Econ241b: PS 3

Casey O'Hara

2018-02-19

1. (2017 Prelim) You are asked to determine how the conditional mean of (log) wage, denoted y, depends on the sex of the individual. Let m be an indicator for male (that is, m takes the value 1 if the individual is male) and let f be the indicator for female (that is, f takes the value 1 if the individual is female). For a sample of n individuals, indexed by i, with n_1 males, consider OLS regression of

$$y_i = \beta_1 m_i + \beta_2 f_i + u_i$$

(a) Is $\beta_1 m_i + \beta_2 f_i$ an exact, or approximate, expression for the mean of the male and female wage distributions?

This is an exact expression for the mean of the male and female wage distributions.

First, note that (presumably) each observation i will have either $m_i = 1, f_i = 0$ or $m_i = 0, f_i = 1$. Therefore, in $\beta_1 m_i + \beta_2 f_i$, β_1 is the mean for males, β_2 is the mean for females. So the equation essentially equates to $\mathbb{E}[y|m, f]$, the mean (log) wage conditional on gender (m and f).

This is similar to section 2.17 in Hansen.

(b) Show that $\widehat{\beta}_1$ is the sample mean of male wages.

$$\begin{aligned}
SSE_n(\boldsymbol{\beta}) &= \sum_{i=1}^{n}(y_i - (\beta_1 m_i + \beta_2 f_i))^2 && \text{(def of SSE)}\\
&= \sum_{i=1}^{n}(y_i^2 - 2y_i(m_i\beta_1 + f_i\beta_2) + m_i^2\beta_1^2 + 2m_i f_i\beta_1\beta_2 + f_i^2\beta_2^2) && \text{(expand)}\\
&= \sum_{i=1}^{n}(y_i^2 - 2y_i(m_i\beta_1 + f_i\beta_2) + m_i^2\beta_1^2 + 0 + f_i^2\beta_2^2) && (m_i \times f_i = 0)\\
&= \sum_{i=1}^{n} y_i^2 - 2\beta_1\sum_{i=1}^{n} y_i m_i - 2\beta_2\sum_{i=1}^{n} y_i f_i + \beta_1^2\sum_{i=1}^{n} m_i^2 + \beta_2^2\sum_{i=1}^{n} f_i^2 && \text{(rearrange)}\\
\frac{\partial SSE_n(\widehat{\boldsymbol{\beta}})}{\partial\widehat{\beta}_1} &= 0 - 2\sum_{i=1}^{n} y_i m_i - 0 + 2\widehat{\beta}_1\sum_{i=1}^{n} m_i^2 + 0 = 0 && \text{(FOC)}\\
\Rightarrow \widehat{\beta}_1 &= \frac{\sum_{i=1}^{n} y_i m_i}{\sum_{i=1}^{n} m_i^2} && \text{(rearrange)}\\
&= \frac{\sum_{j=1}^{n_1} y_j}{\sum_{j=1}^{n_1} 1^2} && (j \in \{i|m_i = 1\}, \text{ ie. male})\\
&= \frac{1}{n_1}\sum_{j=1}^{n_1} y_j = \bar{y}_{male} && \blacksquare
\end{aligned}$$

(note second order condition $\frac{\partial^2 SSE_n(\vec{\beta})}{\partial\beta_1^2} = 2\sum_{i=1}^{n} m_i^2 > 0$ so $\widehat{\beta}_1$ minimizes SSE)

(c) Consider the continuous covariates X (an $n \times k$ matrix) along with y (an $n \times 1$ vector). Describe, in words, the transformations

$$y^* = y - \widehat{\beta}_1 m - \widehat{\beta}_2 f$$
$$X^* = X - m\bar{x}_1^T - f\bar{x}_2^T$$

where x_1 and x_2 are the $k \times 1$ means of the covariates for men and women, respectively.

- $\widehat{\beta}_1 m - \widehat{\beta}_2 f$ is the fitted value to estimate conditional mean of y on m, f. Since y^* subtracts the fitted values $\widehat{y}$ from the observed values y, then y^* is simply the vector of residuals $y - \widehat{y}$.
- $m\bar{x}_1^T$ is an $n \times k$ matrix of the means for each covariate for men, likewise for $f\bar{x}_2^T$ for women. $X^* = X - m\bar{x}_1^T - f\bar{x}_2^T$ subtracts the mean effect of gender on each of the covariates in matrix X, so basically X^* is the matrix of demeaned regressors.

