Correcting for Survival Effects in Cross Section Wage Equations Using NBA Data

by

Peter A Groothuis
Professor
Appalachian State University
Boone, NC

and

James Richard Hill
Professor
Central Michigan University
Mt Pleasant, MI

Fall 2008

Abstract: Cross sectional employment data is not random. Individuals who survive to a longer level of tenure tend to have a higher level of productivity than those who exit earlier. This result suggests that in cross sectional data high productivity workers are over-sampled at high levels of tenure. In wage equations using cross sectional data, results could be biased from the over sampling of high productive workers at long levels of tenure. This survival effect in cross sectional data could possibly bias the coefficient on tenure upwards. We explore techniques to correct for survival bias using a panel study of National Basketball Association players. In particular we focus on a modified Heckman selectivity bias procedure using duration models to correct for survival bias.

Introduction

Wages tend to increase with labor market experience. In a cross sectional wage equation the coefficient on experience is generally found to be positive and the coefficient on experience squared is generally found to be negative. These coefficients show that experience has a positive but concave relation to wages. We suggest that the coefficients on experience and experience squared may be biased because in a cross sectional data set the observations are not random. In particular, we suggest that individuals with greater ability have a higher survival rates in occupations than low ability workers that leads to an over-representation of high ability workers at high levels of experience.

In a series of articles using professional sports data increased performance has been shown to lower the probability of exiting a career (Groothuis and Hill 2004, Groothuis and Hill 2008, and Groothuis and Groothuis 2008.) Groothuis and Hill (2004) focusing on career duration in the National Basketball Association (NBA) found that performance and weight of the basketball player determined career length. They found that the race of the player did not matter. Groothuis and Hill (2008) research focusing on Major League Baseball (MLB) found that career length of both pitchers and hitters depended on performance not the race of the player. Groothuis and Groothuis (2008) found that career length also depended on performance, not family status of the driver. In all cases individuals with high performance had longer career duration than low performance individuals.

In this article using NBA data, we explore techniques to correct for survival effects in wage equations. In section one, we provide a literature review of the theories

of the wage tenure profile explaining why wages are expected to rise with labor market experience. In section two, we review a duration model of career length and discuss how the inverse-Mills ratio can be used to correct for survival bias. In section three, we estimate wage equations for NBA players. In this section we estimate wage equations using OLS estimates and a survival-bias corrected model. We conclude with a discussion of possible extensions of the survival bias model.

Section One: Wages Tenure Profile

Three main theories have been developed to explain the positive relationship between wages and experience. The first model is the human capital model. The human capital model suggests that investment in training occurs early in a career and enhances productivity later in a career (Becker 1962). Thus wages rise as productivity rises. If all training is general, wages should equal productivity in all periods. If training is firm-specific, wages should be higher than measured productivity in the first period and lower than measured productivity in the second period as firm specific human capital is a shared investment.

The second model used to explain the wage-tenure profile is the deferred compensation model. In this model, to induce effort, workers are paid lower than productivity wages in the first period and rewarded with wages above productivity later in their careers for their earlier career efforts (Lazear 1979). In this model wages should be lower than productivity in the first period and higher than productivity in the second period.

The third model suggests that wages rise not as a function of training or deferred compensation but because of a better labor market match. In this model the wage tenure relation is based on the idea that workers who last in a career are either better workers (more productive) or have better worker-employer matches (Abraham and Farber 1987). The cross sectional wage equation thus captures job duration effects that are reflected in increased productivity of the sample who remain and not each individual worker.

We suggest a way to test the implication of the three models by controlling for duration effects by using a modified Heckman selectivity bias procedure. In this study, we use a panel data set of all NBA players from the 1989 through the 1999 season for an eleven year panel. In table 1, we report the means of the variables used; in particular we find the average career length was about six years. In addition, we find that the median career length for the same time period was 5 years with the longest career being 21 seasons. In the next section, we review a stock-flow duration model that supports the claim that workers who last in a career are more productive than those who exit earlier.

Career Duration Analysis

To test for survival effects we estimate semi-parametric hazard functions following Berger and Black (1998), Berger, Black, and Scott (2004), Groothuis and Hill (2004), Groothuis and Hill (2008) and Groothuis and Groothuis (2008); since our data is at the season level we calculate our hazard model as a discrete random variable. As with Berger, Black, and Scott (2004), we model the durations of a single spell and assume a homogeneous environment so that the length of the spell is uncorrelated with the calendar time in which the spell begins. This assumption lets us treat all the players' tenure as the

same regardless of when it occurred in the panel study. For instance, all fourth year players are considered to have the same base line hazard regardless of calendar time so a fourth year player in 1990 has the same baseline hazard as a fourth year player in 1997.

