The dynamics of patent citations

Alan C. Marco[*]

July 3, 2006

Abstract

The use of patent citations as a measure of patent "quality" increased dramatically in recent years. I estimate the hazard of patent citation, and find evidence of unobserved heterogeneity. Hazard estimation provides a means to separate patent quality from citation "inflation."

JEL: C1, O3.

Keywords: patents, patent citations, hazard rate, duration.

[*] Assistant Professor, Department of Economics, Vassar College, 124 Raymond Ave. #592, Poughkeepsie, NY 12604, marco@vassar.edu. I would like to thank the editor and the referee for helpful comments.

Vassar College Economics Working Paper # 84

1 Introduction

In recent years the use of patent citations has proliferated in the economic literature. Applications abound in estimations of patent value, firm value, innovative performance, and strategic behavior. This is due, in part, to the recent availability of NBER patent citation data (Hall, Jaffe and Trajtenberg 2001). That project was undertaken because of a recognition that simple patent counts are noisy measures of innovative output (Trajtenberg 1990).

A patent citation is very similar to a bibliographic citation: an innovation may be partly based on an earlier patented innovation. The inventor is required to disclose all "prior art" related to the patented invention. Particularly novel patented innovations will be the subject of greater citation. For this reason, the number of citations received by a patent (forward citations) has been used in the literature as a measure of the innovative output embodied in the technology.

Three applications of patent citations dominate the innovation literature in economics: measuring patent "quality" (Trajtenberg 1990, Hagedoorn and Cloodt 2003); measuring knowledge flows and spillovers (Jaffe, Trajtenberg and Henderson 1993, Moretti 2004); and investigating strategic behavior by firms (Podolny, Stuart and Hannan 1996, Lanjouw and Schankerman 2001, Marco 2005). These applications impute some real economic value to patent citations: they signal either patent quality (value), or some transfer of knowledge from one party to another.

This note investigates patent citations using parametric and non-parametric hazard estimation. Hazard estimation is valuable in this context because the distribution of patent citations is truncated due to unobserved future citation behavior. Estimates show that there is unobserved heterogeneity in citation rates among patents. The results highlight the need for researchers to understand the consequences of dynamics in relying on citation measures, especially in cross-sectional estimation.

2 Data

Data for the empirical analysis come from the NBER Patent Citation Datafile, described in great detail by Hall et al. (2001). I use a random sample of 20,000 patents issued between 1975 and 1995, including information on backward citations and forward citations.

Backward citations are citations made by a patent to previously issued patents. Forward citations are citations received by a patent from subsequently issued patents. In contrast to backward citations, the number of forward citations changes over time, even beyond the patent expiration. I normalize the number of forward citations at a given date by the patent's age to separate the effects of age and prior citations on the citation hazard rate (citation rate). The heterogeneity results below are robust to different measures of forward citations.

Forward citation counts present a problem in cross-sectional research. Younger patents are bound to have fewer citations than otherwise identical older patents; that is, the distribution of forward citations is truncated as described by Hall et al. (2001). The truncated distribution must be accounted for in some way. Additionally, there is evidence of *citation inflation:* the propensity to cite may be increasing for reasons unrelated to patent quality. The inflation may lead to time-inconsistency in equating citations to patent quality (Hall et al. 2001). This note investigates the distribution of forward citation using a hazard model, which corrects for truncation. Additionally, I make recommendations for using hazard estimation to account for citation inflation.

Control variables include the number of patent claims (explained below), age, the year of patent grant, and indicator variables for technology and type of patent assignee. I include two time dummies to account for changes in the patent system in 1982-1984. Only the forward citation count and age vary over time. Because I use time varying covariates, each patent has multiple observations: one at birth and one for each forward citation. In total, there are over 105,000 citations to the 20,000 patents.

3 Analysis

In discrete terms, the hazard function, $h(t)$, is the probability that a patent will receive a citation given that it has survived for t years since its last citation. It is defined as $h(t) = \frac{f(t)}{1-F(t)}$, where $f(t)$ and $F(t)$ are the probability density and cumulative probability functions for the random variable τ, the time until failure (citation).

I estimate the hazard function using a Weibull distribution because of the flexibility of the functional form:[1]

$$h(t, X) = \lambda\rho(\lambda t)^{\rho-1} \tag{1}$$

where the parameter λ is defined as

$$\lambda = e^{X\beta+\varepsilon}. \tag{2}$$

The parameters ρ and β are estimated using data X. In the Weibull specification, ρ indicates *duration dependence.* If $\rho = 1$, the Weibull reduces to an exponential distribution in which the hazard is constant over time. Positive duration dependence occurs when $\rho > 1$ and indicates that the citation hazard rate increases as the patent ages. $\rho < 1$ indicates negative duration dependence.

