R and Stata Workshop: Using R

Economic and Finance Department
Brunel University London, UK

Michael Willox

2024-06-09

Setup

Use *mus*08*psidextract.dta* to set up panel data in R, and define panel variable id, and time variable t. Consider the following model:

$$lwage_{it} = \alpha + \beta_1 exp_{it} + \beta_2 exp2_{it} + \beta_3 wks_{it} + \beta_4 ed_{it} + \mu_i + \epsilon_{it}$$

Summarize and describe the dataset.

```
data <- haven::read_dta("../Data/mus08psidextract.dta")
summary(data)
```

```
##       exp              wks              occ              ind        
##  Min.   : 1.00   Min.   : 5.00   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:11.00   1st Qu.:46.00   1st Qu.:0.0000   1st Qu.:0.0000  
##  Median :18.00   Median :48.00   Median :1.0000   Median :0.0000  
##  Mean   :19.85   Mean   :46.81   Mean   :0.5112   Mean   :0.3954  
##  3rd Qu.:29.00   3rd Qu.:50.00   3rd Qu.:1.0000   3rd Qu.:1.0000  
##  Max.   :51.00   Max.   :52.00   Max.   :1.0000   Max.   :1.0000  
##      south             smsa              ms               fem        
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:0.0000  
##  Median :0.0000   Median :1.0000   Median :1.0000   Median :0.0000  
##  Mean   :0.2903   Mean   :0.6538   Mean   :0.8144   Mean   :0.1126  
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000  
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
##      union             ed              blk              lwage      
##  Min.   :0.000   Min.   : 4.00   Min.   :0.00000   Min.   :4.605  
##  1st Qu.:0.000   1st Qu.:12.00   1st Qu.:0.00000   1st Qu.:6.395  
##  Median :0.000   Median :12.00   Median :0.00000   Median :6.685  
##  Mean   :0.364   Mean   :12.85   Mean   :0.07227   Mean   :6.676  
##  3rd Qu.:1.000   3rd Qu.:16.00   3rd Qu.:0.00000   3rd Qu.:6.953  
##  Max.   :1.000   Max.   :17.00   Max.   :1.00000   Max.   :8.537  
##        id               t          tdum1             tdum2             tdum3        
##  Min.   :  1   Min.   :1   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:149   1st Qu.:2   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000  
##  Median :298   Median :4   Median :0.0000   Median :0.0000   Median :0.0000  
##  Mean   :298   Mean   :4   Mean   :0.1429   Mean   :0.1429   Mean   :0.1429
```

```
##  3rd Qu.:447   3rd Qu.:6   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :595   Max.   :7   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##      tdum4            tdum5            tdum6            tdum7
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.1429   Mean   :0.1429   Mean   :0.1429   Mean   :0.1429
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##       exp2
##  Min.   :   1.0
##  1st Qu.: 121.0
##  Median : 324.0
##  Mean   : 514.4
##  3rd Qu.: 841.0
##  Max.   :2601.0
```

```
glimpse(data)
```

```
## Rows: 4,165
## Columns: 22
## $ exp   <dbl> 3, 4, 5, 6, 7, 8, 9, 30, 31, 32, 33, 34, 35, 36, 6, 7, 8, 9, 10,~
## $ wks   <dbl> 32, 43, 40, 39, 42, 35, 32, 34, 27, 33, 30, 30, 37, 30, 50, 51, ~
## $ occ   <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ ind   <dbl> 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0~
## $ south <dbl> 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ smsa  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1~
## $ ms    <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0~
## $ fem   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1~
## $ union <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0~
## $ ed    <dbl> 9, 9, 9, 9, 9, 9, 9, 11, 11, 11, 11, 11, 11, 11, 12, 12, 12, 12,~
## $ blk   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1~
## $ lwage <dbl> 5.56068, 5.72031, 5.99645, 5.99645, 6.06146, 6.17379, 6.24417, 6~
## $ id    <dbl> 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 4~
## $ t     <dbl> 1, 2, 3, 4, 5, 6, 7, 1, 2, 3, 4, 5, 6, 7, 1, 2, 3, 4, 5, 6, 7, 1~
## $ tdum1 <dbl> 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1~
## $ tdum2 <dbl> 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0~
## $ tdum3 <dbl> 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0~
## $ tdum4 <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0~
## $ tdum5 <dbl> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0~
## $ tdum6 <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0~
## $ tdum7 <dbl> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0~
## $ exp2  <dbl> 9, 16, 25, 36, 49, 64, 81, 900, 961, 1024, 1089, 1156, 1225, 129~
```

Questions 1. Determine if this panel data is the short or long panel.