To understand how stock data influences a likelihood function we follow the notation of Berger, Black, and Scott (2004). Suppose the probability mass function (pmf) of durations is defined as $f(t,x,\beta)$, where t is the duration of the career, x is a vector of performance and personal characteristics, and β is a vector of parameters. Now denote F(t,x,β) as the cumulative distribution function; then the probability that a career lasts at least t° years is simply 1 - F(t°,x,β). If we define the hazard function as h(t,x,β) ≡ f(t,x,β) / S(t,x,β) where S is the survivor function, $S(t,x,\beta)=\prod_{i=1}^{t-1}[1-h(i,x,\beta)]$ and apply the definition of conditional probabilities, we may express the pmf as

$$f(t_i,x_i,\beta)=\prod_{j=0}^{t_i-1}[1-h(j,x_i,\beta)]h(t_i,x_i,\beta). \qquad (1)$$

If we have a sample of n observations, $\{t_1, t_2, \ldots, t_n\}$, the likelihood function of the sample is

$$L(\beta)=\prod_{i=1}^{n} f\ (t_i,x_i,\beta)=\prod_{i=1}^{n}\left(\prod_{j=1}^{t_i-1}[1-h(j,x_i,\beta)]\,h(t_i,x_i,\beta)\right). \qquad (2)$$

Often it is not possible to observe all careers until they end, hence careers are often right-censored. Let the set A be the set of all observations where the players' careers are completed and the set B be the set of all observations where the careers are right

censored. For the set of right-censored observations, all we know is that the actual length of the career is greater than t_i, the observed length of the career up through the last year. Because we know that the actual length of the career is longer than we observe then the contribution of these observations to the likelihood function is just the survivor function (S).

To introduce stock sampling, let the set C be the set of careers that were in progress when data collection began. For these observations, we know that the career i has lasted for r years before the panel begins so the likelihood must be adjusted by the conditional probability of the career having length r. Of course, some stock-sampled observations may be right-hand censored. Let the set D be the set of all stock-sampled observations that are also right-hand censored. An example of a career that is both right and left censored would be a player that starts his career prior to 1989 and ends his career after 1999. Taking into account all four sets: A, B, C, and D the likelihood function becomes

$$L(\beta) = \prod_{i \in A} \left(\prod_{j=1}^{t_i - 1} [1 - h(j, x_i, \beta)] \, h(t_i, x_i, \beta) \right) \times \prod_{i \in B} \left(\prod_{j=1}^{t_i - 1} [1 - h(j, x_i, \beta)] \right) \times \prod_{i \in C} \left(\prod_{j=r_i}^{t_i - 1} [1 - h(j, x_i, \beta)] \right) h(t_i, x_i, \beta) \times \prod_{i \in D} \left(\prod_{j=r_i}^{t_i - 1} [1 - h(j, x_i, \beta)] \right) \tag{3}$$

In equation (3) the contribution of censored, stock-sampled observations to the likelihood function is strictly from the last two terms; such observations simply provide information about the survivor function between (r,t).

Thus we, as Berger, Black and Scott (2004), have expressed the likelihood function as a function of the hazard functions. All that remains is to specify the form of a hazard function and estimate by means of maximum likelihood estimation. As the hazard function is the conditional probability of exiting NBA given that the NBA career lasted

until the previous season, the hazard function must have a range from zero to one. In principle, any mapping with a range from zero to one will work. For our purposes we choose the probit model. We choose the probit model over the more traditional logit model because the probit model provides the benefit of the ability to calculate the inverse-Mills ratio. We will review the implication of the inverse-Mills ratio in the next section.

The intuition behind the probit model for the hazard function is relatively simple. For each year during the survey in which the player is in NBA, the player either comes back for another season or ends his career. If the career ends, the dependent variable takes on a value of one; otherwise, the dependent variable is zero. The player remains in the panel until the player exits NBA or the panel ends. If the panel ends, we say the worker's spell is right-hand censored. Thus a player who begins his NBA career during the panel and plays for 6 years will enter the data set 6 times: the value of his dependent variable will be zero for the first 5 years (tenure one through five) and be equal to one for the sixth year.

To illustrate a stock sample consider another player who enters the panel with 7 years of NBA job tenure prior to 1990 the first year of the panel, then plays for an additional 3 years for a 10 year career. For this player we ignore his first 7 years of tenure because he is left-hand censored. As the equation of the likelihood function with stock data indicates, the duration of a NBA career prior to the beginning of the panel makes no contribution to the value of the likelihood function. Therefore only years 8 through 10 will enter the data set with the dependent variable taking on the value zero for years 8 and 9 and in the 10^{th} year it takes on a value of one with this player appearing in

the data set a total of 3 times. Note for all players who are right-hand censored, we do not know when their career ends so their dependent variables are always coded as zero.

