The Empirical Human Capital Model

Simple Version

For the simplest version of the model

$$ w_i = A_0 \cdot \exp(r_i \cdot S_i + v_i) $$

where $w_i$ = wage rate, $S_i$ = years of schooling, and $v_i$ is a random disturbance. This model implies

$$ \ln w_i = \ln(A_0) + r_i \cdot S_i + v_i = \beta_0 + \beta_S \cdot S_i + v_i $$

The following results are run on just the 1979 male sample in CPSORG.

```
. reg lnwage ed
```

```
Source |       SS       df       MS              Number of obs =    3096
-------------+------------------------------           F(  1,  3094) =  350.61
Model |  66.7595844     1  66.7595844           Prob > F      =  0.0000
Residual |  589.123863  3094  .190408488           R-squared     =  0.1018
-------------+------------------------------           Adj R-squared =  0.1015
Total |  655.883447  3095  .211917107           Root MSE      =  .43636

------------------------------------------------------------------------------
lnwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ed |   .0499781   .0026691    18.72   0.000     .0447447    .0552115
_cons |   2.202434   .0341515    64.49   0.000     2.135472    2.269396
------------------------------------------------------------------------------
```
More Realistic

We might think of allowing for other variables to enter the this human capital model. For instance, it is though that increases in labor market experience increase human capital and, thus, wages

\( \ln w_i = \beta_0 + \beta_x \cdot S_i + \beta_{ex} \cdot ex_i + v_i \) (2)

```
reg lnwage ed ex
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 3096</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>137.84</td>
<td>2</td>
<td>68.92</td>
<td>F( 2, 3093) = 411.49</td>
</tr>
<tr>
<td>Residual</td>
<td>518.04</td>
<td>3093</td>
<td>.17</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>655.88</td>
<td>3095</td>
<td>.21</td>
<td>R-squared = 0.2102</td>
</tr>
</tbody>
</table>

```
Source | SS    | df  | MS       | Number of obs = 3096 |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>162.48</td>
<td>3</td>
<td>54.16</td>
<td>F( 3, 3092) = 339.40</td>
</tr>
<tr>
<td>Residual</td>
<td>493.40</td>
<td>3092</td>
<td>.16</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>655.88</td>
<td>3095</td>
<td>.21</td>
<td>R-squared = 0.2477</td>
</tr>
</tbody>
</table>

```
| lnwage | Coef.  | Std. Err. | t    | P>|t|  | [95% Conf. Interval] |
|--------|--------|-----------|------|------|----------------------|
| ed     | .0698  | .0027     | 26.02| 0.000| .0645 - .0751        |
| ex     | .0122  | .0006     | 20.60| 0.000| .0101 - .0133        |
| _cons | 1.72    | .040      | 43.35| 0.000| 1.65 - 1.79          |

One of the problems with the above model is that returns to labor market experience are constant when we think there are many reasons to believe they are not. The fix for this is to allow for the rate of return to a year of labor market experience to depend on how much experience you actually have. This can be done by including an experience squared term in the above regression equation

\( \ln w_i = \beta_0 + \beta_x \cdot S_i + \beta_{ex} \cdot ex_i + \beta_{exsq} \cdot (ex_i)^2 + v_i \) (3)

```
reg lnwage ed ex exsq
```

Source | SS    | df  | MS       | Number of obs = 3096 |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>162.48</td>
<td>3</td>
<td>54.16</td>
<td>F( 3, 3092) = 339.40</td>
</tr>
<tr>
<td>Residual</td>
<td>493.40</td>
<td>3092</td>
<td>.16</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>655.88</td>
<td>3095</td>
<td>.21</td>
<td>R-squared = 0.2470</td>
</tr>
</tbody>
</table>