(d) Compare $\widetilde{\alpha}$ from the OLS regression

$$y^* = X^*\widetilde{\alpha} + \widetilde{u}$$

with $\widehat{\alpha}$ from the OLS regression

$$y = \widehat{\beta}_1 m + \widehat{\beta}_2 f + X\widehat{\alpha} + \widehat{u}$$

I believe the two end up being equal.

From the second expression, decompose from long regression to short using equation from notes:

$$\gamma_{1(short)} = \beta_{1(long)} + \text{Var}(x_1)^{-1}\text{Cov}(x_1, x_2)\beta_{2(long)}$$
$$\implies \beta_{1(long)} = \gamma_{1(short)} - \text{Var}(x_1)^{-1}\text{Cov}(x_1, x_2)\beta_{2(long)}$$

Also note:

$$\frac{\mathbb{E}[Xm_i]}{\mathbb{E}[m_i^2]} = \frac{(\sum_{i=1}^n X_i m_i)/n}{(\sum_{i=1}^n m_i^2)/n} = \frac{1}{n_1}\sum_{i=1}^n X_i m_i = \bar{x}_1^T$$

Decomposing the coefficients:

$$\begin{aligned}
\widehat{\beta}_1 &= \mathbb{E}[m_i^2]^{-1}\mathbb{E}[m_i y_i] - \mathbb{E}[m_i^2]^{-1}\mathbb{E}[m_i f_i]\beta_{2(long)} - \mathbb{E}[m_i^2]^{-1}\mathbb{E}[m_i X]\widehat{\alpha} \\
&= \frac{\mathbb{E}[m_i y_i]}{\mathbb{E}[m_i^2]} - \frac{0}{\mathbb{E}[m_i^2]}\widehat{\beta}_{2(long)} - \frac{\mathbb{E}[Xm_i]}{\mathbb{E}[m_i^2]}\widehat{\alpha} && (m_i \times f_i = 0) \\
&= \frac{\mathbb{E}[m_i y_i]}{\mathbb{E}[m_i^2]} - \bar{x}_1^T\widehat{\alpha} && \text{(see note above)} \\
\widehat{\beta}_2 &= \mathbb{E}[f_i^2]^{-1}\mathbb{E}[f_i y_i] - \mathbb{E}[f_i^2]^{-1}\mathbb{E}[m_i f_i]\beta_{1(long)} - \mathbb{E}[f_i^2]^{-1}\mathbb{E}[Xf_i]\widehat{\alpha} \\
&= \frac{\mathbb{E}[f_i y_i]}{\mathbb{E}[f_i^2]} - \bar{x}_2^T\widehat{\alpha} && \text{(by similar math)}
\end{aligned}$$

Using these values for β_1, β_2 in the second equation:

$$\begin{aligned}
y &= \widehat{\beta}_1 m + \widehat{\beta}_2 f + X\widehat{\alpha} + \widehat{u} \\
&= \frac{\mathbb{E}[m_i y_i]}{\mathbb{E}[m_i^2]}m - m\bar{x}_1^T\widehat{\alpha} + \frac{\mathbb{E}[f_i y_i]}{\mathbb{E}[f_i^2]}f - f\bar{x}_2^T\widehat{\alpha} + X\widehat{\alpha} + \widehat{u} && \text{(substitute)} \\
&= \frac{\mathbb{E}[m_i y_i]}{\mathbb{E}[m_i^2]}m + \frac{\mathbb{E}[f_i y_i]}{\mathbb{E}[f_i^2]}f + (X - m\bar{x}_2^T - f\bar{x}_2^T)\widehat{\alpha} + \widehat{u} && \text{(group terms)} \\
&= \frac{\mathbb{E}[m_i y_i]}{\mathbb{E}[m_i^2]}m + \frac{\mathbb{E}[f_i y_i]}{\mathbb{E}[f_i^2]}f + (X - m\bar{x}_2^T - f\bar{x}_2^T)\widehat{\alpha} + \widehat{u} && \text{(group terms)} \\
&= \widehat{\gamma}_1 m + \widehat{\gamma}_2 f + X^*\widehat{\alpha} + \widehat{u} && \text{(substitute)}
\end{aligned}$$

I may have gotten lost in the weeds in the math somewhere. But I believe the $\widehat{\gamma}$ terms here are the same as the $\widehat{\beta}$ terms in part C, i.e. the $\widehat{\beta}$ terms in

$$y^* = y - \widehat{\beta}_1 m - \widehat{\beta}_2 f$$

are not the same as those in

$$y = \widehat{\beta}_1 m + \widehat{\beta}_2 f + X\widehat{\alpha} + \widehat{u}$$

because that $X\widehat{\alpha}$ term would shift the values.