Because the players in the panel have varying degrees of job tenure prior to the beginning of the panel, we identify the hazard function for both long and short careers. The disadvantage to this approach is that the vector γ_t of equation (3) can be very large. In our study it would require 21 dummy variables. We also run into problems with the dummy variable technique because we have too few players who have long careers. To simplify the computation of the likelihood function and be able to keep the long careers, we simply approximate the γ_t vector with a 4^{th} order polynomial of the players' tenure in NBA, which reduces the number of parameters to be estimated from 21 to 4. Thus, the hazard function becomes

$$\Pr(t, x\beta)=\Pr(\phi(t)+ x\beta), \qquad (4)$$

where $\phi(t)$ is a 4^{th} order polynomial of the player's tenure in NBA. The 4^{th} order polynomial therefore includes tenure to the first, second, third, fourth and fifth powers. Once again, we choose the Taylor series approximation technique over using tenure dummies due to the small number of observations for high tenures.[1]

In table 2 we report the results of equation 6 for NBA players. We find that points per minutes, assists per minutes, blocked shots per minute, games per season, weight of player, and higher draft position all lower the probability of exiting a career. These results show that performance lengthens the careers of NBA players suggesting

[1] When higher order polynomials of the fifth and sixth power are included results do not change suggesting that a fourth order polynomial is flexible enough to capture the influence of the base line hazard

that higher performance players are over represented latter in careers. In the next section we explore if survival bias effects wage equations.

Duration effects and Sample Selection Bias

The Heckman (1979) procedure has long been used by economist to control for self selection in a sample. In our case, we suggest that performance increases the likelihood of survival in the NBA. We explore if a modified Heckman procedure controls for survival bias in wage equations. Following Green's (1998) notation consider the basic sample selection model where,

$$d_i^* = z'\gamma + \upsilon \text{ and } d_i = 1 \text{ if } d_i^* > 0;\ d_i = 0 \text{ otherwise,} \qquad (5)$$

$$y_i = x'\beta + \varepsilon_i \qquad (6),$$

and y is only observed if d_i is equal to one. Suppose as well that υ and ε have a bivariate normal distribution with zero means and a correlation of ρ. Then

$$E(y|\ y \text{ is observed}) = E(y|\ d^* > 0) \qquad (7)$$

$$= E(y|\ \upsilon > -z'\gamma) \qquad (8)$$

$$= x'\beta + E(\varepsilon|\ \upsilon > -z'\gamma) \qquad (9)$$

$$= x'\beta + \rho\sigma_\varepsilon \lambda(\alpha_\upsilon) \qquad (10)$$

$$\text{or} \qquad = x'\beta + \beta_\lambda \lambda(\alpha_\upsilon) \qquad (11)$$

Where $\lambda(\alpha_\upsilon)$ is the inverse Mill's ratio. The overall equation of interest becomes

$$y_i|d^* > 0 = x'\beta + \rho\sigma_\varepsilon \lambda(\alpha_\upsilon) + \varepsilon_i, \qquad (12)$$

Thus the marginal effect of a regressor in the observed sample consists of two parts

$$\frac{\partial E(y|\ y \text{ is observed})}{\partial x_k} = \beta_k + \gamma_k (\rho\sigma_\varepsilon / \sigma_\upsilon)\lambda(\alpha_\upsilon) \qquad (13)$$

where the first part β is the marginal effect of the observed equation and γ is the influence the variable has on the likelihood of being observed. Heckman (1979) recommends a two step procedure: first estimate a probit on the selection equation to obtain estimates of γ. Then estimate the OLS regression along with the inverse Mill's ratio to estimate β and β_λ.

In our case the selection equation comes from equation 5 from the above career duration model. The probit estimates the probability of exiting from a career so instead of estimating the probability of being in the sample, as in Heckman's approach, we estimate the probability of exiting the sample. Equation 6 in our analysis is a cross sectional wage equation where the data on workers are selected into the sample with higher skilled workers remaining in the sample while low skill workers are more likely to leave.

In table 3, we report the results of the log wage equation for NBA players. In column one, we report the results of the log wage equation using OLS. In this model we find that tenure and tenure squared follow the traditional relation. In addition, we find that improved performance and height increase wages. In particular, increases in points per minute, assists per minute, and blocks per minute all increase the wage rate. We also find that higher draft numbers correspond with lower wages and the more games played per season correspond with higher wages. In column two, we correct for survival bias using the inverse Mill's ratio as an additional independent variable to correct for survival bias. The results show that the coefficient on the inverse Mill's ratio is positive and

significant. The results suggest that survival bias exists for players in the NBA confirming that high quality players are over sampled at long levels of tenure.

A closer look at the results show that the coefficient on tenure is positive while tenure squared is negative and changes slightly from .240 to .234 after controlling for survival bias suggesting a slightly lower influence of tenure on wages at the magnitude of a 2.5 percent. This result is consistent with the Abraham and Farber's (1987) model that suggest that the tenure variable captures survival effects. Survival bias, however, is relatively modest with tenure still having a large impact on wages. In addition Becker's (1962) human capital model is not supported. It is not supported because in the wage equation tenure and tenure squared does not change when including the performance measures in the equation.