Given that the intercept term in equation 1 of the assignment is common to all units, it suggests that the model being considered is characterized by random effects, which assumes the term u_i is not correlated with the regressors, X_{it}.

Use the distinct command to count distinct values for the time and panel variables. The dataset is short because $n = 595 > T = 7$. The total number of observations, N, is 4165.

```r
distinct_ids <- n_distinct(data$id)
distinct_times <- n_distinct(data$t)

distinct_ids
```

```
## [1] 595
```

```r
distinct_times
```

```
## [1] 7
```

Questions 2. Run pooled OLS, fixed effects and random effects regressions?

Pooled OLS with a single intercept.

```r
ols_model <- lm(lwage ~ exp + exp2 + wks + ed, data = data)
summary(ols_model)
```

```
##
## Call:
## lm(formula = lwage ~ exp + exp2 + wks + ed, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.16058 -0.25035  0.00027  0.26792  2.12969
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.908e+00  6.733e-02  72.894  < 2e-16 ***
## exp          4.468e-02  2.393e-03  18.670  < 2e-16 ***
## exp2        -7.156e-04  5.279e-05 -13.555  < 2e-16 ***
## wks          5.827e-03  1.183e-03   4.927 8.67e-07 ***
## ed           7.604e-02  2.227e-03  34.151  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3908 on 4160 degrees of freedom
## Multiple R-squared:  0.2836, Adjusted R-squared:  0.2829
## F-statistic: 411.6 on 4 and 4160 DF,  p-value: < 2.2e-16
```

Fixed effects with unit specific intercepts and unit-specific, time-invariant error term that is uncorrelated with the explanatory variables. u_i is assumed to be correlated with the regressors, X_{it}.

```r
fe_model <- plm(lwage ~ exp + exp2 + wks + ed, data = data,
                index = c("id", "t"), model = "within")
summary(fe_model)
```

```
## Oneway (individual) effect Within Model
##
## Call:
```

```
## plm(formula = lwage ~ exp + exp2 + wks + ed, data = data, model = "within",
##     index = c("id", "t"))
##
## Balanced Panel: n = 595, T = 7, N = 4165
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -1.8120877 -0.0511129  0.0037112  0.0614251  1.9434064
##
## Coefficients:
##         Estimate  Std. Error t-value  Pr(>|t|)
## exp   1.1379e-01  2.4689e-03 46.0888 < 2.2e-16 ***
## exp2 -4.2437e-04  5.4632e-05 -7.7678 1.036e-14 ***
## wks   8.3588e-04  5.9967e-04  1.3939    0.1634
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    240.65
## Residual Sum of Squares: 82.632
## R-Squared:      0.65663
## Adj. R-Squared: 0.59916
## F-statistic: 2273.74 on 3 and 3567 DF, p-value: < 2.22e-16
```

Random effects with single intercept and a unit specific error term (random effect), which is uncorrelated with the explanatory variables, varies across units, is constant over time for each unit and is separate from the idiosyncratic error term.

```
re_model <- plm(lwage ~ exp + exp2 + wks + ed, data = data,
                index = c("id", "t"), model = "random")
summary(re_model)
```

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## Call:
## plm(formula = lwage ~ exp + exp2 + wks + ed, data = data, model = "random",
##     index = c("id", "t"))
##
## Balanced Panel: n = 595, T = 7, N = 4165
##
## Effects:
##                   var std.dev share
## idiosyncratic 0.02317 0.15220 0.185
## individual    0.10209 0.31952 0.815
## theta: 0.8228
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -2.0439674 -0.1057049  0.0070993  0.1147499  2.0875838
##
## Coefficients:
##                Estimate  Std. Error  z-value Pr(>|z|)
## (Intercept)  3.8294e+00  9.3634e-02  40.8974   <2e-16 ***
```

```
## exp          8.8861e-02  2.8178e-03  31.5360   <2e-16 ***
## exp2        -7.7257e-04  6.2262e-05 -12.4083   <2e-16 ***
## wks          9.6577e-04  7.4329e-04   1.2993   0.1938
## ed           1.1171e-01  6.0572e-03  18.4426   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    260.94
## Residual Sum of Squares: 151.35
## R-Squared:      0.42
## Adj. R-Squared: 0.41945
## Chisq: 3012.45 on 4 DF, p-value: < 2.22e-16
```

Questions 3. Does this model have multicollinearity or heteroscedasticity?

Pooled OLS exhibits multicollinearity.