If that's the case, then I think I can say

$$\begin{aligned}
y &= \widehat{\gamma}_1 m + \widehat{\gamma}_1 f + X^*\widehat{\alpha} + \widehat{u} && \text{(from above)} \\
y - \widehat{\beta}_1 m - \widehat{\beta}_1 f &= X^*\widehat{\alpha} + \widehat{u} && \text{(sub } \widehat{\gamma}_i = \widehat{\beta}_i \text{ from part c)} \\
y^* &= X^*\widehat{\alpha} + \widehat{u} && \text{(sub } y^* \text{ def)}
\end{aligned}$$

For this to be true and $y^* = X^*\widetilde{\alpha} + \widetilde{u}$ to be true, since the error terms are mean zero and uncorrelated with X, then necessarily $\widehat{\alpha} = \widetilde{\alpha}$.

This seemed like a long route to a simple answer, so I look forward to seeing the answer key...

2. Computational Exercise
   Read through 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 estimate the model in columns (1), (2) and (3) of Table 3 of Charness and Kuhn. Calculate classic standard errors (the authors report cluster-robust standard errors, so your estimated standard errors will not match those in the table).

   Finally, for the models of columns (2) and (3) test the hypothesis that the coefficient on relative wage equals zero and provide the p-value for the estimated test statistic.

Stata output (code in following section)

For Stata, I used the `reg` function, e.g. `reg e1 w1 rel_w1 if pubwage == 1` to regress effort vs. wages and relative wages filtered for public wage observations, and then using `eststo` and `esttab` to store and display the model information. See code in next section.

Table 1: Type 1 workers

	(1)	(2)	(3)
	e1	e1	e1
w1	.589	.581	.594
	(.0431)	(.0535)	(.0509)
rel_w1		.00722	
		(.0317)	
rel_w1_low			-.00706
			(.034)
N	555	555	555
R^2	0.252	0.252	0.252

Table 2: Type 2 workers

	(1)	(2)	(3)
	e2	e2	e2
w2	.536	.549	.523
	(.0391)	(.066)	(.0414)
rel_w2		-.0131	
		(.0532)	
rel_w2_low			.0986
			(.105)
N	555	555	555
R^2	0.253	0.253	0.255

Standard errors in parentheses

Here are the hypothesis test outputs from Stata arranged into a table. For both worker types, neither of the tested hypotheses (relative wages = 0) are statistically significant:

Here are the hypothesis test outputs from Stata `test` function (p-value is "Prob > F" value). In all cases, it looks like the tested hypotheses (relative wages = 0) cannot be rejected, due to high p values.

Table 3: p-values

	rel wage	rel wage * (rel wage < 0)
Type 1 workers	0.8202	0.8355
Type 2 workers	0.8057	0.3489

Matlab output

In Matlab, I calculated the β terms three different ways, defining y as $e1$ and X as a matrix with regressors depending on the model. I included a constant in the X matrix to facilitate the first method:

- using matrix algebra: $\beta = (X'X)^-1(X'y)$
- using function `regress()`: `regress(y, X)`
- using function `fitlm()`: e.g. `lm1 = fitlm(wage_pub, 'e1 ~ w1')`

The latter also reports standard errors, R^2 values, etc. See code in next section.

Note, I cleaned up the output a bit by removing extra lines and spaces… all the outputs match well with the Stata outputs and the columns in the published paper.

