UNIVERSITY OF NEVADA LAS VEGAS

LEE BUSINESS SCHOOL

ECO 772- ECONOMETRICS II

PROBLEM SET #03

STUDENT NAME: DANIEL EMAASIT

MAJOR: CIVIL ENGINEERING

INSTRUCTOR NAME: Prof. Ian McDonough

DATE SUBMITTED: 04/11/2015

SPRING 2015

Problem 9.4

The following equation explains weekly hours of television viewing by a child in terms of the child's age, mother's education, father's education, and number of siblings:

$tvhours = \beta0 + \beta1age + \beta2age2 + \beta3motheduc + \beta4fatheduc + \beta5sibs + u.$

We are worried that tvhours is measured with error in our survey. Let tvhours denote the reported hours of television viewing per week.

Part (i)

What do the classical errors-in-variables (CEV) assumptions require in this application?

The CEV assumptions requires that the measurement error have zero mean and be uncorrelated with tvhours and each explanatory variable in the equation.

Part (ii)

Do you think the CEV assumptions are likely to hold? Explain.

I don't think the CEV assumptions are ikely to hold. This is because tvhours depends directly on the explanatory variables.

Problem 15.2

Suppose that you wish to estimate the effect of class attendance on student performance, as in Example 6.3. A basic model is

$stndfnl = \beta0 + \beta1atndrte + \beta2priGPA + \beta3ACT + u,$

where the variables are defined as in Chapter 6.

Part (i)

Let dist be the distance from the students' living quarters to the lecture hall. Do you think dist is uncorrelated with u?

Yes, it's possible for distance and the error term to be uncorrelated. This is because classrooms are randomly assigned without consideration for where students live.

Part (ii)

Assuming that dist and u are uncorrelated, what other assumption must dist satisfy to be a valid IV for atndrte?

The other assumption is that distance must be partially correlated with the varibale "atndrte".

Part (iii)

Suppose, as in equation (6.18), we add the interaction term priGPA.atndrte:

stndfnl = β0 + β1atndrte + β2priGPA + β3ACT + β4priGPA.atndrte + u.

If atndrte is correlated with u, then, in general, so is priGPA.atndrte. What might be a good IV for priGPA.atndrte?

A good IV for may be the interaction term between GPA and distance, priGPAβ*dist.

Problem 15.4

Suppose that, for a given state in the United States, you wish to use annual time series data to estimate the effect of the state-level minimum wage on the employment of those 18 to 25 years old (EMP). A simple model is

gEMPt = b0 + b1gMINt + b2gPOPt + b3gGSPt + b4gGDPt + ut,

where MINt is the minimum wage, in real dollars, POPt is the population from 18 to 25 years old, GSPt is gross state product, and GDPt is U.S. gross domestic product. The g prefix indicates the growth rate from year t - 1 to year t, which would typically be approximated by the difference in the logs.

Part (i)

If we are worried that the state chooses its minimum wage partly based on unobserved (to us) factors that affect youth employment, what is the problem with OLS estimation?

This means that the Minimum wage will be correlated with the unobserved error term thereby making estimates from OLS to be biased and inconsistent.

Part (ii)

Let USMINt be the U.S. minimum wage, which is also measured in real terms. Do you think gUSMINt is uncorrelated with ut?

The U.S. minimum wage may be uncorrelated with the error term because the Gross Domestic Factor controls for the overall performance of the US economy.

Part (iii)

By law, any state's minimum wage must be at least as large as the U.S. minimum. Explain why this makes gUSMINt a potential IV candidate for gMINt.

This is because the State minimum wage is correlated with the US minimum wage, that is, as the national wage increases so must the State wage. It is also uncorrelated with the disturbance term.

Problem 15.6

Part (i)

In the model with one endogenous explanatory variable, one exogenous explanatory variable, and one extra exogenous variable, take the reduced form for y2 (15.26), and plug it into the structural equation (15.22). This gives the reduced form for y1:

$y1 = \beta0 + \beta1z1 + \beta2z2 + v1.$

$y1 = \alpha0 + \alpha1z1 + \alpha2z2 + v1.$

Find the αj in terms of the βj and the πj.

