

1. We will investigate the following claim (which appears in nearly this form in Studenmund):

If an equation that contains a lagged dependent variable as a regressor has a serially correlated error term, then OLS estimates of the coefficients will be inconsistent.

To investigate, we define a regression model in a slightly different way. Let Y_t be a strictly stationary and ergodic random variable that is observed into the infinite past. Assume that $\mathbb{E}[Y_t] = 0$ and $\mathbb{E}[Y_t^2] < \infty$. Define

$$\beta = \frac{\text{Cov}(Y_{t-1}, Y_t)}{\text{Var}(Y_t)}$$

and $U_t = Y_t - \beta Y_{t-1}$ so that we can write

$$Y_t = \beta Y_{t-1} + U_t \quad \text{for } t = 1, \dots, n.$$

- (a) Show that $\mathbb{E}[U_t] = 0$:

$$\begin{aligned} \mathbb{E}[U_t] &= \mathbb{E}[Y_t - \beta Y_{t-1}] && \text{(substitute)} \\ &= \mathbb{E}[Y_t] - \beta \mathbb{E}[Y_{t-1}] && \text{(linearity of } \mathbb{E}) \\ \text{But } \mathbb{E}[Y_t] &= \mathbb{E}[Y_{t-1}] = 0 && \text{(stationarity of } Y) \\ \Rightarrow \mathbb{E}[U_t] &= 0 && \blacksquare \end{aligned}$$

Show that $\text{Cov}(Y_{t-1}, U_t) = 0$:

$$\begin{aligned} \text{Cov}(Y_{t-1}, U_t) &= \mathbb{E}[Y_{t-1}U_t] - \mathbb{E}[Y_{t-1}]\mathbb{E}[U_t] && \text{(def of Cov)} \\ &= \mathbb{E}[Y_{t-1}U_t] - 0 \times 0 && (\mathbb{E}[U_t] = \mathbb{E}[Y_{t-1}] = 0) \\ &= \mathbb{E}[Y_{t-1}(Y_t - \beta Y_{t-1})] && \text{(subst for } U_t) \\ &= \mathbb{E}[Y_{t-1}Y_t] - \frac{\text{Cov}(Y_{t-1}, Y_t)}{\text{Var}(Y_t)} \mathbb{E}[(Y_{t-1})^2] && \text{(rearrange, subs for } \beta) \\ &= \mathbb{E}[Y_{t-1}Y_t] - \frac{\mathbb{E}[Y_{t-1}Y_t] - \mathbb{E}[Y_{t-1}]\mathbb{E}[Y_t]}{\mathbb{E}[Y_t^2] - \mathbb{E}[Y_t]^2} \mathbb{E}[Y_{t-1}^2] \\ &= \mathbb{E}[Y_{t-1}Y_t] - \frac{\mathbb{E}[Y_{t-1}Y_t]}{\mathbb{E}[Y_t^2]} \mathbb{E}[Y_{t-1}^2] && (\mathbb{E}[Y_t] = 0) \end{aligned}$$

Note: $\beta = \frac{\text{Cov}(Y_{t-1}, Y_t)}{\text{Var}(Y_t)}$ is constant, and $\text{Cov}(Y_{t-j}, Y_t)$ does not depend on t , only j , since strict stationarity implies covariance stationarity; therefore $\text{Cov}(Y_{t-1}, Y_t)$ is constant and indep of t . So:

$$\begin{aligned} \Rightarrow \text{Var}(Y_t) &= \frac{\beta}{\text{Cov}(Y_{t-1}, Y_t)} \text{ is constant, indep of } t \\ \Rightarrow \text{Var}(Y_t) &= \text{Var}(Y_{t-1}) \\ \Rightarrow \mathbb{E}[Y_t^2] &= \mathbb{E}[Y_{t-1}^2] && (\mathbb{E}[Y_t] = \mathbb{E}[Y_{t-1}] = 0) \\ \Rightarrow \text{Cov}(Y_{t-1}, U_t) &= \mathbb{E}[Y_{t-1}Y_t] - \frac{\mathbb{E}[Y_{t-1}Y_t]}{\mathbb{E}[Y_t^2]} \mathbb{E}[Y_t^2] \\ &= \mathbb{E}[Y_{t-1}Y_t] - \mathbb{E}[Y_{t-1}Y_t] = 0 && \blacksquare \end{aligned}$$

- (b) Show that the OLS estimator B from a regression of Y_t on Y_{t-1} , for $t = 1, \dots, n$, is consistent for β .