I test for unobserved heterogeneity by estimating a frailty model. The hazard function for observation j for patent n is specified as

$$h(t_{nj}|X_{nj}, \alpha_n) = \alpha_n h(t_{nj}|X_{nj}), \tag{3}$$

where α_n (the "frailty") follows a gamma distribution with mean one and variance θ (the degree of heterogeneity). $\theta = 0$ implies no unobserved heterogeneity—the standard Weibull hazard model. The null hypothesis of $\theta = 0$ can be tested using a Likelihood Ratio test.

The frailty model is essentially a random effects model for hazard estimation. If there exists unobserved heterogeneity among patents—making some more prone to citation than others based on unobservable characteristics—then an interesting result can occur. One can observe a *decreasing* duration dependence in the population, even when individual-level duration dependence is *increasing*.[2]

Estimation of Equation 3 proceeds via maximum likelihood, with censored observations incorporated much like the Tobit model (Greene 1993). The log-likelihood function is

$$\ln L = \sum_{uncensored} h\left(t|\beta, \rho, \theta\right) + \sum_{all} \ln\left(1 - F\left(t|\beta, \rho, \theta\right)\right). \tag{4}$$

The first column of Table 1 reports the standard hazard estimation using a Weibull distribution; the second column estimates a frailty model. Coefficients are expressed as hazard ratios, where a value above one indicates a positive impact on the hazard rate.

The current number of forward citations (normalized by age) tends to increase the hazard rate of receiving a forward citation. This potentially accounts for some heterogeneity because forward citations may beget forward citations. The number of claims and the number of backward citations also increase the hazard rate. Backward citations account for patenting areas where the propensity to cite is higher, and more claims may represent patents with broader scope (Lanjouw and Schankerman 2001). As expected, patent age has a negative impact on the citation rate and citation rates differ significantly across technology classes.

Three results merit emphasis. First, the significance of θ indicates the presence of unobserved heterogeneity: there is positive duration dependence for individual patents while the population rate exhibits negative duration dependence (Figure 1). That is, individual patents have increasing hazard rates, but the frailty effect generates a lognormal-like distribution for the population. That there exists unobserved heterogeneity should not come as a surprise to scholars who use patent citations as a measure for patent quality. If quality is otherwise unobservable, then one should expect θ to be significant if quality and citations are correlated.[3]

Second, even though age has a negative impact on the hazard rate, individual hazard rates are hump-shaped in age (Figure 2). Importantly, the citation frequency in the population falls after three years whereas the estimated citation hazard rate increases for the first seven years.[4] Again, this is a consequence of the frailty effect.

Lastly, the parameter coefficients on the technology dummies in Table 1 suggest that there are differences in the hazard rate across technologies; non-parametric estimates confirm this. Figure 3 shows the Kaplan-Meier estimated hazard rates for different technology groups. While the overall shape of the hazards is similar, the first panel shows that citation rates for Medical patents and Computer and Communication patents differ from those for other technology classes. Even within classes, there are differences among subclasses (e.g., Medical patents in panel two).

4 Conclusion

This note shows that there is unobserved heterogeneity in the rate of patent citation, and that individual hazards exhibit positive duration dependence despite the average hump-shaped hazard rate over the age of the patent. The results highlight the need for care in interpreting the marginal effects of forward citations on other variables of interest, such as firm value. At the least they must be interpreted in light of the fact that the citation rate changes over the life of a patent; at best, one would want to model citation dynamics. An alternative would be to include as regressors those factors that influence the rate of forward citations, e.g., the industry or technology group, the year of patent grant, or the observed life of the patent. That approach has two limitations. First, it does not account for the clear non-linearity in citation rates. Second, it does not allow the researcher to separately identify direct effects of patent citations from the independent effects of other regressors, such as industry or technology.

Researchers have expressed concern that citation inflation may bias the estimated effect of citations on other variables of interest. Indeed, it is difficult to separate the effects of inflation from those of patent quality. If researchers are able to ascertain those factors that are correlated to citation inflation rather than quality, then hazard estimation suggests the rendering of new metrics for patent quality.

By estimating a hazard rate based only on inflationary factors, residuals can be used to measure latent patent quality. For example, the difference between actual citation lags and predicted citation lags may be a better measure of patent quality than simple citation counts. Alternatively, the ratio of observed citations to predicted citations may represent a proxy of patent quality. The advantage of such constructs is that they can be calculated as time-invariant values that are not subject to censoring.