$y1 = \beta0 + \beta1(\pi0 + \pi1z1 + \pi2z2 + v2) + \beta2z1 + u1$

$y1 = (\beta0 + \beta1\pi0) + (\beta1\pi1 + \beta2)z1 + \beta1\pi2z2 + u1 + \beta1v2$, hence

$\alpha0 = \beta0 + \beta1\pi0,$

$\alpha1 = \beta1\pi1 + \beta2$, and

$\alpha2 = \beta1\pi2.$

Part (ii)

Find the reduced form error, v1, in terms of u1, v2, and the parameters.

$v1 = u1 + \beta1v2.$

Part (iii)

How would you consistently estimate the αj?

v1 has zero mean and is uncorrelated with z1 and z2, which means that OLS consistently estimates the αj.

Problem C15.3

Use the data in CARD.RAW for this exercise.

Part (i)

The equation we estimated in Example 15.4 can be written as

$\log(wage) = \beta0 + \beta1educ + \beta2exper + ... + u,$

where the other explanatory variables are listed in Table 15.1. In order for IV to be consistent, the IV for educ, nearc4, must be uncorrelated with u. Could nearc4 be correlated with things in the error term, such as unobserved ability? Explain.

It's possible for proximity to college to be correlated with things in the error term such as unobserrved ability. One could argue that the presence of a college can motivate an individual to work harder in school so that they can one day make to college themselves.

Part (ii)

For a subsample of the men in the data set, an IQ score is available. Regress IQ on nearc4 to check whether average IQ scores vary by whether the man grew up near a four-year college. What do you conclude?

```
# read the data into R
library(haven)

card_data <- read_dta("CARD.dta")

# Regress IQ on nearc4
lm_model1 <- lm(iq~nearc4, data = card_data)

# Put the model in a tabular format
library(stargazer)

stargazer(lm_model1, type = "text", title = "Regression of IQ on Proximity to
College")

##
## Regression of IQ on Proximity to College
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                  iq
## ------------------------------------------------
## nearc4                         2.596***
##                                 (0.745)
##
## Constant                      100.611***
##                                 (0.627)
##
## ------------------------------------------------
## Observations                     2,061
## R2                               0.006
## Adjusted R2                      0.005
## Residual Std. Error     15.382 (df = 2059)
## F Statistic          12.128*** (df = 1; 2059)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

The model shows that the predicted IQ is 2.6 points higher if the man grew up near a four-year college. This estimate is statistically significant at a 0.01 level of confidence.

Part (iii)

Now, regress IQ on nearc4, smsa66, and the 1966 regional dummy variables reg662, ... ,reg669. Are IQ and nearc4 related after the geographic dummy variables have been partialled out? Reconcile this with your findings from part (ii).

```
# Regress IQ on nearc4, smsa66, and reg662, ... ,reg669.
lm_model2 <- lm(iq~nearc4 + smsa66 + reg662 + reg663 + reg664 + reg665 +
reg666 + reg667 + reg668 + reg669, data = card_data)

stargazer(lm_model2, type = "text", title = "Regression of IQ on More
Variables")
```

```
##
## Regression of IQ on More Variables
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                 iq
## ------------------------------------------------
## nearc4                          0.348
##                                (0.814)
##
## smsa66                          1.089
##                                (0.809)
##
## reg662                          1.099
##                                (1.650)
##
## reg663                         -1.559
##                                (1.623)
##
## reg664                         -0.543
##                                (1.916)
##
## reg665                        -8.475***
##                                (1.666)
##
## reg666                        -7.421***
##                                (1.974)
##
## reg667                        -8.394***
##                                (1.830)
##
## reg668                         -2.925
##                                (2.345)
##
## reg669                         -2.892
##                                (1.797)
##
```

```
## Constant                    104.773***
##                               (1.625)
##
## ------------------------------------------------
## Observations                   2,061
## R2                             0.063
## Adjusted R2                    0.058
## Residual Std. Error     14.969 (df = 2050)
## F Statistic          13.700*** (df = 10; 2050)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

The model shows that nearc4 is statistically insignificant. This means that proximity to a four-year college does not influence the IQ score, controling for region and environment while growing up.

Part (iv)

From parts (ii) and (iii), what do you conclude about the importance of controlling for smsa66 and the 1966 regional dummies in the log(wage) equation?

It is important because it allows control for differences in access to colleges that might also be correlated with ability.

Problem C15.5

Use the data in CARD.RAW for this exercise.

