-value=1.2602847×10^{-4}<0.05), thus the variance of $\hat{\beta}_1$ is higher than if we were to not use sibs (coefficient would be 0.1306448, and standard error would be 0.0320385). This is the same for $\hat{\beta}_2$ as well, since sibs is correlated with educ.

e

The correlation between $\hat{educ}$ and sibs is -0.9294818. Thus, as explained in d, R^2 is likely very close to 1, so multicollinearity problems are likely.