```
vif_values <- car::vif(ols_model)
vif_values
```

```
##       exp      exp2       wks        ed
## 18.771894 18.768104  1.003096  1.050545
```

The *VIF* measures how much the variance of a regression coefficient is inflated due to multicollinearity with other predictors. A *VIF* value greater than 10 is often considered indicative of high multicollinearity, which can affect the stability and interpretation of the regression coefficients.

Although the mean *VIF* of 9.90 is just below 10, the individual *VIF* values for exp and exp^2 are of more concern. Centering variables can help reduce multicollinearity. This involves subtracting the mean of a variable from each of its values and then using this centered variable in the regression. Note that $estatvif$ does not work with $xtreg$ combined with the fe or re options.

Here, the bar over the variable represents the centred or demeaned variable.

$$\overline{lwage_{it}} = \alpha + \beta_1 \overline{exp_{it}} + \beta_2 \overline{exp_{it}^2} + \beta_3 \overline{wks_{it}} + \beta_4 \overline{ed_{it}} + \mu_i + \epsilon_{it}$$

```
# Generate de-meaned variables
data <- data %>%
  mutate(across(c(exp, exp2, wks, ed), ~ . - mean(.),
                .names = "centered_{col}"))

centered_ols_model <- lm(lwage ~ centered_exp + centered_exp2 +
                           centered_wks + centered_ed, data = data)
summary(centered_ols_model)
```

```
##
## Call:
## lm(formula = lwage ~ centered_exp + centered_exp2 + centered_wks +
##     centered_ed, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -2.16058 -0.25035  0.00027  0.26792  2.12969
##
## Coefficients:
##                 Estimate Std. Error  t value Pr(>|t|)
## (Intercept)    6.676e+00  6.056e-03 1102.465  < 2e-16 ***
## centered_exp   4.468e-02  2.393e-03   18.670  < 2e-16 ***
## centered_exp2 -7.156e-04  5.279e-05  -13.555  < 2e-16 ***
## centered_wks   5.827e-03  1.183e-03    4.927 8.67e-07 ***
## centered_ed    7.604e-02  2.227e-03   34.151  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3908 on 4160 degrees of freedom
## Multiple R-squared:  0.2836, Adjusted R-squared:  0.2829
## F-statistic: 411.6 on 4 and 4160 DF,  p-value: < 2.2e-16
```

```
car::vif(centered_ols_model)
```

```
##  centered_exp centered_exp2  centered_wks   centered_ed
##     18.771894     18.768104      1.003096      1.050545
```

```
bptest(ols_model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  ols_model
## BP = 40.252, df = 4, p-value = 3.838e-08
```

The result, *Prob > chi2 = 0.9763* indicates that the null hypothesis that the residuals are homoscedastic and cannot be rejected at standard levels of statistical significance.

Questions 4. Which method is suitable for this model, pooled OLS regression or a random effects model?

```
ols_model <- lm(lwage ~ exp + exp2 + wks + ed, data = data)
fe_model <- plm(lwage ~ exp + exp2 + wks + ed, data = data,
                index = c("id", "t"), model = "within")
re_model <- plm(lwage ~ exp + exp2 + wks + ed, data = data,
                index = c("id", "t"), model = "random")

summary(ols_model)
```

```
##
## Call:
## lm(formula = lwage ~ exp + exp2 + wks + ed, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -2.16058 -0.25035  0.00027  0.26792  2.12969
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.908e+00  6.733e-02  72.894  < 2e-16 ***
## exp          4.468e-02  2.393e-03  18.670  < 2e-16 ***
## exp2        -7.156e-04  5.279e-05 -13.555  < 2e-16 ***
## wks          5.827e-03  1.183e-03   4.927 8.67e-07 ***
## ed           7.604e-02  2.227e-03  34.151  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3908 on 4160 degrees of freedom
## Multiple R-squared:  0.2836, Adjusted R-squared:  0.2829
## F-statistic: 411.6 on 4 and 4160 DF,  p-value: < 2.2e-16
```