Part (i)

In Table 15.1, the difference between the IV and OLS estimates of the return to education is economically important. Obtain the reduced form residuals, v2, from the reduced form regression educ on nearc4, exper, expersq, black, smsa, south, smsa66, reg662, ..., reg669 --- see Table15.1. Use these to test whether educ is exogenous; that is, determine if the difference between OLS and IV is statistically significant.

```
# The reduced form is obtained as follows
lm_model3 <- lm(lwage ~ educ + nearc4 + exper + expersq + black + smsa +
south + smsa66 + reg662 + reg663 + reg664 + reg665 + reg666 + reg667 + reg668
+ reg669, data = card_data)

# Obtain the residuals
v_residuals <- resid(lm_model3)

# Add the residuals to the original equation
x <- card_data
lm_model4 <- lm(x$lwage ~ x$educ + x$nearc4 + x$exper + x$expersq + x$black +
x$smsa + x$south + x$smsa66 + x$reg662 + x$reg663 + x$reg664 + x$reg665 +
x$reg666 + x$reg667 + x$reg668 + x$reg669 + v_residuals)
```

```r
# Display the model results
summary(lm_model4)
```

```
## Coefficients:
##                Estimate Std. Error    t value Pr(>|t|)
## (Intercept)  4.616e+00  5.735e-17  8.048e+16   <2e-16 ***
## x$educ       7.444e-02  2.703e-18  2.754e+16   <2e-16 ***
## x$nearc4     1.825e-02  1.302e-17  1.402e+15   <2e-16 ***
## x$exper      8.473e-02  5.107e-18  1.659e+16   <2e-16 ***
## x$expersq   -2.285e-03  2.441e-19 -9.362e+15   <2e-16 ***
## x$black     -2.002e-01  1.409e-17 -1.420e+16   <2e-16 ***
## x$smsa       1.348e-01  1.554e-17  8.672e+15   <2e-16 ***
## x$south     -1.476e-01  2.003e-17 -7.370e+15   <2e-16 ***
## x$smsa66     1.999e-02  1.564e-17  1.278e+15   <2e-16 ***
## x$reg662     9.628e-02  2.768e-17  3.479e+15   <2e-16 ***
## x$reg663     1.467e-01  2.712e-17  5.408e+15   <2e-16 ***
## x$reg664     5.658e-02  3.213e-17  1.761e+15   <2e-16 ***
## x$reg665     1.307e-01  3.231e-17  4.045e+15   <2e-16 ***
## x$reg666     1.456e-01  3.507e-17  4.152e+15   <2e-16 ***
## x$reg667     1.222e-01  3.467e-17  3.525e+15   <2e-16 ***
## x$reg668    -5.318e-02  3.958e-17 -1.344e+15   <2e-16 ***
## x$reg669     1.198e-01  2.995e-17  4.001e+15   <2e-16 ***
## v_residuals  1.000e+00  1.409e-17  7.096e+16   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.87e-16 on 2992 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 4.232e+32 on 17 and 2992 DF,  p-value: < 2.2e-16
```

The coefficient on the residual is 1.0 with a t-statistic of 7. Hence the difference in the estimates of the return to education is statistically significant.

Part (ii)

Estimate the equation by 2SLS, adding nearc2 as an instrument. Does the coefficient on educ change much?

```r
# Estimate the equation by 2SLS
library(AER)

formula1 <- lwage ~ educ + nearc4 + exper + expersq + black + smsa + south +
smsa66 + reg662 + reg663 + reg664 + reg665 + reg666 + reg667 + reg668 +
reg669

inst1 <- ~ nearc2 + nearc4 + exper + expersq + black + smsa + south + smsa66
+ reg662 + reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + reg669
```

```
fit_iv1 <- ivreg(formula1, inst = inst1, data = card_data)
summary(fit_iv1)


##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.0064101  3.0347630   0.332 0.740194
## educ         0.2913608  0.1823486   1.598 0.110188
## nearc4      -0.0511381  0.0636416  -0.804 0.421730
## exper        0.1742174  0.0758554   2.297 0.021704 *
## expersq     -0.0024738  0.0005035  -4.913 9.45e-07 ***
## black        0.0027750  0.1727387   0.016 0.987184
## smsa         0.0475167  0.0793709   0.599 0.549441
## south       -0.1364209  0.0403343  -3.382 0.000728 ***
## smsa66       0.0144578  0.0309818   0.467 0.640780
## reg662       0.1133383  0.0560548   2.022 0.043273 *
## reg663       0.1527250  0.0533502   2.863 0.004230 **
## reg664       0.0311647  0.0664473   0.469 0.639094
## reg665       0.1898515  0.0804519   2.360 0.018348 *
## reg666       0.2113100  0.0881069   2.398 0.016531 *
## reg667       0.1692320  0.0785547   2.154 0.031295 *
## reg668      -0.1668247  0.1229901  -1.356 0.175072
## reg669       0.0742009  0.0700568   1.059 0.289615
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.562 on 2993 degrees of freedom
## Multiple R-Squared: -0.595,  Adjusted R-squared: -0.6036
## Wald test: 22.99 on 16 and 2993 DF,  p-value: < 2.2e-16
```