Using $x = Y_{t-1}$ and law of large numbers:

$$\begin{aligned}
B &= \left(\sum_{t=1}^n (Y_{t-1} Y'_{t-1}) \right)^{-1} \sum_{t=1}^n (Y_{t-1} Y_t) \\
&= \beta + \left(\sum_{t=1}^n (Y_{t-1} Y'_{t-1}) \right)^{-1} \sum_{t=1}^n (Y_{t-1} U_t) && \text{(subst } Y_t = \beta Y_{t-1} + U_t) \\
&= \beta + \left(\frac{1}{n} \sum_{t=1}^n (Y_{t-1} Y'_{t-1}) \right)^{-1} \frac{1}{n} \sum_{t=1}^n (Y_{t-1} U_t) && \text{(mult by } \frac{1/n}{1/n}) \\
\frac{1}{n} \sum_{t=1}^n (Y_{t-1} Y'_{t-1}) &= \frac{1}{n} \sum_{t=1}^n (Y_{t-1}^2) \xrightarrow{p} \mathbb{E}[Y_{t-1}^2] = \mathbb{E}[Y_t^2] = \text{Var}(Y_t) && \text{(from part a)} \\
\frac{1}{n} \sum_{t=1}^n (Y_{t-1} U_t) &\xrightarrow{p} 0 && \text{(from part a)} \\
\Rightarrow B &\xrightarrow{p} \beta + \text{Var}(Y_t)^{-1} \times 0 = \beta && \text{(B is consistent for } \beta) \blacksquare
\end{aligned}$$

- (c) Show that, without further assumptions, U_t is serially correlated.

Serially correlated means $\text{Cov}(U_t, U_{t-1}) \neq 0$.

$$\begin{aligned}
Y_{t-1} &= \beta Y_{t-2} + U_{t-1} \\
\text{Cov}(U_t, U_{t-1}) &= \mathbb{E}[U_t \times (Y_{t-1} - \beta Y_{t-2})] \\
&= \mathbb{E}[U_t Y_{t-1}] - \beta \mathbb{E}[U_t Y_{t-2}] \\
&= 0 - \beta \mathbb{E}[U_t Y_{t-2}] && (\mathbb{E}[U_t Y_{t-1}] = 0 \text{ from part a}) \\
&= -\beta \mathbb{E}[(Y_t - \beta Y_{t-1}) Y_{t-2}] && \text{(subst for } U_t) \\
&= -\beta \mathbb{E}[Y_t Y_{t-2}] + \beta^2 \mathbb{E}[Y_{t-1} Y_{t-2}]
\end{aligned}$$

Unless $\beta = 0$, and/or the lag 1 and lag 2 correlations of Y are both zero ($\text{Cov}(Y_t, Y_{t-1}) = \text{Cov}(Y_t, Y_{t-2}) = 0$), then U_t is serially correlated.

- (d) What do you conclude about the validity of the quotation?

If an equation that contains a lagged dependent variable as a regressor has a serially correlated error term, then OLS estimates of the coefficients will be inconsistent.

In our analysis in part c, we showed B is consistent for β , though we didn't need to assume $\beta = 0$ nor that the Y terms are uncorrelated. So I believe this shows the quotation must be incorrect.

