Taking the Model to the Data

October 4, 2015

An estimable model

Lets think of a nice way to bring model to data

$$w_{it} = \varphi(S, X)\, e^{\theta_i + \varepsilon_{it}}$$

where

- S : schooling
- X : experience
- $\theta_i + \varepsilon_{it}$: error term
- φ: human capital production function or hedonic pricing function
- Most models assume it is production function

Take logs

$$\log(w_{it}) = \log(\varphi(S, X)) + \theta_i + \varepsilon_{it}$$

Lets assume that

- ε_{it} just represents measurement error
  It is irrelevant to the agent
- No tuition (or at least income during school$\simeq$tuition paid)
- Separability

$$\varphi(S,X) = f(S)g(X)$$

Then

$$\begin{aligned} V_i(S;r) &= \int_S^\infty e^{-rt}\varphi(S,t-S)\,e^{\theta_i}dt \\ &= \int_S^\infty e^{-rt}f(S)g(t-S)e^{\theta_i}dt \\ &= e^{-rS+\theta_i}f(S)\int_S^\infty e^{-r(t-S)}g(t-S)\,dt \\ &= e^{-rS+\theta_i}f(S)\int_0^\infty e^{-rX}g(X)\,dX \end{aligned}$$

Internal rate of return comparing schooling levels S and $S+d$ is defined as $\rho(S, S+d)$

$$V_i(S+d;\rho(S,S+d)) = V_i(S;\rho(S,S+d))$$

$$e^{-\rho(S,S+d)(S+d)+\theta_i} f(S+d) \int_0^\infty e^{-\rho(S,S+d)X} g(X)\, dX$$
$$= e^{-\rho(S,S+d)S+\theta_i} f(S) \int_0^\infty e^{-\rho(S,S+d)X} g(X)\, dX$$

$$e^{-\rho(S,S+d)S} f(S+d) = e^{-\rho(S,S+d)S} f(S)$$

Thus

$$\rho\left(S, S+d\right) = \frac{\log\left(f(S+d)\right) - \log\left(f(S)\right)}{d}$$

To estimate this model we can just run a nonparametric regression since

$$\log\left(w_{it}\right) = \log(f(S)) + \log(g(X)) + \theta_i + \varepsilon_{it}$$

From this you can get the internal rate of return to schooling

People often assume that

$$\log(f(S)) = \beta S$$

Then β is the internal rate of return to schooling

The phrase “returns to” has taken on a much broader meaning

Often use dummies instead of linear term

The Mincer Model

Suppose each period you spend some time working and the rest investing in human capital

Let

- $K(t)$ be percentage of time spent investing in human capital
- $h(0)$ human capital at birth

Suppose the human capital production function is

$$\dot{H} = \rho H(t)K(t)$$

Solving the differential equation yields

$$\log(H(t)) = \log(H(0)) + \int_0^t \rho K(t)dt$$

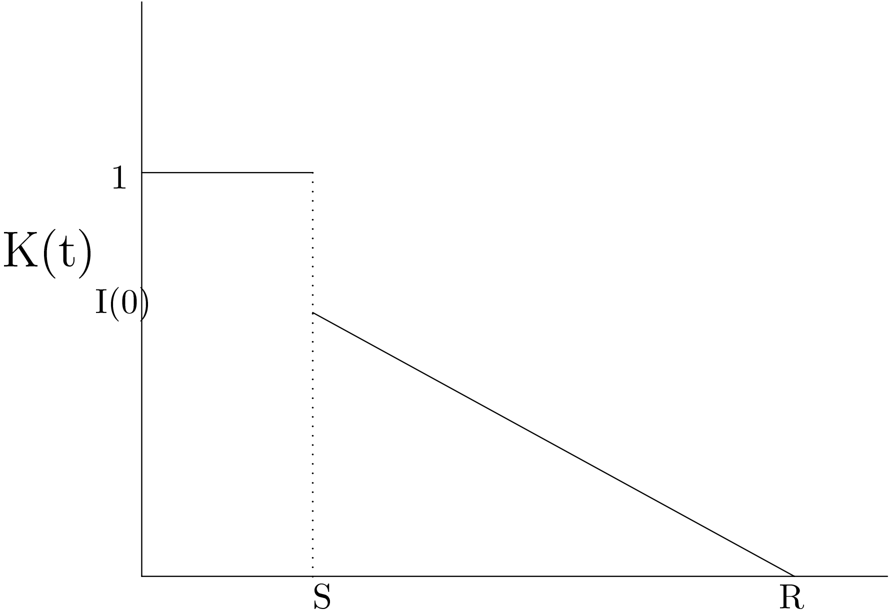
Normalizing human capital rental rate to 1