The 2SLS estimate of the coefficient on educ is now 0.29 with a standard error = 0.18. This estimate is bigger than the previous one.

Part (iii)

Test the single overidentifying restriction from part (ii).

Problem C15.8

Use the data in 401KSUBS.RAW for this exercise.

Part (i)

Estimate the equation by OLS and discuss the estimated effect of p401k.

```
# Read the data into R
ksubs_data <- read_dta("401KSUBS.dta")

# Fit a linear model by OLS
```

```r
fit_lm <- lm(pira ~ p401k + inc + incsq + age, data = ksubs_data)

# Tabulate the model results
stargazer(fit_lm, type = "text", title = "Model Results using OLS")
```

```
##
## Model Results using OLS
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                pira
## ------------------------------------------------
## p401k                         0.052***
##                               (0.010)
##
## inc                           0.009***
##                               (0.001)
##
## incsq                        -0.00002***
##                               (0.00000)
##
## age                           0.009***
##                               (0.0004)
##
## Constant                      -0.399***
##                               (0.020)
##
## ------------------------------------------------
## Observations                    9,275
## R2                              0.179
## Adjusted R2                     0.179
## Residual Std. Error       0.395 (df = 9270)
## F Statistic           505.752*** (df = 4; 9270)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

The model shows that having a 401(k) plan increases the probability of having an individual retirement account by 0.052.

Part (ii)

For the purposes of estimating the ceteris paribus tradeoff between participation in two different types of retirement savings plans, what might be a problem with ordinary least squares?

The model in part(i) does not account for the fact that different people may have different saving preferences. People who are savers tend to have both a 401(k) plan as well as an IRA. This makes the error term positively correlated with p401k.

Part (iii)

The variable e401k is a binary variable equal to one if a worker is eligible to participate in a 401(k) plan. Explain what is required for e401k to be a valid IV for p401k. Do these assumptions seem reasonable?

Condition 1: e401k has to be correlated with p401k.

Condition 2: e401k has to be uncorrelated with the unobserved residual error.

Part (iv)

Estimate the reduced form for p401k and verify that e401k has significant partial correlation with p401k. Since the reduced form is also a linear probability model, use a heteroskedasticity-robust standard error.

```
# Estimate the reduced form
fit_lm2 <- lm(p401k ~ e401k + inc + incsq + age + agesq, data = ksubs_data)

# obtain the model results using a heteroskedasticity-robust standard error
coeftest(fit_lm2, vcov = vcovHC)

##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  5.9149e-02  4.6213e-02  1.2799  0.200599
## e401k        6.8885e-01  7.9932e-03 86.1793 < 2.2e-16 ***
## inc          1.1117e-03  3.4496e-04  3.2226  0.001275 **
## incsq        1.8410e-06  2.6945e-06  0.6832  0.494476
## age         -4.7205e-03  2.2449e-03 -2.1028  0.035509 *
## agesq        5.2037e-05  2.5730e-05  2.0224  0.043162 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Eligibility for 401k is statistically significant correlated with the propbability of having a 401k at a 0.01 level of confidence.