```
Source | SS    | df  | MS       | Number of obs = 3096 |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>162.48</td>
<td>3</td>
<td>54.16</td>
<td>F( 3, 3092) = 339.40</td>
</tr>
<tr>
<td>Residual</td>
<td>493.40</td>
<td>3092</td>
<td>.16</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>655.88</td>
<td>3095</td>
<td>.21</td>
<td>R-squared = 0.2470</td>
</tr>
</tbody>
</table>

```
| lnwage | Coef.  | Std. Err. | t    | P>|t|  | [95% Conf. Interval] |
|--------|--------|-----------|------|------|----------------------|
| ed     | .0630  | .0027     | 23.51| 0.000| .0577 - .0685        |
| ex     | .0360  | .0020     | 17.97| 0.000| .0320 - .0400        |
| exsq | -.0007 | .0000     | -12.43| 0.000| -.0008 - -.0006      |
| _cons | 1.65    | .040      | 41.94| 0.000| 1.57 - 1.72          |

2
One Step Closer To Reality

One problem with the latest version of the model is our estimates of the expected wage conditional on educational attainment and labor market experience is the same for blacks, whites, and Hispanics. We could relax this by allowing different intercepts for individuals from different racial or ethnic groups. These different intercepts could reflect different average levels of initial human capital \( (A_0) \) or the effects of discrimination.

\[
\ln w_i = \beta_0 + \beta_{\text{black}} \cdot \text{black}_i + \beta_{\text{hisp}} \cdot \text{Hisp}_i + \beta_s \cdot S_i + \beta_{\text{ex}} \cdot \text{ex}_i + \beta_{\text{ex}, sq} \cdot (\text{ex}_i)^2 + v_i
\]

In this model the variables \( \text{black} \) and \( \text{Hisp} \) are known as dummy variables or indicator variables.

In model (4) there are effectively three intercept terms.

Intercept for whites: \( \beta_0 \)

Intercept for blacks: \( \beta_0 + \beta_{\text{black}} \)

Intercept for Hispanics: \( \beta_0 + \beta_{\text{hisp}} \)

An equivalent model in terms of information content would be

\[
\ln w_i = \alpha_{\text{white}} \cdot \text{white}_i + \alpha_{\text{black}} \cdot \text{black}_i + \alpha_{\text{hisp}} \cdot \text{Hisp}_i + \beta_s \cdot S_i + \beta_{\text{ex}} \cdot \text{ex}_i + \beta_{\text{ex}, sq} \cdot (\text{ex}_i)^2 + v_i
\]

In this model the intercept for whites is

Intercept for whites: \( \alpha_{\text{white}} \)

Intercept for blacks: \( \alpha_{\text{black}} \)

Intercept for Hispanics: \( \alpha_{\text{hisp}} \)
The results from the model (4) estimation are

```
> reg lnwage black hisp ed ex exsq
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>166.661253</td>
<td>5</td>
<td>33.3322506</td>
</tr>
<tr>
<td>Residual</td>
<td>489.222194</td>
<td>3090</td>
<td>.158324335</td>
</tr>
<tr>
<td>Total</td>
<td>655.883447</td>
<td>3095</td>
<td>.211917107</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>lnwage</th>
<th>Coef. [95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>black</td>
<td>-.1330717 [-.1889619, -.0771816]</td>
</tr>
<tr>
<td>hisp</td>
<td>-.0836429 [-.1481287, -.0191571]</td>
</tr>
<tr>
<td>ed</td>
<td>.0602155 [.0548355, .0655955]</td>
</tr>
<tr>
<td>ex</td>
<td>.0362958 [.0323783, .0402132]</td>
</tr>
<tr>
<td>exsq</td>
<td>-.0005539 [-.0006395, -.0004682]</td>
</tr>
<tr>
<td>_cons</td>
<td>1.691663 [1.612142, 1.771185]</td>
</tr>
</tbody>
</table>

And the results from the model (5) specification

```
> reg lnwage black white hisp ed ex exsq, nocon
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>24871.5301</td>
<td>6</td>
<td>4145.25502</td>
</tr>
<tr>
<td>Residual</td>
<td>489.222194</td>
<td>3090</td>
<td>.158324335</td>
</tr>
<tr>
<td>Total</td>
<td>25360.7523</td>
<td>3096</td>
<td>8.19145747</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>lnwage</th>
<th>Coef. [95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>black</td>
<td>1.558592 [1.468316, 1.648868]</td>
</tr>
<tr>
<td>white</td>
<td>1.691663 [1.612142, 1.771185]</td>
</tr>
<tr>
<td>hisp</td>
<td>1.608021 [1.518385, 1.697656]</td>
</tr>
<tr>
<td>ed</td>
<td>.0602155 [.0548355, .0655955]</td>
</tr>
<tr>
<td>ex</td>
<td>.0362958 [.0323783, .0402132]</td>
</tr>
<tr>
<td>exsq</td>
<td>-.0005539 [-.0006395, -.0004682]</td>
</tr>
</tbody>
</table>

Note the relationship between the model (4) and model (5) coefficients.
(Hypothesis) Test Involving More than Just One Coefficient Value (The F-test)

Note that

\[ \ln w_i = \beta_0 + \beta_1 \cdot S_i + \beta_{\text{ex}} \cdot \text{ex}_i + \beta_{\text{ex,sq}} \cdot (\text{ex}_i)^2 + \nu_i \]

is just a restricted version of

\[ \ln w_i = \beta_0 + \beta_{\text{black}} \cdot \text{black}_i + \beta_{\text{hisp}} \cdot \text{hisp}_i + \beta_1 \cdot S_i + \beta_{\text{ex}} \cdot \text{ex}_i + \beta_{\text{ex,sq}} \cdot (\text{ex}_i)^2 + \nu_i \]

Basically model (3) is equivalent to model (4) with the values of \( \beta_{\text{black}} = \beta_{\text{hisp}} = 0 \).

The F-test provides a way of testing these and other linear restrictions. The F-test is based on the test statistic

\[
F = \frac{\frac{SSE^* - SSE}{r}}{\frac{SSE}{n - (k + 1)}} \sim F(r, n - (k + 1))
\]

where

- \( r = \) the number of restrictions (the number of equal signs)
- \( SSE^* = \) the error sum of squares from a restricted model
- \( SSE = \) the error sum of squares from the unrestricted model

If we had the error sum of squares from the restricted and unrestricted models we could compute the value of the test statistic. It turns out that there is a much easier way to compute the value of the test statistic. If I divide both the numerator and denominator of the F-statistic by \( \text{SST} \) (Total Sum of Squares) we get

\[
F = \frac{1}{r} \left[ \frac{SSE^*}{\text{SST}} \right] \frac{SSE}{\text{SST}} = \frac{1}{r} \left[ 1 - R^2 - (1 - R^2) \right] = \frac{R^2 - R^2}{r} = \frac{1 - R^2}{n - (k + 1)}
\]

Basically all I need to do to conduct the test is estimate both the restricted and the unrestricted models, get the R-squared terms from these regressions, form the F-statistic and go to the tables.
It is now possible to test the restrictions implied by model (3) given only the output from
the regression table provided in this handout.

\[
F = \frac{\left( \frac{R^2 - R_0^2}{r} \right)}{\frac{1 - R^2}{n - (k + 1)}} = \frac{0.2541 - 0.2477}{2} \approx 13.26
\]

The p-value associated with this test statistic is infinitesimal ⇒ reject the restrictions.

It turns out there is also a very easy way to compute this test statistic (and associated
p-value) using STATA’s post estimation commands.

Simply estimate the model