Of course, the proposed methods for measuring patent quality are predicated upon a correlation between unobserved heterogeneity in patent citation and patent quality. This is an area left for future research.

Notes

[1] I tested the Weibull distribution against exponential and lognormal distributions. The Weibull performed better on a range of criteria including the Akaike Infomation Criterion (AIC).

[2] In mortality studies, some subjects may be more "frail" than others. Frail patients die early, leaving a more robust population alive. If the frailty is unobservable *ex ante*, then there will be an apparent decrease in mortality over analysis time. It is a type of fallacy of composition. In the patent context, highly cited patents drop out early and the clock on analysis time (duration) is reset. As the duration of a spell increases, weakly cited patents remain, leading to an apparently lower citation rate.

[3] This should not be taken to mean that evidence of unobserved heterogeneity implies that patents do—in fact—measure quality; only that the two are consistent.

[4] Adding age^2 does not impact this result.

References

Greene, William H., *Econometric Analysis, 2nd edition*, New York: MacMillan, 1993.

Hagedoorn, John and Myriam Cloodt, "Measuring Innovative Performance: Is There an Advantage in Using Multiple Indicators?," *Research Policy*, September 2003, *32* (8), 1365–79.

Hall, Bronwyn, Adam B. Jaffe, and Manuel Trajtenberg, "The NBER Patent Citation File: Lessons, Insights and Methodological Tools," *NBER Working Paper Number 8498*, October 2001.

Jaffe, Adam B., Manuel Trajtenberg, and Rebecca Henderson, "Geographic localization of knowledge spillovers as evidenced by patent citations," *Quarterly Journal of Economics*, 1993, *108* (3), 577–598.

Lanjouw, Jean O. and Mark Schankerman, "Characteristics of Patent Litigation: A Window on Competition," *RAND Journal of Economics*, Spring 2001, *32* (1), 129–151.

Marco, Alan C., "The Option Value of Patent Litigation: Theory and Evidence," *Review of Financial Economics*, 2005, *14* (3-4), 323–351.

Moretti, Enrico, "Workers' Education, Spillovers, and Productivity: Evidence from Plant-Level Production Functions," *American Economic Review*, June 2004, *94* (3), 656–90.

Podolny, Joel M., Toby E. Stuart, and Michael T. Hannan, "Networks, knowledge, and niches: Competition in the worldwide semiconductor industry, 1984-1991," *American Journal of Sociology*, November 1996, *102* (3), 659–689.

Trajtenberg, Manuel, "A Penny for Your Quotes: Patent Citations and the Value of Innovations," *RAND Journal of Economics*, Spring 1990, *21* (1), 172–87.

Table 1: Forward Citation Rate: Hazard Estimation

Variable	Weibull	Frailty
Claims	1.004 (0.000) *	1.002 (0.000) *
Forward/Age	1.290 (0.002) *	1.546 (0.006) *
Age	0.659 (0.001) *	0.670 (0.001) *
Citations made	1.004 (0.000) *	1.004 (0.001) *
Foreign	0.956 (0.007) *	0.970 (0.008) *
Individual	1.014 (0.030)	1.013 (0.038)
Grant year	1.062 (0.001) *	1.073 (0.002) *
Pre-1982 grant	1.510 (0.019) *	1.663 (0.026) *
Post-1984 grant	0.793 (0.010) *	0.730 (0.012) *
Chemicals	1.079 (0.011) *	1.043 (0.013) *
Comp./Comm.	1.267 (0.014) *	1.154 (0.017) *
Medical	1.232 (0.015) *	1.155 (0.019) *
Electronics	1.049 (0.011) *	1.039 (0.014) *
Mechanical	0.994 (0.010)	0.984 (0.012)
p, Weibull parameter	3.606 (0.013) *	3.492 (0.008) *
θ, heterogeneity parameter	--	0.063 (0.002) *
Log-likelihood	106,305	109,549
LR test $\chi^2(14)$	128,660	76,240

Notes:

105,327 citations for 20,000 patents.

Dependent variable is duration until failure. All specifications assume the Weibull distribution.

Coefficients expressed as hazard ratios. Standard errors in parentheses. * significant at 1%.

Likelihood-ratio test of $\theta = 0$: $\chi^2(1) = 6487$.