Part (v)

Now, estimate the structural equation by IV and compare the estimate of ??1 with the OLS estimate. Again, you should obtain heteroskedasticity-robust standard errors.

```
# Estimate the structural equation by IV
formula2 <- pira ~ p401k + inc + incsq + age + agesq
inst2 <- ~ e401k + inc + incsq + age + agesq
fit_iv2 <- ivreg(formula2, inst = inst2, data = ksubs_data)
stargazer(fit_iv2, type = "text")

##
```

```
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                   pira
## ------------------------------------------------
## p401k                              0.021
##                                   (0.013)
##
## inc                               0.009***
##                                   (0.001)
##
## incsq                            -0.00002***
##                                   (0.00000)
##
## age                                -0.001
##                                   (0.003)
##
## agesq                             0.0001***
##                                   (0.00004)
##
## Constant                          -0.207***
##                                   (0.069)
##
## ------------------------------------------------
## Observations                       9,275
## R2                                 0.179
## Adjusted R2                        0.178
## Residual Std. Error      0.395 (df = 9269)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

The IV estimate for p401k is less than the OLS estimate.

Part (vi)

Test the null hypothesis that p401k is in fact exogenous, using a heteroskedasticityrobust test.

```
# Obtain the reduced form residuals from part (iv)
v_residuals2 <- resid(fit_lm2)

# Add the residuals to the structural equation and run OLS
fit_lm3 <- lm(ksubs_data$pira ~ ksubs_data$p401k + ksubs_data$inc +
ksubs_data$incsq + ksubs_data$age + ksubs_data$agesq + v_residuals2)

# obtain the model results using a heteroskedasticity-robust test
coeftest(fit_lm3, vcov = vcovHC)

##
```

```
## t test of coefficients:
##
##                      Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)      -2.0731e-01  6.5297e-02 -3.1749  0.001504 **
## ksubs_data$p401k  2.0701e-02  1.3227e-02  1.5651  0.117597
## ksubs_data$inc    8.9982e-03  4.9292e-04 18.2548 < 2.2e-16 ***
## ksubs_data$incsq -2.4136e-05  3.9014e-06 -6.1866 6.409e-10 ***
## ksubs_data$age   -1.1466e-03  3.2436e-03 -0.3535  0.723717
## ksubs_data$agesq  1.1207e-04  3.8268e-05  2.9286  0.003414 **
## v_residuals2      7.4794e-02  1.9108e-02  3.9143 9.133e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The model shows that the residual is statistically significant with t value of 3.9. This means that p401k is endogenous.

Problem C15.9

The purpose of this exercise is to compare the estimates and standard errors obtained by correctly using 2SLS with those obtained using inappropriate procedures. Use the data file WAGE2.RAW.

Part (i)

Use a 2SLS routine to estimate the equation

log(wage) = β0 + β1educ + β2exper + β3tenure + β4black + u,

where sibs is the IV for educ. Report the results in the usual form.

```
# Read the data into R
wage_data <- read_dta("WAGE2.dta")

# Use 2SLS
formula3 <- lwage ~ educ + exper + tenure + black
inst3 <- ~sibs + exper + tenure + black
fit_iv3 <- ivreg(formula3, inst = inst3, data = wage_data)
stargazer(fit_iv3, type = "text")
```

```
##
```

```
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                  lwage
## ------------------------------------------------
## educ                           0.094***
##                                 (0.034)
##
## exper                           0.021**
##                                 (0.008)
##
## tenure                         0.012***
##                                 (0.003)
##
## black                          -0.183***
##                                 (0.050)
##
## Constant                       5.216***
##                                 (0.543)
##
## ------------------------------------------------
## Observations                      935
## R2                               0.169
## Adjusted R2                      0.165
## Residual Std. Error       0.385 (df = 930)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Part (ii)

Now, manually carry out 2SLS. That is, first regress educ on sibs, exper, tenure, and black and obtain the fitted values. Then, run the second stage regression log(wage) on educ estimate, exper, tenure, and black. Verify that the estimate ?? coefficients are identical to those obtained from part (i), but that the standard errors are somewhat different. The standard errors obtained from the second stage regression when manually carrying out 2SLS are generally inappropriate.

```
# First regression
fit_lm4 <- lm(educ ~ sibs + exper + tenure + black, data = wage_data)

# Obtain the fitted values
educ_fitted <- fitted(fit_lm4)

# Then run the second stage regression
fit_lm5 <- lm(wage_data$lwage ~ educ_fitted + wage_data$exper +
wage_data$tenure + wage_data$black)