```
summary(fe_model)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = lwage ~ exp + exp2 + wks + ed, data = data, model = "within",
##     index = c("id", "t"))
##
## Balanced Panel: n = 595, T = 7, N = 4165
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -1.8120877 -0.0511129  0.0037112  0.0614251  1.9434064
##
## Coefficients:
##         Estimate  Std. Error t-value  Pr(>|t|)
## exp   1.1379e-01  2.4689e-03 46.0888 < 2.2e-16 ***
## exp2 -4.2437e-04  5.4632e-05 -7.7678 1.036e-14 ***
## wks   8.3588e-04  5.9967e-04  1.3939    0.1634
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    240.65
## Residual Sum of Squares: 82.632
## R-Squared:      0.65663
## Adj. R-Squared: 0.59916
## F-statistic: 2273.74 on 3 and 3567 DF, p-value: < 2.22e-16
```

```
summary(re_model)
```

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## Call:
## plm(formula = lwage ~ exp + exp2 + wks + ed, data = data, model = "random",
##     index = c("id", "t"))
##
```

```
## Balanced Panel: n = 595, T = 7, N = 4165
##
## Effects:
##                   var std.dev share
## idiosyncratic 0.02317 0.15220 0.185
## individual    0.10209 0.31952 0.815
## theta: 0.8228
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -2.0439674 -0.1057049  0.0070993  0.1147499  2.0875838
##
## Coefficients:
##                Estimate  Std. Error  z-value Pr(>|z|)
## (Intercept)  3.8294e+00  9.3634e-02  40.8974   <2e-16 ***
## exp          8.8861e-02  2.8178e-03  31.5360   <2e-16 ***
## exp2        -7.7257e-04  6.2262e-05 -12.4083   <2e-16 ***
## wks          9.6577e-04  7.4329e-04   1.2993   0.1938
## ed           1.1171e-01  6.0572e-03  18.4426   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    260.94
## Residual Sum of Squares: 151.35
## R-Squared:      0.42
## Adj. R-Squared: 0.41945
## Chisq: 3012.45 on 4 DF, p-value: < 2.22e-16
```

Based on the results, a random effects model appears to more suitable since the Wald chi-squared statistic (3012.45) with a p-value of 0.0000 indicates that the model is also highly significant. For the fixed effects model, the F-statistic ($F(3, 3567) = 2273.74$) with a p-value of 0.0000 indicates that the overall model is highly significant. Moreover, In the fixed effects model, $\rho = 0.9789$, which indicates a high degree of correlation within groups (individuals). This suggests that there are individual-specific effects that need to be accounted for. Ignoring these effects in an OLS model (with or without robust standard errors) would lead to biased and inconsistent estimates.

The suitability of a fixed or random effects model depends on whether the regressors are correlated with the error term, which is addressed in the next question.

Questions 5. Compare the random effect and the fixed effect model, which one is better?

```r
# Hausman test
phtest(fe_model, re_model)
```

```
##
##  Hausman Test
##
## data:  lwage ~ exp + exp2 + wks + ed
## chisq = 6191.4, df = 3, p-value < 2.2e-16
## alternative hypothesis: one model is inconsistent
```

Based on the $\chi^2 = 6191.43$ and p-value $= 0.0000$, we can reject the null hypothesis that there is no correlation between the regressors and the error. This implies that,

$$\mathbb{E}[\epsilon_{it} \mid X_{i1}, X_{i2}, \ldots] \neq 0.$$

Questions 6. Export the above regression results to Excel, Word or Latex. (only need to output one).

Consider another model:

$$lwage_{it} = \alpha + \beta_1 exp_{it} + \beta_2 exp_{it}^2 + \beta_3 wks_{it} + \beta_4 ed_{it} + \beta_5 occ_{it} + \epsilon_{it}$$