2. Stochastic Processes - Consider two sequences formed from the i.i.d. mean zero finite variance sequence $\{\varepsilon_t\}_{t \geq 0}$. The first is $\{g_t\}_{t \geq 1}$ formed as

$$g_t = t\varepsilon_t,$$

the second is $\{h_t\}_{t \geq 1}$ formed as

$$h_t = \varepsilon_t \varepsilon_{t-1}.$$

- (a) Determine if each sequence is independent and identically distributed (i.i.d.).

For $\{g_t\}_{t \geq 1}$:

$$\mathbb{E}[g_t] = \mathbb{E}[t\varepsilon_t] = t\mathbb{E}[\varepsilon_t] = 0$$

and

$$\text{Var}(g_t) = \mathbb{E}[g_t^2] - \mathbb{E}[g_t]^2 = \mathbb{E}[t^2 \varepsilon_t^2] = t^2 \mathbb{E}[\varepsilon_t^2] = 0$$

Since mean = 0, but variance is dependent on t , $\{g_t\}_{t \geq 1}$ is not iid.

For $\{h_t\}_{t \geq 1}$:

$$\mathbb{E}[h_t] = \mathbb{E}[\varepsilon_t \varepsilon_{t-1}] = \mathbb{E}[\varepsilon_t] \mathbb{E}[\varepsilon_{t-1}] = 0$$

and

$$\text{Var}(h_t) = \mathbb{E}[h_t^2] - \mathbb{E}[h_t]^2 = \mathbb{E}[\varepsilon_t^2 \varepsilon_{t-1}^2]$$

Mean and variance are independent of t ; but since each element in the sequence shares a term with the prior element (i.e. $\varepsilon_1 \varepsilon_2, \varepsilon_2 \varepsilon_3, \varepsilon_3 \varepsilon_4, \dots$) the terms are not independent.

- (b) Determine if each sequence is a stationary martingale difference sequence (m.d.s.).

A vector sequence is a martingale difference sequence if $\mathbb{E}[U_t | U_{t-1}, \dots, U_1] = 0$, i.e. the expected value with respect to the past is zero.

For $\{g_t\}_{t \geq 1}$: expected value is independent of t , so it is an m.d.s. However, as shown in part a, the variance changes with t so it is not stationary.

For $\{h_t\}_{t \geq 1}$: expected value is also independent of t , so it is also an m.d.s. In this case, the variance is not dependent on t , so the distribution is constant over time, so this is a stationary m.d.s.

- (c) Determine if each sequence is white noise.

White noise is covariance stationary, with $\mathbb{E}[Z_t] = 0$ and $\text{Cov}(Z_t, Z_{t-j}) = 0$ for all $j \neq 0$ (serially uncorrelated).

For $\{g_t\}_{t \geq 1}$: we've already seen $\mathbb{E}[g_t] = 0$.

$$\text{Cov}(g_t, g_{t-j}) = \mathbb{E}[g_t g_{t-j}] - \mathbb{E}[g_t] \mathbb{E}[g_{t-j}] = t(t-j) \mathbb{E}[\varepsilon_t \varepsilon_{t-j}] = t(t-j) \mathbb{E}[\varepsilon_t] \mathbb{E}[\varepsilon_{t-j}] = 0$$

We've seen that the variance is time-dependent (from part a), but I think the sequence can still be *covariance* stationary? In which case g_t would be white noise.

For $\{h_t\}_{t \geq 1}$: $\mathbb{E}[h_t] = 0$, and variance is constant with time. Is it covariance stationary?

$$\text{Cov}(h_t, h_{t-j}) = \mathbb{E}[h_t h_{t-j}] - \mathbb{E}[h_t] \mathbb{E}[h_{t-j}] = \mathbb{E}[\varepsilon_t \varepsilon_{t-1} \times \varepsilon_{t-j} \varepsilon_{t-j-1}]$$

Since ε_t are iid and $\mathbb{E}[\varepsilon_t] = 0$, then

$$\text{Cov}(h_t, h_{t-j}) = \mathbb{E}[\varepsilon_t] \mathbb{E}[\varepsilon_{t-1}] \mathbb{E}[\varepsilon_{t-j}] \mathbb{E}[\varepsilon_{t-j-1}] = 0$$

Therefore, h_t is definitely white noise!