# Display the model results
stargazer(fit_lm5, type = "text")
```

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                  lwage
## ------------------------------------------------
## educ_fitted                    0.094***
##                                 (0.035)
##
## exper                           0.021**
##                                 (0.009)
##
## tenure                         0.012***
##                                 (0.003)
##
## black                          -0.183***
##                                 (0.052)
##
## Constant                       5.216***
##                                 (0.569)
##
## ------------------------------------------------
## Observations                      935
## R2                               0.089
## Adjusted R2                      0.085
## Residual Std. Error       0.403 (df = 930)
## F Statistic            22.747*** (df = 4; 930)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

The standard errors in part(i) and part(ii) are 0.034 and 0.035 respectively.

Part (iii)

Now, use the following two-step procedure, which generally yields inconsistent parameter estimates of the ??j, and not just inconsistent standard errors. In step one, regress educ on sibs only and obtain the fitted values, say educ estimated. (Note that this is an incorrect first stage regression.) Then, in the second step, run the regression of log(wage) on educ_estimated, , experi , tenure, and black. How does the estimate from this incorrect, two-step procedure compare with the correct 2SLS estimate of the return to education?

```
# First regression
fit_lm6 <- lm(educ ~ sibs, data = wage_data)

# Obtain the fitted values
educ_fitted2 <- fitted(fit_lm6)

# Then run the second stage regression
fit_lm7 <- lm(wage_data$lwage ~ educ_fitted2 + wage_data$exper +
wage_data$tenure + wage_data$black)
```

```
# Display the model results
stargazer(fit_lm7, type = "text")
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                lwage
## ------------------------------------------------
## educ_fitted2                 0.070***
##                               (0.026)
##
## exper                         -0.0004
##                               (0.003)
##
## tenure                       0.014***
##                               (0.003)
##
## black                        -0.242***
##                               (0.042)
##
## Constant                     5.771***
##                               (0.360)
##
## ------------------------------------------------
## Observations                    935
## R2                             0.089
## Adjusted R2                    0.085
## Residual Std. Error      0.403 (df = 930)
## F Statistic           22.747*** (df = 4; 930)
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

The incorrent method shows a coefficient on 0.07 for estimated eduction. This is lower than the estimate from the correct method of 0.094.

Problem C15.10

Use the data in HTV.RAW for this exercise.

Part (i)

Run a simple OLS regression of log(wage) on educ. Without controlling for other factors, what is the 95% confidence interval for the return to another year of education?

```
# Read the data into R
htv_data <- read_dta("HTV.dta")

# Run a simple OLS
```

```
fit_lm8 <- lm(lwage ~ educ, data = htv_data)
stargazer(fit_lm8, type = "text", ci = TRUE, ci.level = 0.95)
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                lwage
## ------------------------------------------------
## educ                          0.101***
##                             (0.088, 0.114)
##
## Constant                      1.092***
##                             (0.921, 1.263)
##
## ------------------------------------------------
## Observations                    1,230
## R2                              0.162
## Adjusted R2                     0.161
## Residual Std. Error       0.544 (df = 1228)
## F Statistic           236.622*** (df = 1; 1228)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

The 95% confidence interval is [0.088, 0.114].

Part (ii)

The variable ctuit, in thousands of dollars, is the change in college tuition facing students from age 17 to age 18. Show that educ and ctuit are essentially uncorrelated. What does this say about ctuit as a possible IV for educ in a simple regression analysis?

```
# show that educ and ctuit are uncorrelated
fit_lm9 <- lm(educ ~ ctuit, data = htv_data)
summary(fit_lm9)
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.03836    0.06717 194.117   <2e-16 ***
## ctuit       -0.04945    0.08352  -0.592    0.554
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.355 on 1228 degrees of freedom
## Multiple R-squared:  0.0002853,  Adjusted R-squared:  -0.0005288
## F-statistic: 0.3505 on 1 and 1228 DF,  p-value: 0.5539
```

The t-statistic of -0.59 is less than the critical t-value of 1.96. We fail to reject the null hypothesis that the estimate for ctuit is zero. We conclude that educ and ctuit are uncorrelated. This means that ctuit is not a good IV for educ.

Part (iii)

Now, add to the simple regression model in part (i) a quadratic in experience and a full set of regional dummy variables for current residence and residence at age 18. Also include the urban indicators for current and age 18 residences. What is the estimated return to a year of education?