The regression results are output to an html file.

```
ols_model <- lm(lwage ~ exp + exp2 + wks + ed, data = data)
fe_model <- plm(lwage ~ exp + exp2 + wks + ed, data = data,
                index = c("id", "t"), model = "within")
re_model <- plm(lwage ~ exp + exp2 + wks + ed, data = data,
                index = c("id", "t"), model = "random")

# Convert summaries to data frames
ols_df <- tidy(ols_model)
fe_df <- tidy(fe_model)
re_df <- tidy(re_model)

# invisible({capture.output({
# stargazer(ols_model, fe_model, re_model,
#             type = "text", out = "regression_results.text")
# })})

stargazer(ols_model, fe_model, re_model,
            type = "text", out = "regression_results.text")
```

```
##
## ===============================================================================
##                                        Dependent variable:
##                    ---------------------------------------------------------
##                                              lwage
##                              OLS                          panel
##                                                           linear
##                              (1)                (2)               (3)
## -------------------------------------------------------------------------------
## exp                       0.045***           0.114***          0.089***
##                            (0.002)            (0.002)           (0.003)
##
## exp2                      -0.001***          -0.0004***        -0.001***
##                           (0.0001)            (0.0001)          (0.0001)
##
## wks                       0.006***             0.001             0.001
##                            (0.001)            (0.001)           (0.001)
##
## ed                        0.076***                               0.112***
```

```
##                                (0.002)                                                (0.006)
##
## Constant                      4.908***                                               3.829***
##                                (0.067)                                                (0.094)
##
## ------------------------------------------------------------------------------------------------
## Observations                    4,165                        4,165                     4,165
## R2                              0.284                        0.657                     0.420
## Adjusted R2                     0.283                        0.599                     0.419
## Residual Std. Error     0.391 (df = 4160)
## F Statistic         411.623*** (df = 4; 4160) 2,273.736*** (df = 3; 3567) 3,012.453***
## ================================================================================================
## Note:                                                          *p<0.1; **p<0.05; ***p<0.01
```

```
# stargazer(ols_model, fe_model, re_model,
#             type = "latex", out = "regression_results.tex")
#
# stargazer(ols_model, fe_model, re_model,
#             type = "html", out = "regression_results.html")
```

Note that using the *capture.output()* function prevents *stargazer* from displaying the latex, html, or text output that it generates in the output file rendered by R Markdown, in this case a pdf.

Questions 7. Consider an endogenous variable $\beta_5 occ_{it}$, and $south_{it}$ and fem_{it} as instrumental variables.

Since wages likely influence occupation, there is a high probability that *occ* is endogenous and, therefore, correlated with the error term.

```
first_stage <- lm(occ ~ south + fem + exp + exp2 + wks + ed, data = data)
summary(first_stage)
```

```
##
## Call:
## lm(formula = occ ~ south + fem + exp + exp2 + wks + ed, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.16579 -0.22445 -0.02906  0.35421  0.95926
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.170e+00  6.790e-02  31.963  < 2e-16 ***
## south       -3.821e-02  1.343e-02  -2.845  0.00446 **
## fem         -1.449e-01  1.923e-02  -7.534    6e-14 ***
## exp         -6.156e-03  2.381e-03  -2.585  0.00978 **
## exp2         6.941e-05  5.250e-05   1.322  0.18623
## wks         -1.595e-03  1.181e-03  -1.350  0.17699
## ed          -1.145e-01  2.235e-03 -51.212  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.3886 on 4158 degrees of freedom
## Multiple R-squared:  0.3966, Adjusted R-squared:  0.3957
## F-statistic: 455.5 on 6 and 4158 DF,  p-value: < 2.2e-16
```

In the first-stage regression the endogenous variable *occ* is regressed on the instruments *south* and fem, along with any other exogenous variables. The coefficients of *south* and fem are relatively large compared to the coefficients for the other regressors and they statistically significant, which indicates that the instruments are correlated with the endogenous variable.

The F-statistic for the joint significance of the instruments (*south* and fem) and other exogenous variables is large and statistically significant. It is also worth noting that the R-squared value of the first-stage regression is 0.3966, which suggests that the instruments and other exogenous variables explain a substantial amount of the variation in *occ*.

Questions 8. Run 2SLS and GMM.