3. Martingale Sequences

You join a research project designed to predict stock prices. One member of your team, using previous research proposes a regression analysis of the form:

$$p_t = x_t' \beta + u_t$$

where p_t measures the stock price for period t and the regressors include lagged values of both firm characteristics, such as research and development expenditures, and the stock price. Let $w_t = \{x_t, x_{t-1}, \dots\}$ be the information set that consists of the current and all past values of the regressors.

- (a) In one discussion, you state that the stock price, p_t , is a martingale with respect to w_t . Your colleagues ask you not only to define what it means for the price to be a martingale, but also to explain what implications your martingale statement has for the proposed regression analysis. Please respond.

A martingale is a process in which the expected value now is the same as the observed value in the previous time period; in other words, the conditional expected value now, given the sequence of observations prior to now, is just the most recent observation.

For this case, $EE[p_t|w_t] = p_{t-1}$. If p_t is a martingale with respect to w_t , then in our regression, the only thing we need to include in our $w_t = \{x_t, x_{t-1}, \dots\}$ is the price in the last period. Our regression equation would simply be:

$$p_t = \beta p_{t-1} + u_t, \quad \text{where } \beta = 1$$

- (b) The second stage of the analysis is to predict future prices and, hence future price changes, given w_t . What prediction of future prices results from the martingale model? What is the corresponding prediction for future price changes? (Hint, use the Law of Iterated Expectations to determine $E[p_{t+1}|w_t]$ to begin.)

$$\begin{aligned} \text{Note: } w_{t+1} &= \{x_{t+1}, x_t, x_{t-1}, \dots\} = \{x_{t+1}, w_t\} \\ \Rightarrow E[p_{t+1}|w_{t+1}] &= E[p_{t+1}|x_{t+1}, w_t] = p_t && (p_{t+1} \text{ is martingale WRT } w_{t+1}) \\ &= E[E[p_{t+1}|x_{t+1}, w_t]|w_t] = E[p_t|w_t] && (\text{law of iter. exp.}) \\ &= p_{t-1} && (p_t \text{ is martingale WRT } w_t) \end{aligned}$$

This can be used recursively to predict future prices; for any future price p_{t+j} , the expected value $E[p_{t+j}|w_{t+j}] = p_{t-1}$, the most recent observed price.

This also implies that changes in price from one period to the next,

$$E[p_{t+j} - p_{t+j-1}|w_{t+j}] = 0,$$

so the sequence can also be represented as a martingale difference sequence.

- (c) Upon hearing of your results, another researcher remarks that he understood that stock prices follow a random walk. He states that the results you found confirm his understanding, as the statement that stock prices are a martingale (with respect to w_t) implies that stock prices are a random walk. The other researchers turn to you to assess the accuracy of this statement. Please respond.

A random walk is a sequence of cumulative sums, with U_t independent white noise (iid, $E[U_t] = 0$, $Cov(U_t, U_{t-j}) = 0 \forall j \neq 0$):

$$p_t = p_{t-1} + U_t, p_{t-1} = p_{t-2} + U_{t-1}, \dots$$

I think in our price example, the U_t terms are not necessarily independent white noise.

random walk \Rightarrow martingale sequence

but

martingale sequence \nRightarrow random walk

So while our price example *might* be a true random walk, it is not *necessarily* a random walk.

4. Computational Exercise

Return to the paper by Charness and Kuhn listed on the syllabus. Write programs in both Matlab and Stata (the results from each program should match) that again estimate the models in columns (6) and (7) of Table 3 of Charness and Kuhn. Calculate cluster-robust standard errors (so your estimated standard errors should match those in the table). Test the hypothesis that all four slope coefficients are equal and provide the p -value for the estimated test statistic.