Type 1 workers

```
lm1 = Linear regression model:
    e1 ~ 1 + w1
Estimated Coefficients:
                   Estimate       SE       tStat      pValue
    (Intercept)    0.10896     0.052721    2.0667      0.03923
    w1             0.58867     0.043111    13.655    8.644e-37
Number of observations: 555, Error degrees of freedom: 553
Root Mean Squared Error: 0.762
R-squared: 0.252,  Adjusted R-Squared 0.251
F-statistic vs. constant model: 186, p-value = 8.64e-37

lm2 = Linear regression model:
    e1 ~ 1 + w1 + rel_w1
Estimated Coefficients:
                   Estimate        SE        tStat       pValue
    (Intercept)      0.12153    0.076438       1.59       0.11242
    w1               0.58148    0.053495      10.87    4.5202e-25
    rel_w1         0.0072182    0.031742    0.22741       0.82019
Number of observations: 555, Error degrees of freedom: 552
Root Mean Squared Error: 0.763
R-squared: 0.252,  Adjusted R-Squared 0.25
F-statistic vs. constant model: 93.1, p-value = 1.45e-35
p2 = 0.8202


lm3 = Linear regression model:
    e1 ~ 1 + w1 + rel_w1_low
Estimated Coefficients:
                    Estimate        SE        tStat        pValue
    (Intercept)      0.097328    0.076922      1.2653        0.2063
    w1                0.59429    0.050907      11.674    2.7001e-28
    rel_w1_low     -0.0070601    0.033984    -0.20775        0.8355
```

```
Number of observations: 555, Error degrees of freedom: 552
Root Mean Squared Error: 0.763
R-squared: 0.252,  Adjusted R-Squared 0.25
F-statistic vs. constant model: 93.1, p-value = 1.46e-35
p3 = 0.8355
```

Type 2 workers

```
lm1 = Linear regression model:
    e2 ~ 1 + w2
Estimated Coefficients:
                   Estimate       SE       tStat       pValue
    (Intercept)    0.10211     0.079095     1.291       0.19725
    w2             0.53594     0.039118    13.701    5.4003e-37
Number of observations: 555, Error degrees of freedom: 553
Root Mean Squared Error: 0.94
R-squared: 0.253,  Adjusted R-Squared 0.252
F-statistic vs. constant model: 188, p-value = 5.4e-37

lm2 = Linear regression model:
    e2 ~ 1 + w2 + rel_w2
Estimated Coefficients:
                   Estimate        SE        tStat        pValue
    (Intercept)     0.089509    0.094282     0.94937       0.34285
    w2               0.54901    0.065984      8.3204    6.8987e-16
    rel_w2         -0.013095    0.053221    -0.24605       0.80573
Number of observations: 555, Error degrees of freedom: 552
Root Mean Squared Error: 0.941
R-squared: 0.253,  Adjusted R-Squared 0.251
F-statistic vs. constant model: 93.7, p-value = 9.05e-36
p2 = 0.8057


lm3 = Linear regression model:
    e2 ~ 1 + w2 + rel_w2_low
Estimated Coefficients:
                   Estimate       SE        tStat       pValue
    (Intercept)     0.13427    0.086223     1.5572       0.11999
    w2              0.52312    0.041444     12.622    2.8701e-32
    rel_w2_low     0.098636     0.10522    0.93744       0.34894
Number of observations: 555, Error degrees of freedom: 552
Root Mean Squared Error: 0.94
R-squared: 0.255,  Adjusted R-Squared 0.252
F-statistic vs. constant model: 94.3, p-value = 6.01e-36
p3 = 0.3489
```

R outputs

In R, I used the function `lm()` to calculate linear regressions for each model, e.g. `lm1c <- lm(e1 ~ w1 + rel_w1_low, data = wage_pub)`. See code in next section.

I used the `broom::tidy()` function to clean up the standard model outputs into a nice data frame for easier printing using `knitr::kable()`.

All the outputs match well with the Stata outputs and the columns in the published paper.

term	estimate	std.error	statistic	p.value	model
(Intercept)	0.1089562	0.0527207	2.066671	0.0392301	e1 ~ w1
w1	0.5886740	0.0431110	13.654840	0.0000000	e1 ~ w1

term	estimate	std.error	statistic	p.value	model
(Intercept)	0.1215327	0.0764380	1.5899519	0.1124182	e1 ~ w1 + rel_w1
w1	0.5814829	0.0534951	10.8698364	0.0000000	e1 ~ w1 + rel_w1
rel_w1	0.0072182	0.0317415	0.2274064	0.8201919	e1 ~ w1 + rel_w1