The first estimation results are from the *ivreg()* command from the R package *AER*. The results are very close to those for Stata's *ivregress* command. R's *summary()* command allows for heteroscedastic consistent standard errors to be calculated by adding the option *vcov = vcovHC(iv_model, type = "HC1")*. Heteroscedastic and autocorrelation consistent standard errors can be calculated using the *vcov = sandwich* option.

```
# 2SLS

iv_model <- ivreg(lwage ~ exp + exp2 + wks + ed + occ | south +
                  fem + exp + exp2 + wks + ed, data = data)
summary(iv_model)
```

```
##
## Call:
## ivreg(formula = lwage ~ exp + exp2 + wks + ed + occ | south +
##     fem + exp + exp2 + wks + ed, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.70018 -0.98165  0.02119  0.86680  3.80979
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.0395202  0.8221234  -1.264   0.2061
## exp          0.0599443  0.0077145   7.770 9.80e-15 ***
## exp2        -0.0008969  0.0001659  -5.406 6.81e-08 ***
## wks          0.0084283  0.0036934   2.282   0.0225 *
## ed           0.3995287  0.0437927   9.123  < 2e-16 ***
## occ          2.8571993  0.3819344   7.481 8.96e-14 ***
##
## Diagnostic tests:
##                   df1  df2 statistic  p-value
## Weak instruments    2 4158     33.51 3.66e-15 ***
## Wu-Hausman          1 4158    680.76  < 2e-16 ***
## Sargan              1   NA      0.04    0.841
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.215 on 4159 degrees of freedom
## Multiple R-Squared: -5.924,  Adjusted R-squared: -5.932
## Wald test: 45.26 on 5 and 4159 DF,  p-value: < 2.2e-16
```

```
summary(iv_model, vcov = vcovHC(iv_model, type = "HC1"))
```

```
##
## Call:
## ivreg(formula = lwage ~ exp + exp2 + wks + ed + occ | south +
##     fem + exp + exp2 + wks + ed, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.70018 -0.98165  0.02119  0.86680  3.80979
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.0395202  0.8598233  -1.209    0.227
## exp          0.0599443  0.0081855   7.323 2.89e-13 ***
## exp2        -0.0008969  0.0001808  -4.961 7.28e-07 ***
## wks          0.0084283  0.0033933   2.484    0.013 *
## ed           0.3995287  0.0457957   8.724  < 2e-16 ***
## occ          2.8571993  0.3993847   7.154 9.91e-13 ***
##
## Diagnostic tests:
##                   df1  df2 statistic  p-value
## Weak instruments    2 4158     33.51 3.66e-15 ***
## Wu-Hausman          1 4158    680.76  < 2e-16 ***
## Sargan              1   NA      0.04    0.841
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.215 on 4159 degrees of freedom
## Multiple R-Squared: -5.924,  Adjusted R-squared: -5.932
## Wald test: 56.61 on 5 and 4159 DF,  p-value: < 2.2e-16
```

```
summary(iv_model, vcov = sandwich)
```

```
##
## Call:
## ivreg(formula = lwage ~ exp + exp2 + wks + ed + occ | south +
##     fem + exp + exp2 + wks + ed, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.70018 -0.98165  0.02119  0.86680  3.80979
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.0395202  0.8592038  -1.210    0.226
## exp          0.0599443  0.0081796   7.328 2.78e-13 ***
## exp2        -0.0008969  0.0001807  -4.965 7.14e-07 ***
## wks          0.0084283  0.0033908   2.486    0.013 *
```

```
## ed           0.3995287  0.0457627   8.730  < 2e-16 ***
## occ          2.8571993  0.3990970   7.159 9.55e-13 ***
##
## Diagnostic tests:
##                   df1  df2 statistic  p-value
## Weak instruments    2 4158     31.93 1.73e-14 ***
## Wu-Hausman          1 4158    864.31  < 2e-16 ***
## Sargan              1   NA      0.04    0.841
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.215 on 4159 degrees of freedom
## Multiple R-Squared: -5.924,  Adjusted R-squared: -5.932
## Wald test: 56.69 on 5 and 4159 DF,  p-value: < 2.2e-16
```

```
# GMM
# Define the formula and instruments
formula <- lwage ~ exp + exp2 + wks + ed + occ
instruments <- ~ exp + exp2 + wks + ed + south + fem

# Fit the GMM model
gmm_model <- gmm(formula, x = instruments, data = data)
summary(gmm_model)
```

```
##
## Call:
## gmm(g = formula, x = instruments, data = data)
##
##
## Method:  twoStep
##
## Kernel:  Quadratic Spectral(with bw =  3.13717 )
##
## Coefficients:
##              Estimate     Std. Error   t value      Pr(>|t|)
## (Intercept)  -1.03576650   2.19656614  -0.47153895   0.63725591
## exp           0.05981487   0.01976800   3.02584308   0.00247941
## exp2         -0.00089328   0.00043604  -2.04861618   0.04049966
## wks           0.00840921   0.00463976   1.81242489   0.06992057
## ed            0.39941983   0.11702288   3.41317705   0.00064210
## occ           2.85605081   1.02294287   2.79199445   0.00523843
##
## J-Test: degrees of freedom is 1
##                 J-test     P-value
## Test E(g)=0:    0.0067649  0.9344488
##
## Initial values of the coefficients
##   (Intercept)           exp          exp2           wks            ed
## -1.0395202184  0.0599442884 -0.0008969382  0.0084282504  0.3995287062
##           occ
##  2.8571993178
```