```
# Add more variables to the simple regression
fit_lm10 <- lm(lwage ~ educ + exper + expersq + ne + nc + west + ne18 + nc18
+ west18 + urban + urban18, data = htv_data)
stargazer(fit_lm10, type = "text")
```

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                 lwage
## ------------------------------------------------
## educ                           0.137***
##                                 (0.009)
##
## exper                          0.112***
##                                 (0.027)
##
## expersq                        -0.003**
##                                 (0.001)
##
## ne                              -0.017
##                                 (0.086)
##
## nc                              -0.017
##                                 (0.071)
##
## west                             0.018
##                                 (0.081)
##
## ne18                            0.156*
##                                 (0.087)
##
## nc18                             0.011
##                                 (0.073)
##
## west18                          -0.030
##                                 (0.086)
##
## urban                          0.205***
```

```
##                                 (0.042)
##
## urban18                        0.126***
##                                 (0.049)
##
## Constant                       -0.507**
##                                 (0.241)
##
## ------------------------------------------------
## Observations                     1,230
## R2                               0.219
## Adjusted R2                      0.212
## Residual Std. Error       0.527 (df = 1218)
## F Statistic           30.978*** (df = 11; 1218)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

The estimated return to a year of education is 13.7%.

Part (iv)

Again using ctuit as a potential IV for educ, estimate the reduced form for educ. [Naturally, the reduced form for educ now includes the explanatory variables in part (iii).] Show that ctuit is now statistically significant in the reduced form for educ.

```
# Use ctuit as a potential IV
fit_lm11 <- lm(educ ~ ctuit + exper + expersq + ne + nc + west + ne18 + nc18
+ west18 + urban + urban18, data = htv_data)
stargazer(fit_lm11, type = "text")
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                educ
## ------------------------------------------------
## ctuit                         -0.165***
##                                (0.060)
##
## exper                         -0.874***
##                                (0.080)
##
## expersq                       0.016***
##                                (0.004)
##
## ne                             -0.375
##                                (0.270)
##
## nc                             -0.142
##                                (0.224)
##
```

```
## west                             0.622**
##                                  (0.253)
##
## ne18                             0.653**
##                                  (0.272)
##
## nc18                              0.232
##                                  (0.229)
##
## west18                           -0.448*
##                                  (0.271)
##
## urban                            -0.077
##                                  (0.131)
##
## urban18                         -0.989***
##                                  (0.151)
##
## Constant                        21.242***
##                                  (0.459)
##
## ------------------------------------------------
## Observations                      1,230
## R2                                0.509
## Adjusted R2                       0.504
## Residual Std. Error      1.657 (df = 1218)
## F Statistic         114.701*** (df = 11; 1218)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Change in tuition is now statistically significant at 0.01 level of confidence. Hence an increase in tuition fees reduces eduction by 0.165.

Part (v)

Estimate the model from part (iii) by IV, using ctuit as an IV for educ. How does the confidence interval for the return to education compare with the OLS CI from part (iii)?

```
#Estimate using IV method
formula4 <- lwage ~ educ + exper + expersq + ne + nc + west + ne18 + nc18 +
west18 + urban + urban18
inst4 <- ~ctuit + exper + expersq + ne + nc + west + ne18 + nc18 + west18 +
urban + urban18
fit_iv4 <- ivreg(formula4, inst = inst4, data = htv_data)
stargazer(fit_iv4, type = "text", ci = TRUE, ci.level = 0.95)

##
```

```
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                  lwage
## ------------------------------------------------
## educ                             0.250**
##                            (0.011, 0.489)
##
## exper                            0.209*
##                            (-0.003, 0.422)
##
## expersq                         -0.005**
##                           (-0.009, -0.0004)
##
## ne                               0.029
##                            (-0.174, 0.232)
##
## nc                               0.003
##                            (-0.151, 0.157)
##
## west                            -0.054
##                            (-0.281, 0.172)
##
## ne18                             0.076
##                            (-0.171, 0.323)
##
## nc18                            -0.021
##                            (-0.187, 0.145)
##
## west18                           0.023
##                            (-0.188, 0.235)
##
## urban                          0.215***
##                            (0.125, 0.304)
##
## urban18                          0.237*
##                            (-0.018, 0.493)
##
## Constant                        -2.894
##                            (-7.954, 2.165)
##
## ------------------------------------------------
## Observations                     1,230
## R2                               0.120
## Adjusted R2                      0.112
## Residual Std. Error      0.560 (df = 1218)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

The confidence interval in the IV method is [0.011, 0.489] which is wider than the one from OLS.

Part (vi)

Do you think the IV procedure from part (v) is convincing?

The IV estimate has a large standared error thereby showing that the IV procedure is not convincing. Infact ctuit was found not to correlated with eduction.