term	estimate	std.error	statistic	p.value	model
(Intercept)	0.0973283	0.0769219	1.2652863	0.2063026	e1 ~ w1 + rel_w1_low
w1	0.5942863	0.0509073	11.6738908	0.0000000	e1 ~ w1 + rel_w1_low
rel_w1_low	-0.0070601	0.0339836	-0.2077515	0.8354996	e1 ~ w1 + rel_w1_low

term	estimate	std.error	statistic	p.value	model
(Intercept)	0.1021090	0.0790951	1.290966	0.197255	e2 ~ w2
w2	0.5359437	0.0391182	13.700627	0.000000	e2 ~ w2

term	estimate	std.error	statistic	p.value	model
(Intercept)	0.0895085	0.0942824	0.9493662	0.3428499	e2 ~ w2 + rel_w2
w2	0.5490124	0.0659835	8.3204459	0.0000000	e2 ~ w2 + rel_w2
rel_w2	-0.0130953	0.0532211	-0.2460539	0.8057319	e2 ~ w2 + rel_w2

term	estimate	std.error	statistic	p.value	model
(Intercept)	0.1342696	0.0862226	1.5572423	0.1199861	e2 ~ w2 + rel_w2_low
w2	0.5231221	0.0414443	12.6222822	0.0000000	e2 ~ w2 + rel_w2_low
rel_w2_low	0.0986357	0.1052176	0.9374444	0.3489400	e2 ~ w2 + rel_w2_low

Code

Stata code

```
set more off
clear

// Set working directory

cd "~/github/econ_courses/econ241b/assts/asst3"

// The key variables:
//     e1 = effort of type 1 workers
//     w1 = wage of type 1 workers
//     w2 = wage of type 2 workers
//     pubwage = flag for public-wage regime

clear all
use "prob_set_3.dta"

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

///////////////////////////////////////////////
// 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 1: effort vs wage
// Col 2: effort vs wage and relative wage (symmetric model)
// Col 3: effort vs wage and relative wage (< 0, asymmetric model)
eststo clear

reg e1 w1             if pubwage == 1
eststo col_a1

reg e1 w1 rel_w1      if pubwage == 1
eststo col_a2
test rel_w1 = 0

reg e1 w1 rel_w1_low if pubwage == 1
eststo col_a3
test rel_w1_low = 0
```

```
esttab col_a1 col_a2 col_a3 using table3a.tex, title(B: Type 1 workers) ///
    noconst b(%10.3g) 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 1: effort vs wage
// Col 2: effort vs wage and relative wage (symmetric model)
// Col 3: effort vs wage and relative wage (< 0, asymmetric model)
eststo clear

reg e2 w2            if pubwage == 1
eststo col_b1

reg e2 w2 rel_w2     if pubwage == 1
eststo col_b2
test rel_w2 = 0

reg e2 w2 rel_w2_low if pubwage == 1
eststo col_b3
test rel_w2_low = 0

esttab col_b1 col_b2 col_b3 using table3b.tex, title(B: Type 2 workers) ///
    noconst b(%10.3g) se r2 nostar replace

///////////////////////////////////////////////
// Export data to .csv (for Matlab and R)
///////////////////////////////////////////////

outsheet period wrk2id  wrk1id  w1  w2  e1  e2  pubwage ///
    using prob_set3.csv, comma replace
```

Matlab code

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

filename = '/Users/ohara/github/econ_courses/econ241b/assts/asst3/prob_set3.csv';

raw = readtable(filename);

%% set up data: filter to public, create a constant vector
wage_pub = raw(raw.pubwage == 1, :);

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

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% TYPE 1 WORKER
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% Effort vs wages, type 1 worker
% set up X1 to be a constant plus own wage regressor
X1 = [const wage_pub.w1];
y  = wage_pub.e1;

% NOTE: calculating coefficients three different ways:
% * using matrix math.
% * using 'regress' function.
% * using 'fitlm' function.  All three returned identical coefficients.
% For the assignment results, I am using the output of 'fitlm' call.

b1 = (X1' * X1)^-1 * (X1' * y);
col_a1 = regress(y, X1);
lm1 = fitlm(wage_pub, 'e1 ~ w1')


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% Effort vs wages and rel wages, type 1 worker
% set up X2 to be a constant plus own wage regressor plus relative wage
wage_pub.rel_w1 = wage_pub.w1 - wage_pub.w2;