The results for GMM and IVGMM are identical. The results for the 2SLS are nearly identical to the results from the GMM and IVGMM.

Questions 9. Test if occ_{it} is endogenous or not, and examine $south_{it}$, fem_{it} are valid instrumental variables.

The *ivreg()* command from the *ivreg* package (not to be confused with the *ivreg()* command from the *AER* package, which will not permit one to perform a Hausman or Sargan test) does not require a second step to perform either the Hausman or Sargan tests. You only need to use the standard *summary()* command.

```
iv_model <- ivreg(lwage ~ exp + exp2 + wks + ed + occ | south +
                  fem + exp + exp2 + wks + ed, data = data)

# Overidentification test
summary(iv_model)
```

```
##
## Call:
## ivreg(formula = lwage ~ exp + exp2 + wks + ed + occ | south +
##     fem + exp + exp2 + wks + ed, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.70018 -0.98165  0.02119  0.86680  3.80979
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.0395202  0.8221234  -1.264   0.2061
## exp          0.0599443  0.0077145   7.770 9.80e-15 ***
## exp2        -0.0008969  0.0001659  -5.406 6.81e-08 ***
## wks          0.0084283  0.0036934   2.282   0.0225 *
## ed           0.3995287  0.0437927   9.123  < 2e-16 ***
## occ          2.8571993  0.3819344   7.481 8.96e-14 ***
##
## Diagnostic tests:
##                   df1  df2 statistic  p-value
## Weak instruments    2 4158     33.51 3.66e-15 ***
## Wu-Hausman          1 4158    680.76  < 2e-16 ***
## Sargan              1   NA      0.04    0.841
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.215 on 4159 degrees of freedom
## Multiple R-Squared: -5.924,  Adjusted R-squared: -5.932
## Wald test: 45.26 on 5 and 4159 DF,  p-value: < 2.2e-16
```

The small p-value for the Wu-Hausman test indicates that the OLS suffers from endogeneity due to at least one regressor. However, the Sargan test indicates that the choice of the instrumental variables do not cause the model to be overidentified. You can find more information about the *ivreg()* command here, https://cran.r-project.org/web/packages/ivreg/vignettes/Diagnostics-for-2SLS-Regression.html

Questions 10. Store OLS, 2SLS and GMM regression results in R.

```
stargazer(ols_model, iv_model, gmm_model, type = "text",
          keep.stat = c("n", "rsq"), star.cutoffs = c(0.05, 0.01, 0.001))
```

```
##
## ===============================================
##                       Dependent variable:
##                   ----------------------------
##                          lwage              NA
##                     OLS    instrumental   GMM
##                              variable
##                     (1)        (2)         (3)
## -----------------------------------------------
## exp             0.045***    0.060***    0.060**
##                  (0.002)    (0.008)     (0.020)
##
## exp2            -0.001***  -0.001***    -0.001*
##                 (0.0001)    (0.0002)   (0.0004)
##
## wks             0.006***     0.008*      0.008
##                  (0.001)    (0.004)     (0.005)
##
## ed              0.076***    0.400***    0.399***
##                  (0.002)    (0.044)     (0.117)
##
## occ                        2.857***    2.856**
##                            (0.382)     (1.023)
##
## Constant        4.908***     -1.040      -1.036
##                  (0.067)    (0.822)     (2.197)
##
## -----------------------------------------------
## Observations     4,165       4,165       4,165
## R2               0.284       -5.924
## ===============================================
## Note:             *p<0.05; **p<0.01; ***p<0.001
```

The *stargazer()* command stores the results in a nicely formatted text file.