$$w(t) = H(t)\,(1 - K(t))$$

People leave school at time S

Mincer then assumes that

- Retire from the labor market at time R
- $K(t) = 1$ while in school
- $K(t)$ is linear in the labor force
- $K(R) = 0$



K(t)
1
I(0)
S
R

Let x be experience, i.e. $x = t - S$ and let $I(x) = K(t)$

$$H(S) = H(0)e^{\rho S}$$
$$I(x) = I(0) - \frac{I(0)x}{R - S}$$

so

$$\begin{aligned} \log(H(x)) &= \log(H(S)) + \rho \int_0^x I(\xi) d\xi \\ &= \log(H(0)) + \rho S + \rho \left[I(0)x - \frac{I(0)x^2}{2(R - S)} \right] \\ &\approx \beta_0 + \beta_1 S + \beta_2 x + \beta_3 x^2 \end{aligned}$$

The famous Mincer specification $\beta_1 = \rho =$ internal rate of return to schooling

This is a huge empirical success

Problems:

- Is $K(t)$ really linear?
- Is β_1 the same for everyone?
- Not “structural” in the classic sense
- Implies everyone is exactly indifferent between all levels of schooling-in which case schooling should be very sensitive to anything

Next we will consider an empirical model that relax some of these assumptions.

Willis and Rosen(1978)

They think of schooling based on the Roy model of comparative advantage

Keep things very simple

2 Schooling Choices:

Attend College or not

Let

- V_{ci} be present value of earnings as a college graduate
- V_{Hi} be present value of earnings as a high school graduate

Go to college if $V_{ci} > V_{Hi}$

Assume exponential growth in earnings

For college

$$W_{cit} = \begin{cases} 0 & t \leq s \\ \overline{Y}_{ci} e^{g_{ci}(t-s)} & t > s \end{cases}$$

- c represents college
- s is number of years it takes to get a college degree
- t represents age (measured as years since high school graduation)
- i is an individual
- $\overline{Y}_{ci}$ is initial wage
- g_{ci} growth rates in wages

High School:

$$W_{Hit} = \overline{Y}_{Hi} e^{g_{Hi}(t)}$$

with these terms being defined analogously

Let r_i be individual specific interest rate

Then putting it together

$$\begin{aligned}
V_{ci} &= \int_0^\infty e^{-r_i t} W_{cit} dt \\
&= \int_s^\infty e^{-r_i t} \overline{Y}_{ci} e^{g_{ci}(t-s)} dt \\
&= e^{-r_i s} \overline{Y}_{ci} \int_0^\infty e^{(g_{ci}-r_i)t} dt \\
&= \frac{e^{-r_i s} \overline{Y}_{ci}}{r_i - g_{ci}} \\
V_{Hi} &= \int_0^\infty e^{-r_i t} \overline{Y}_{Hi} e^{g_{Hi} t} dt \\
&= \frac{\overline{Y}_{Hi}}{r_i - g_{Hi}}
\end{aligned}$$

As

- $\overline{Y}_{ci} \uparrow \Longrightarrow$ college $\uparrow$
- $g_{ci} \uparrow \Longrightarrow$ college $\uparrow$
- $r_i \uparrow \Longrightarrow$ college $\downarrow$

They allow for heterogeneity in:

- interest rates (r_i)
- initial wages ($\overline{Y}_{ci}, \overline{Y}_{Hi}$)
- growth rates (g_{ci}, g_{Hi})

Go to college if $V_{ci} > V_{Hi}$

or

$$-r_i s + \log\left(\overline{Y}_{ci}\right) - \log\left(r_i - g_{ci}\right) > \log\left(\overline{Y}_{Hi}\right) - \log\left(r_i - g_{Hi}\right)$$