Stata

My output matches that of the Charness and Kuhn (2007) Table 3, with the exception of the error for `rel_w1_low` (I got .0274, the paper reports .026).

The code is essentially identical to exercise 4, with the exception of adding an option of `reg ..., robust cluster(wrk1id)` or `(wrk2id)`.

Table 1: A: Type 1 workers

	(1)	(2)
	e1	e1
<code>_Iw1_1</code>	.499 (.08262)	.42 (.1038)
<code>_Iw1_2</code>	1.278 (.1718)	1.119 (.1715)
<code>_Iw1_3</code>	1.544 (.4949)	1.308 (.442)
<code>rel_w1_low</code>	-.02431 (.02738)	.01501 (.04194)
N	555	555
R^2	0.295	0.604

Table 2: B: Type 2 workers

	(1)	(2)
	e2	e2
<code>_Iw2_1</code>	.3467 (.1129)	.2594 (.1457)
<code>_Iw2_2</code>	1.045 (.1337)	.9421 (.1626)
<code>_Iw2_3</code>	1.718 (.204)	1.5 (.2103)
<code>_Iw2_4</code>	1.488 (.3375)	1.575 (.2986)
<code>rel_w2_low</code>	.05617 (.05982)	.009565 (.08853)
N	555	555
R^2	0.318	0.654

Results of Stata hypothesis tests

The Wald test statistics are smaller and the p-values higher (though not apparent in the Stata output) than last week's assignment, which is to be expected since our cluster robust standard errors are higher than the unclustered standard errors we calculated last week.

Column 6a:

```
. test _Iw1_1 = _Iw1_2 = _Iw1_3
( 1)  _Iw1_1 - _Iw1_2 = 0
( 2)  _Iw1_1 - _Iw1_3 = 0

      F( 2,    36) =    14.65
      Prob > F =    0.0000
```

Column 7a:

```
. test _Iw1_1 = _Iw1_2 = _Iw1_3
( 1)  _Iw1_1 - _Iw1_2 = 0
( 2)  _Iw1_1 - _Iw1_3 = 0

      F( 2,    36) =    22.77
      Prob > F =    0.0000
```

Column 6b:

```
. test _Iw2_1 = _Iw2_2 = _Iw2_3 = _Iw2_4
( 1)  _Iw2_1 - _Iw2_2 = 0
( 2)  _Iw2_1 - _Iw2_3 = 0
( 3)  _Iw2_1 - _Iw2_4 = 0

      F( 3,    36) =    19.52
      Prob > F =    0.0000
```

Column 7b:

```
. test _Iw2_1 = _Iw2_2 = _Iw2_3 = _Iw2_4
( 1)  _Iw2_1 - _Iw2_2 = 0
( 2)  _Iw2_1 - _Iw2_3 = 0
( 3)  _Iw2_1 - _Iw2_4 = 0

      F( 3,    36) =    14.96
      Prob > F =    0.0000
```

Matlab output

All Matlab output matches Stata output. I found and fixed the error from last week's assignment that caused the Wald test statistics to be slightly off from the Stata results.

Matrix calculated coefficients for table 3, column 6a

0.4990	1.2778	1.5445	-0.0243
--------	--------	--------	---------

Cluster robust standard errors for table 3, column 6a

0.0826	0.1718	0.4949	0.0274
--------	--------	--------	--------

Wald test and p-value, col 6a:

14.6530

6.4388e-07

Matrix calculated coefficients for table 3, column 7a

0.4200	1.1188	1.3080	0.0150
--------	--------	--------	--------

Cluster robust standard errors for table 3, column 7a

0.1038	0.1715	0.4420	0.0419
--------	--------	--------	--------

Wald test and p-value, col 7a:

22.7717

3.5212e-10

Matrix calculated coefficients for table 3, column 6b

0.3467	1.0450	1.7184	1.4877	0.0562
--------	--------	--------	--------	--------

Cluster robust standard errors for table 3, column 6b

0.1129	0.1337	0.2040	0.3375	0.0598
--------	--------	--------	--------	--------

Wald test and p-value, col 6b:

19.5207

5.2512e-12

Matrix calculated coefficients for table 3, column 7b

0.2594	0.9421	1.5005	1.5753	0.0096
--------	--------	--------	--------	--------

Cluster robust standard errors for table 3, column 7b

0.1457	0.1626	0.2103	0.2986	0.0885
--------	--------	--------	--------	--------

Wald test and p-value, col 7b:

14.9627

2.4919e-09

Stata code

```
set more off
clear

// Set working directory

cd "~/github/econ_courses/econ241b/assts/asst5"

clear all
use "prob_set_5.dta"

// drop non-public wages
drop if pubwage != 1

// drop unused columns
keep wrk1id wrk2id period e1 e2 w1 w2

////////////////////
// Part A: Type 1 Workers (low productivity)
////////////////////

// create relative wage variable
gen rel_w1 = w1 - w2

// create indicator for wage 1 <= wage 2
gen w1_low = (w1 <= w2)

// calculate rel wage * dummy
gen rel_w1_low = rel_w1 * w1_low

// regress variables for public wages for low productivity workers (1).
// Col 6: effort vs wage and relative wage (< 0, asymmetric model)
//   with fixed effects of period
// Col 7: effort vs wage and relative wage (< 0, asymmetric model)
//   with fixed effects of worker and period
eststo clear

xi: reg e1 i.w1 rel_w1_low i.period, robust cluster(wrk1id)
eststo col_a6

test _lw1_1 = _lw1_2 = _lw1_3

xi: reg e1 i.w1 rel_w1_low i.period i.wrk1id, robust cluster(wrk1id)
eststo col_a7
test _lw1_1 = _lw1_2 = _lw1_3

esttab col_a6 col_a7 using table3a67.tex, ///
    drop(*wrk1id* *period* *cons*) title(A: Type 1 workers) noconst b(%10.4g) ///
    se r2 nostar replace
```



```

////////////////////////////////////
// Part B: Type 2 Workers (high productivity)
////////////////////////////////////

// create relative wage variable
gen rel_w2 = w2 - w1

// create indicator for wage 1 <= wage 2
gen w2_low = (w2 <= w1)

// calculate rel wage * dummy
gen rel_w2_low = rel_w2 * w2_low

// regress variables for public wages, for high productivity workers (2).
// Col 6: effort vs wage and relative wage (< 0, asymmetric model)
//   with fixed effects of period
// Col 7: effort vs wage and relative wage (< 0, asymmetric model)
//   with fixed effects of worker and period
eststo clear

xi: reg e2 i.w2 rel_w2_low i.period, robust cluster(wrk2id)
eststo col_b6
test _lw2_1 = _lw2_2 = _lw2_3 = _lw2_4

xi: reg e2 i.w2 rel_w2_low i.period i.wrk2id, robust cluster(wrk2id)
eststo col_b7
test _lw2_1 = _lw2_2 = _lw2_3 = _lw2_4

esttab col_b6 col_b7 using table3b67.tex, ///
    drop(*wrk2id* *period* *cons*) title(B: Type 2 workers) noconst b(%10.4g) ///
    se r2 nostar replace