Conclusion

Cross sectional employment data is not random. Individuals who survive to a longer level of tenure tend to have a higher level of productivity then those who exit earlier. These results suggest that in cross sectional data high productivity workers are over-sampled at high levels of tenure. To correct for survival bias we use a modified Heckman technique. We find that in a log wage equation of NBA players that survival bias is present and influences the coefficients on tenure but does not influence the coefficients on tenure squared or performance measures. Future research will attempt to measure survival bias in a panel study instead of a cross sectional wage equation.

TABLE 1
Means

Variable	Means (standard deviation)
Career length	5.78 (3.89)
White	.21 (0.41)
Height	79.2 (4.1)
Weight	221.9 (29.6)
Games per season	57.0 (24.2))
Points per minute	.385 (.135)
Assists per minute	.087 (.062)
Rebounds per minute	.178 (.085)
Steals per minute	.033 (.018)
Turnovers per Minute	.074 (0.28)
Blocks per minute	.021 (0.22)
Draft Number	24.3 (23.4)
Number observations Number of players	3886 1113
Salary 1998	$3,737,229 (3,436,466)
Number of players	295

TABLE 2

NBA Career Duration Semi-parametric Analysis
1989-1999

Variable	Probit Model
Constant	1.75 (4.29)
Weight	-.007 (4.50)
Games per season	-.026 (20.49))
Points per minute	-.410 (2.04)
Assists per minute	-2.11 (3.23)
Rebounds per minute	-.475 (1.09)
Steals per minute	-1.59 (1.07)
Turnovers per Minute	.012 (0.16)
Blocks per minute	-.548 (0.34)
Draft Number	.003 (2.31)
Chi-squared	692.30
Number observations	3889

--First through fourth order tenure polynomials are also included to provide for general functional form of baseline hazard. They are jointly significant.
--The numbers in parentheses are absolute value of t-ratios.

Table 3
Log Wages 1998

Variable	**OLS**	**OLS**
Constant	7.78 (5.11)	8.50 (5.51)
Tenure	.240 (6.02)	.234 (5.88)
Tenure Squared	-.013 (4.96)	-.012 (4.60)
Games per season	.020 (5.54)	.016 (3.96)
Points per minute	2.59 (6.19)	2.60 (6.27)
Assists per minute	5.20 (4.70)	4.97 (4.50)
Rebounds per minute	.77 (0.98)	.46 (0.56)
Steals per minute	-2.12 (0.68)	-2.24 (0.72)
Turnovers per Minute	.010 (0.14)	.029 (0.39)
Blocks per minute	7.88 (2.62)	7.67 (2.94)
Draft Number	-.010 (3.80)	-.009 (3.96)
Height	.050 (2.70)	.045 (2.46)
Inverse Mills Ratio		.497 (2.31)
R^2	.46	.47
Number observations	295	295

References

Abraham, Katherine G. amd Henery Farber, 1987, "Job Duration, Seniority, and Earnings," The American Economic Review, Vol. 77 No. 3 pp.287-297.

Becker, Gary, 1962. "Investment in Human Capital: A Theoretical Analysis," Journal of Political Economy, vol.70, no.5 pp. 9-49.

Berger, Mark C. and Dan A. Black, 1999. "The Duration of Medicaid Spells: An Analysis Using Flow and Stock Samples," The Review of Economics and Statistics, vol.80, no.4, pp. 667-674.

Berger, Mark C., Dan A. Black, and Frank Scott, 2005. "Is There Job Lock? Evidence from the Pre-HIPAA Era," Southern Economic Journal, vol.70 no.4 pp.953-976

Groothuis, Peter A. and Jana Groothuis "Nepotism or Family Tradition: A Study of NASCAR Drivers, Journal of Sports Economics, forthcoming.

Groothuis, Peter A. and J. Richard Hill, 2004. "Exit Discrimination in the NBA: A Duration Analysis of Career Length," Economic Inquiry, vol.42 no.2. pp. 341-349.

Groothuis, Peter A. and J. Richard Hill, 2008. "Exit Discrimination in Major League Baseball: 1990-2004, Southern Economic Journal, (75) 2: 574:590.

Heckman, J., 1979. "Sample Selection Bias as a Specification Error," Econometrica, (47), pp. 153-61.

Hill, James Richard, 2004. "Pay Discrimination in the NBA Revisited," Quarterly Journal of Business and Economics, Vol.43, Nos. 1&2, 81-92.

Lazear, Edward, 1979. "Why is There Mandatory Retirement?" Journal of Political Economy, Vol. 87:1261-1284.