They cheat at this point and assume that

$$\log(V_{ci}) - \log\left(V_{Hi}\right) \approx \alpha_0 + \alpha_1\left[\log\left(\overline{Y}_{ci}\right) - \log\left(\overline{Y}_{Hi}\right)\right] + \alpha_2 g_{ci} + \alpha_3 g_{Hi} + \alpha_4 r_i$$

Assume that

$$
\begin{aligned}
\log\left(\overline{Y}_{ci}\right) &= X_i'\beta_c + u_{1i} \\
\log\left(\overline{Y}_{Hi}\right) &= X_i'\beta_H + u_{2i} \\
g_{ci} &= X_i'\gamma_c + u_{3i} \\
g_{Hi} &= X_i'\gamma_H + u_{4i} \\
r_i &= Z_i'\delta + u_{5i}
\end{aligned}
$$

with all of the error terms normally distributed

Data is

- NBER-Thorndike-Hagen Survey of 1968-1971
- Male World War II Veterans who applied for Army Air Corps
- Not random
- Data on schooling
- Wage at two different points in time

Define

$$\begin{aligned} I_i &= \log(V_{ci}) - \log(V_{Hi}) \\ &= \alpha_0 + \alpha_1 \left[\log\left(\overline{Y}_{ci}\right) - \log\left(\overline{Y}_{Hi}\right)\right] + \alpha_2 g_{ci} + \alpha_3 g_{Hi} + \alpha_4 r_i \\ &= W_i' \pi + \omega_i \end{aligned}$$

where

$$\begin{aligned} W_i &= (X_i, Z_i) \\ \omega_i &= \alpha_1 (u_{1i} - u_{2i}) + \alpha_2 u_{3i} + \alpha_3 u_{3i} + \alpha_4 u_{4i} \end{aligned}$$

Let σ_ω be the standard deviation of ω_i

They estimate this model in 3 steps as we did in the Roy model section (ignoring the last stage getting the variance/cov matrix)

TABLE 2

College Selection Rules: Probit Analysis

Variable	Reduced Form (16)		Structure (26)		Structure (29)	
	Coefficient	t	Coefficient	t	Coefficient	t
Constant	.0485	.20	.1512	.22	.1030	.17
Background:						
Father's ED	−.0145	−.41	−.0168	−.54	−.0152	−.49
Father's ED^2	.0037	2.05	.0038	2.26	.0037	2.26
DK ED	−.4059	−3.96	−.3924	−2.79	−.4001	−2.91
Manager	.1897	2.17	.1825	2.13	.1871	2.21
Clerk	.0556	.54	.0561	.59	.0554	.59
Foreman	.0182	.19	.0210	.23	.0200	.22
Unskilled	−.0910	−.85	−.0948	−.89	−.0928	−.87
Farmer	−.2039	−2.12	−.2256	−2.27	−.2094	−2.14
DK job	−.0413	−.19	−.0629	−.29	−.0609	−.28
Catholic	−.1144	−1.91	−.0982	−1.51	−.1083	−1.66
Jew	−.0293	−.23	.0143	.12	−.0158	−.14
Old sibs	−.0162	−.93	−.0162	−.93	−.0161	−.93
Young sibs	.0122	.63	.0096	.49	.0112	.57
Mother works:						
Full 5	.1039	.66	.1168	.81	.1104	.76
Part 5	.2179	1.42	.2106	1.52	.2156	1.56

None 5	.0655	.63	.0677	.65	.0661	.64
Full 14	.2898	2.29	.2884	2.30	.2888	2.33
Part 14	.2709	2.20	.2768	2.02	.2693	2.03
None 14	.1980	1.91	.1990	1.92	.1966	1.92
H.S. shop	−.4411	−6.14	−.4397	−3.74	−.4379	−3.90
Ability:						
Read	.0047	1.67	...	...	...	...
NR read	−.2575	−1.41	...	...	...	...
Mech	−.0070	−4.29	...	...	...	...
NR mech	−3.0236	−1.04	...	...	...	...
Math	.0244	12.34	...	...	...	...
NR math	−.7539	−5.75	...	...	...	...
Dext	.0019	.72	...	...	...	...
NR dext	2.2797	.47	...	...	...	...
Earnings:						
ln $(\bar{y}_a/\bar{y}_b)$	...	...	5.1486	2.25	...	...
g_a	...	...	138.3850	1.83	7.6632	.11
g_b	...	...	−44.2697	−1.28	71.8981	2.34
ln $y_a(t)/y_b(t)$	...	...	...	...	5.1501	2.57
Observations	3611		3611		3611	
Limit observations	791		791		791	
Nonlimit observations	2820		2820		2820	
−2 ln (likelihood ratio)	579.5		568.8		576.6	
χ^2 degree freedom	28		23		23	