```

Matlab code

Main script

```
%% Import data from text file.
clear all

filename = '~/github/econ_courses/econ241b/assts/asst5/prob_set5.csv';

raw = readtable(filename);
% NOTE: this is the same data as for asst 3; I did not save the Stata
% edits made for assignment 4.

%% set up data
% keep just public wages and drop the pubwage column
wage = raw(raw.pubwage == 1, :);
wage.pubwage = [];

% add in w11-w14, w21-w24 columns
wage.w11 = wage.w1 == 1;
wage.w12 = wage.w1 == 2;
wage.w13 = wage.w1 == 3;
wage.w14 = wage.w1 == 4;
wage.w21 = wage.w2 == 1;
wage.w22 = wage.w2 == 2;
wage.w23 = wage.w2 == 3;
wage.w24 = wage.w2 == 4;

nrows = length(wage.w1);
const = ones(nrows, 1);

%% TYPE 1 WORKER
wage.rel_w1 = wage.w1 - wage.w2;
wage.rel_w1_low = wage.rel_w1 .* (wage.rel_w1 < 0);

%% manual regression with period fixed effects and clustering

% Dummy variable expansion for period variable, then cut one column to
% avoid singular matrix at the end
d_period = dummyvar(wage.period);
d_period(:, 1) = [];

% Set up X matrix including constant, regressors, and fixed effects matrix
X6a = [wage.w11 wage.w12 wage.w13 wage.rel_w1_low d_period const];
y = wage.e1;

disp('Matrix calculated coefficients for table 3, column 6a')
B6a = (X6a' * X6a)^-1 * (X6a' * y);
disp(B6a(1:4')) %%% these are the coefficients on wage 1, 2, 3, and difference
% 0.4990 1.2778 1.5445 -0.0243

% Calc residuals
```

```

e_i6a = y - X6a * B6a;

%% Calc covariance matrix with clustering

V_b6a = calc_clust_V (X6a, B6a, e_i6a, wage.wrk1id);

se_b6a = sqrt(diag(V_b6a));
disp('Cluster robust standard errors for table 3, column 6a')
disp(se_b6a(1:4))

%% Use Wald test to check whether coefficients are equal

[wald_6a, p_6a] = calc_wald(X6a, V_b6a, B6a, e_i6a, 2);
disp('Wald test and p-value, col 6a:')
disp(wald_6a); disp(p_6a)

%% Manual regression: fixed effects on period and worker ID

% Dummy variable expansion for worker 1 variable, then cut two columns to
% avoid singular matrix at the end.
d_wrk1id = dummyvar(wage.wrk1id);
sum_test = sum(d_wrk1id) ~= 0;
d_wrk1id = d_wrk1id(:, sum_test);
d_wrk1id(:, 37) = [];
d_wrk1id(:, 1) = [];
%%
% Set up X matrix including constant, regressors, and fixed effects matrix
X7a = [wage.w11 wage.w12 wage.w13 wage.rel_w1_low d_period d_wrk1id const];
y = wage.e1;

disp('Matrix calculated coefficients for table 3, column 7a')
B7a = (X7a' * X7a)^-1 * (X7a' * y);
disp(B7a(1:4)) %%% these are the coefficients on wage 1, 2, 3, and difference
% 0.4990    1.2778    1.5445   -0.0243

% Calc residuals
e_i7a = y - X7a * B7a;

%% Calc covariance matrix with clustering

V_b7a = calc_clust_V (X7a, B7a, e_i7a, wage.wrk1id);

se_b7a = sqrt(diag(V_b7a));
disp('Cluster robust standard errors for table 3, column 7a')
disp(se_b7a(1:4))

%% Use Wald test to check whether coefficients are equal

[wald_7a, p_7a] = calc_wald(X7a, V_b7a, B7a, e_i7a, 2);
disp('Wald test and p-value, col 7a:')
disp(wald_7a); disp(p_7a)