X2 = [const wage_pub.w1 wage_pub.rel_w1];
y  = wage_pub.e1;

b2 = (X2' * X2)^-1 * (X2' * y);
col_a2 = regress(y, X2);
lm2 = fitlm(wage_pub, 'e1 ~ w1 + rel_w1')

% Test hypothesis on rel_w1
p2 = coefTest(lm2, [0 0 1])

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% Effort vs wages and low rel wages, type 1 worker
% set up X3 to be a constant plus own wage regressor
%   plus (rel wage below zero)
```

```
wage_pub.rel_w1_low = wage_pub.rel_w1 .* (wage_pub.rel_w1 < 0);

X3 = [const wage_pub.w1 wage_pub.rel_w1_low];
y  = wage_pub.e1;

b3 = (X3' * X3)^-1 * (X3' * y);
col_a3 = regress(y, X3);
lm3 = fitlm(wage_pub, 'e1 ~ w1 + rel_w1_low')

% Test hypothesis on rel_w1_low
p3 = coefTest(lm3, [0 0 1])

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% TYPE 2 WORKER
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% Effort vs wages, type 2 worker
% set up X1 to be a constant plus own wage regressor
X1 = [const wage_pub.w2];
y  = wage_pub.e2;

b1 = (X1' * X1)^-1 * (X1' * y);
col_a1 = regress(y, X1);
lm1 = fitlm(wage_pub, 'e2 ~ w2')


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% Effort vs wages and rel wages, type 2 worker
% set up X2 to be a constant plus own wage regressor plus relative wage
wage_pub.rel_w2 = wage_pub.w2 - wage_pub.w1;

X2 = [const wage_pub.w2 wage_pub.rel_w2];
y  = wage_pub.e2;

b2 = (X2' * X2)^-1 * (X2' * y);
col_a2 = regress(y, X2);
lm2 = fitlm(wage_pub, 'e2 ~ w2 + rel_w2')

% Test hypothesis on rel_w2
p2 = coefTest(lm2, [0 0 1])

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% Effort vs wages and low rel wages, type 2 worker
% set up X3 to be a constant plus own wage regressor plus
%   (rel wage below zero)

wage_pub.rel_w2_low = wage_pub.rel_w2 .* (wage_pub.rel_w2 < 0);

X3 = [const wage_pub.w2 wage_pub.rel_w2_low];
y  = wage_pub.e2;

b3 = (X3' * X3)^-1 * (X3' * y);
col_a3 = regress(y, X3);
lm3 = fitlm(wage_pub, 'e2 ~ w2 + rel_w2_low')
```

```
% Test hypothesis on rel_w2_low
p3 = coefTest(lm3, [0 0 1])
```

R code

```r
suppressPackageStartupMessages(library(tidyverse))

raw <- read_csv('prob_set3.csv', col_types = cols())

### Filter to public; mutate new columns for rel wage and rel wage < 0.
wage_pub <- raw %>%
  filter(pubwage == 1) %>%
  mutate(rel_w1 = w1 - w2,
         rel_w1_low = rel_w1 * (rel_w1 < 0),
         rel_w2 = w2 - w1,
         rel_w2_low = rel_w2 * (rel_w2 < 0))

### Use lm() to calculate linear regressions for each model.
lm1a <- lm(e1 ~ w1,              data = wage_pub)
lm1b <- lm(e1 ~ w1 + rel_w1,     data = wage_pub)
lm1c <- lm(e1 ~ w1 + rel_w1_low, data = wage_pub)

lm2a <- lm(e2 ~ w2,              data = wage_pub)
lm2b <- lm(e2 ~ w2 + rel_w2,     data = wage_pub)
lm2c <- lm(e2 ~ w2 + rel_w2_low, data = wage_pub)

### Join the models into a list for easy iteration in a loop.
model_list <- list(lm1a, lm1b, lm1c, lm2a, lm2b, lm2c)

for(mdl in model_list) {
  ### Loop over all models, use broom::tidy to clean up model output,
  ### and add a column to note the model as a formula.
  mdl_text <- mdl$call %>% as.character()
  mdl_df <- mdl %>%
    broom::tidy() %>%
    mutate(model = mdl_text[2])
  ### Print output using knitr::kable().
  print(knitr::kable(mdl_df))
}
```