NOTE.—t is asymptotic t-statistic: DK: Don't know, dummy variable; NR: No response, dummy variable; other variables are defined in Appendix A.

Step 2

Slightly more complicated then before because we have growth as well

$$E\left(\log\left(Y_{ci}\left(0\right)\right) \mid W_i, I_i > 0\right) = X_i'\beta_c + \rho_1 \frac{\phi\left(\frac{X_i'\pi}{\sigma_\omega}\right)}{\Phi\left(\frac{X_i'\pi}{\sigma_\omega}\right)}$$

$$E\left(\log\left(Y_{Hi}\left(0\right)\right) \mid W_i, I_i > 0\right) = X_i'\beta_h + \rho_2 \frac{\phi\left(\frac{X_i'\pi}{\sigma_\omega}\right)}{1 - \Phi\left(\frac{X_i'\pi}{\sigma_\omega}\right)}$$

$$E\left(\frac{\log(Y_{ci}\left(T\right)) - \log\left(Y_{ci}\left(0\right)\right)}{T} \mid W_i, I_i > 0\right) = X_i'\gamma_c + \rho_3 \frac{\phi\left(\frac{X_i'\pi}{\sigma_\omega}\right)}{\Phi\left(\frac{X_i'\pi}{\sigma_\omega}\right)}$$

$$E\left(\frac{\log(Y_{Hi}\left(T\right)) - \log\left(Y_{Hi}\left(0\right)\right)}{T} \mid W_i, I_i < 0\right) = X_i'\gamma_H + \rho_4 \frac{\phi\left(\frac{X_i'\pi}{\sigma_\omega}\right)}{1 - \Phi\left(\frac{X_i'\pi}{\sigma_\omega}\right)}$$

Regressor	Dependent Variable					
	$\ln \bar{y}_a$ (1)	$\ln \bar{y}_b$ (2)	g_a (3)	g_b (4)	$\ln y_a(\bar{t})$ (5)	$\ln y_b(\bar{t})$ (6)
Constant	8.7124	2.8901	.1261	.2517	10.3370	7.5328
	(16.51)	(1.37)	(3.90)	(2.11)	(5.52)	(2.08)
Read	.0009	−.0019	.0001	.0003	.0027	.0057
	(1.21)	(−1.17)	(1.11)	(3.20)	(2.80)	(3.28)
NR read	.0791	.0506	−.0034	−.0046	.0033	−.0402
	(1.24)	(.58)	(−.76)	(−.89)	(.04)	(−.42)
Mech	−.0002	−.0005	−.0001	−.0001	−.0021	−.0017
	(−.48)	(−.54)	(−2.16)	(−1.13)	(−3.59)	(−1.73)
NR mech	...	.1969	...	.0002	...	.2196
		(.69)		(.01)		(.68)
Math	.0015	−.0013	.0001	−.0000	.0030	−.0019
	(2.02)	(.74)	(1.18)	(−.20)	(3.31)	(−1.00)
NR math	−.1087	.0562	.0015	.0006	−.0877	.0712
	(−1.94)	(.83)	(.38)	(.15)	(−1.24)	(.96)
Dext	.0008	−.0019	−.0000	.0003	.0002	.0036
	(1.03)	(−1.21)	(−.78)	(2.77)	(.16)	(2.19)
NR dext	.0751	...	−.0004	...	.1466	...
	(.28)		(−.02)		(.43)	
Exp	−.0523	.4260	−.0028	−.0154	−.0129	.0776
	(−1.49)	(3.10)	(−1.11)	(−1.93)	(−.29)	(.53)
Exp^2	.0015	−.0067	.0000	.0002	−.0000	−.0012
	(2.22)	(−2.95)	(.21)	(1.82)	(−.01)	(−.49)
Year 48	−.0020	−.0156	...	...	...	...
	(−.48)	(−1.72)				
Year 69	...	...	...	...	−.0067	.0039
					(−.26)	(.09)
S13–15	.1288	...	−.0062	...	.0168	...
	(5.15)		(−3.49)		(.52)	
S16	.0760	...	.0026	...	.1095	...
	(3.82)		(1.79)		(4.26)	
S20	.1318	...	.0049	...	.2560	...
	(4.10)		(2.13)		(6.15)	
λ_a	−.1069	...	.0058	...	.0206	...
	(−3.21)		(2.45)		(.49)	
λ_b	...	−.0558	...	.0118	...	.2267
		(−.66)		(2.39)		(2.48)
R^2	.0750	.0439	.1578	.0513	.0740	.0358