```

```

%% TYPE 2 WORKER
% Effort vs wages and rel wages, type 1 worker
% set up X2 to be a constant plus own wage regressor plus relative wage
wage.rel_w2 = wage.w2 - wage.w1;
wage.rel_w2_low = wage.rel_w2 .* (wage.rel_w2 < 0);

%% manual regression

% Set up X matrix including constant, regressors, and fixed effects matrix
X6b = [wage.w21 wage.w22 wage.w23 wage.w24 wage.rel_w2_low d_period const];
y = wage.e2;

disp('Matrix calculated coefficients for table 3, column 6b')
B6b = (X6b' * X6b)^-1 * (X6b' * y);
disp(B6b(1:5)) %% these are the coefficients on wage 1, 2, 3, and difference
% 0.3467 1.0450 1.7184 1.4877 0.0562

% Calc residuals
e_i6b = y - X6b * B6b;

%% Calc covariance matrix with clustering

V_b6b = calc_clust_V (X6b, B6b, e_i6b, wage.wr2id);

se_b6b = sqrt(diag(V_b6b));
disp('Cluster robust standard errors for table 3, column 6b')
disp(se_b6b(1:5))

%% Use Wald test to check whether coefficients are equal

[wald_6b, p_6b] = calc_wald(X6b, V_b6b, B6b, e_i6b, 3);
disp('Wald test and p-value, col 6b:')
disp(wald_6b); disp(p_6b)

%% Manual regression: fixed effects on period and worker ID

% Dummy variable expansion for worker 2 variable, then cut two columns to
% avoid singular matrix at the end.
d_wrk2id = dummyvar(wage.wr2id);
sum_test = sum(d_wrk2id) ~= 0;
d_wrk2id = d_wrk2id(:, sum_test);
d_wrk2id(:, 37) = [];
d_wrk2id(:, 1) = [];
%
% Set up X matrix including constant, regressors, and fixed effects matrix
X7b = [wage.w21 wage.w22 wage.w23 wage.w24 wage.rel_w2_low d_period d_wrk2id const];
y = wage.e2;

disp('Matrix calculated coefficients for table 3, column 7b')
B7b = (X7b' * X7b)^-1 * (X7b' * y);

```

```

disp(B7b(1:5)') %% these are the coefficients on wage 1, 2, 3, and difference
%    0.2594    0.9421    1.5005    1.5753    0.0096

% Calc residuals
e_i7b = y - X7b * B7b;

%% Calc covariance matrix with clustering

V_b7b = calc_clust_V (X7b, B7b, e_i7b, wage.wrk2id);

se_b7b = sqrt(diag(V_b7b));
disp('Cluster robust standard errors for table 3, column 7b')
disp(se_b7b(1:5)')

%% Use Wald test to check whether coefficients are equal

[wald_7b, p_7b] = calc_wald(X7b, V_b7b, B7b, e_i7b, 3);
disp('Wald test and p-value, col 7b:')
disp(wald_7b); disp(p_7b)

```

Matlab functions

```

function [V_b] = calc_clust_V (X, B, e_i, clust_vec)
%% Takes arguments of X, Beta, residual vector e_i, and a clustering
% vector, returns the covariance matrix V_hat_beta.

% X = X6a; B = b6a; clust_vec = wage.wrk1id;

    clust_index = unique(clust_vec);

    n_obs = length(clust_vec);
    k_regr = length(B);
    G_gps = length(clust_index);

    omega_sum = zeros(k_regr, k_regr);
    for i = 1:length(clust_index)
        clust_id = clust_index(i);

        X_gp = X(clust_vec == clust_id, :);
        e_gp = e_i(clust_vec == clust_id);
        omega_gp = X_gp' * e_gp * e_gp' * X_gp;

        omega_sum = omega_sum + omega_gp;
    end

    % calc the adjustment factor
    a_n = (n_obs - 1)/(n_obs - k_regr) * G_gps / (G_gps - 1);

    V_b = a_n * inv(X' * X) * omega_sum * (X' * X)^(-1);

end

```

```

function [wald, p_val] = calc_wald(X, V, B, e_i, q)
% given regressor matrix X, covariance matrix V, coefficient vector B,
% error terms vector e_i, and number of coefficients to test q (assumes
% testing B(1:q) against one another to see if they are identical), returns
% Wald test value and p-value.

R = [ones(q, 1) -eye(q) zeros(q, length(B) - q - 1)];
r = zeros(q, 1);

[n, k] = size(X);

ssr = sum(e_i.^2);
dof = n - k;
s2 = ssr / dof;

wald = (R * B - r)' * (R * V * R')^-1 * (R * B - r) / q;

p_val = 1 - fcdof(wald, q, n - k);

end

```