```
. reg lnwage black hisp ed ex exsq
```

After estimating the model you can use the test command to get a variety of test statistics

```
. test black=0
( 1)  black = 0.0

F(  1,  3090) =  21.79
Prob > F =  0.0000
```

This is the test statistic associated with the hypothesis that the coefficient on \textit{black} is
equal to zero.

```
. test hisp=0, accum
( 1)  black = 0.0
( 2)  hisp = 0.0

F(  2,  3090) =  13.20
Prob > F =  0.0000
```
This is a test statistic associated with testing the restrictions implied by model (3) (i.e. $\beta_{\text{black}} = \beta_{\text{Hisp}} = 0$). Note the use of the accum option.

**Another Example**

Suppose we thought that returns to school vary by race. We could estimate a model that allowed returns to school to vary by race by including race-schooling **interaction variables** into the model (4) specification.

\[
\ln w_i = \beta_0 + \beta_{\text{black}} \cdot \text{black}_i + \beta_{\text{hisp}} \cdot \text{Hisp}_i + \beta_{\text{black-s}} \cdot (\text{black}_i \cdot S_i) + \beta_{\text{Hisp-s}} \cdot (\text{Hisp}_i \cdot S_i) \\
+ \beta_s \cdot S_i + \beta_{\text{ex}} \cdot \text{ex}_i + \beta_{\text{ex-sq}} \cdot (\text{ex}_i)^2 + v_i
\]  

In this model the rate of return to a year of school for whites is $\beta_s$. For blacks and Hispanics the rates of return to a year of schooling $\beta_s + \beta_{\text{black-s}}$ and $\beta_s + \beta_{\text{Hisp-s}}$ respectively.

Note that model (4) is just a restricted version of this model. The restrictions implied by (4) are $\beta_{\text{black-s}} = \beta_{\text{Hisp-s}} = 0$

We can test these restrictions using the F-Statistic. As with the previous example there are two approaches

1. Estimate the restricted and unrestricted models, obtain the R-squared terms, and construct the test statistic. (You will need to be able to do this for the exam).
2. Estimate the unrestricted model and use STATA’s test command.
I will use the second approach

```
. gen black_ed=black*ed
. gen hisp_ed=hisp*ed
. reg lnwage black hisp black_ed hisp_ed ed ex exsq
```

```
Source |       SS       df       MS              Number of obs =    3096
--------+------------------------------           F(  7,  3088) =  150.53
Model |  166.862028     7  23.8374325           Prob > F      =  0.0000
Residual |   489.02142  3088  .158361859           R-squared     =  0.2544
Total |  655.883447  3095  .211917107           Adj R-squared =  0.2527

------------------------------------------------------------------------------
  lnwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   black |  -.2529284   .1115613    -2.27   0.023    -.4716703   -.0341866
   hisp |  -.1117459   .0983421    -1.14   0.256    -.3045685    .0810766
black_ed |   .0104171   .0093896     1.11   0.267    -.0079934    .0288275
hisp_ed |   .0025325   .0088625     0.29   0.775    -.0148444    .0199094
    ed |   .0592706   .0029542    20.06   0.000     .0534781     .065063
    ex |   .0362238    .002003    18.09   0.000     .0322966    .0401511
exsq |  -.0005515   .0000438   -12.58   0.000    -.0006375   -.0004656
   _cons |   1.703744   .0427701    39.83   0.000     1.619884    1.787605
------------------------------------------------------------------------------

. test black_ed=0
( 1)  black_ed = 0.0

F(  1,  3088) =  1.23
    Prob > F =  0.2673

. test hisp_ed=0, accum
( 1)  black_ed = 0.0
( 2)  hisp_ed = 0.0

F(  2,  3088) =  0.63
    Prob > F =  0.5306
```

On the basis of this test statistic we cannot reject restrictions implied by model (4).

As an additional exercise you should compute the value of the test statistic manually. If it is not approximately equal to 0.63 you did something wrong.