Step 3

Finally we go back to the "Structural probit"

$$\log(V_{ci}) - \log(V_{Hi}) \approx \alpha_0 + \alpha_1 \left[X_i'\beta_c - X_i'\beta_H\right] + \alpha_2 X_i'\gamma_c + \alpha_3 X_i'\gamma_H + \alpha_4 Z_i'\delta + \omega_i$$

Standard errors must be corrected

TABLE 2

College Selection Rules: Probit Analysis

Variable	Reduced Form (16)		Structure (26)		Structure (29)	
	Coefficient	t	Coefficient	t	Coefficient	t
Constant	.0485	.20	.1512	.22	.1030	.17
Background:						
Father's ED	−.0145	−.41	−.0168	−.54	−.0152	−.49
Father's ED^2	.0037	2.05	.0038	2.26	.0037	2.26
DK ED	−.4059	−3.96	−.3924	−2.79	−.4001	−2.91
Manager	.1897	2.17	.1825	2.13	.1871	2.21
Clerk	.0556	.54	.0561	.59	.0554	.59
Foreman	.0182	.19	.0210	.23	.0200	.22
Unskilled	−.0910	−.85	−.0948	−.89	−.0928	−.87
Farmer	−.2039	−2.12	−.2256	−2.27	−.2094	−2.14
DK job	−.0413	−.19	−.0629	−.29	−.0609	−.28
Catholic	−.1144	−1.91	−.0982	−1.51	−.1083	−1.66
Jew	−.0293	−.23	.0143	.12	−.0158	−.14
Old sibs	−.0162	−.93	−.0162	−.93	−.0161	−.93
Young sibs	.0122	.63	.0096	.49	.0112	.57
Mother works:						
Full 5	.1039	.66	.1168	.81	.1104	.76
Part 5	.2179	1.42	.2106	1.52	.2156	1.56

None 5	.0655	.63	.0677	.65	.0661	.64
Full 14	.2898	2.29	.2884	2.30	.2888	2.33
Part 14	.2709	2.20	.2768	2.02	.2693	2.03
None 14	.1980	1.91	.1990	1.92	.1966	1.92
H.S. shop	−.4411	−6.14	−.4397	−3.74	−.4379	−3.90
Ability:						
Read	.0047	1.67	...	...	...	...
NR read	−.2575	−1.41	...	...	...	...
Mech	−.0070	−4.29	...	...	...	...
NR mech	−3.0236	−1.04	...	...	...	...
Math	.0244	12.34	...	...	...	...
NR math	−.7539	−5.75	...	...	...	...
Dext	.0019	.72	...	...	...	...
NR dext	2.2797	.47	...	...	...	...
Earnings:						
ln $(\bar{y}_a/\bar{y}_b)$	...	...	5.1486	2.25	...	...
g_a	...	...	138.3850	1.83	7.6632	.11
g_b	...	...	−44.2697	−1.28	71.8981	2.34
ln $y_a(t)/y_b(t)$	...	...	...	...	5.1501	2.57
Observations		3611		3611		3611
Limit observations		791		791		791
Nonlimit observations		2820		2820		2820
−2 ln (likelihood ratio)		579.5		568.8		576.6
χ^2 degree freedom		28		23		23

NOTE.—t is asymptotic t-statistic: DK: Don't know, dummy variable; NR: No response, dummy variable; other variables are defined in Appendix A.