Figure 1: Estimated Forward Citation Hazard Rate

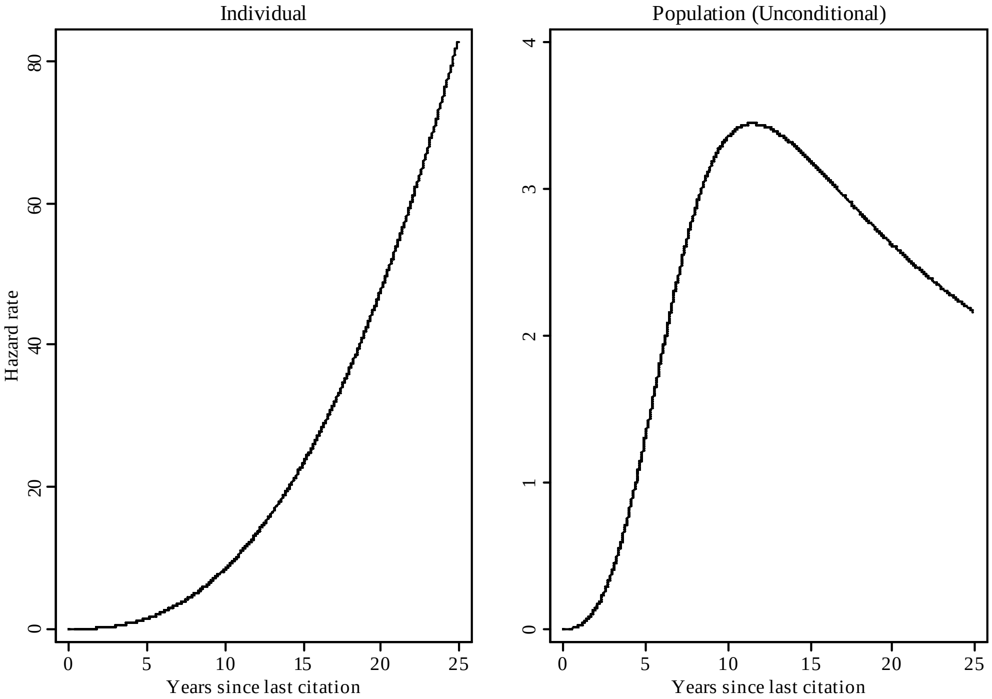


Figure 2: Forward Citations by Age of Patent

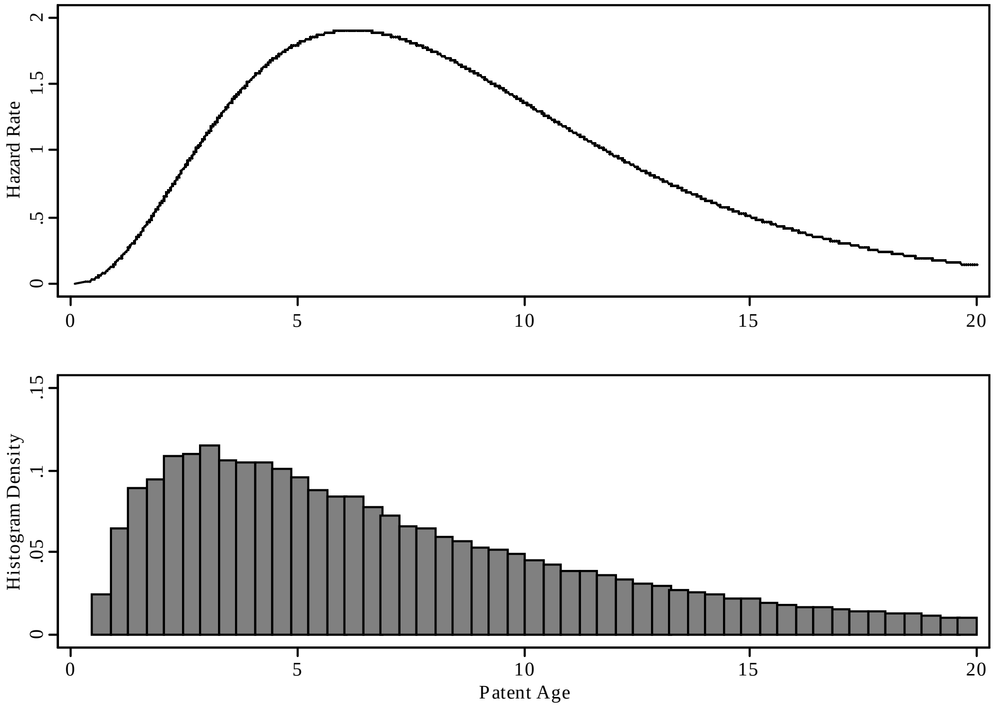
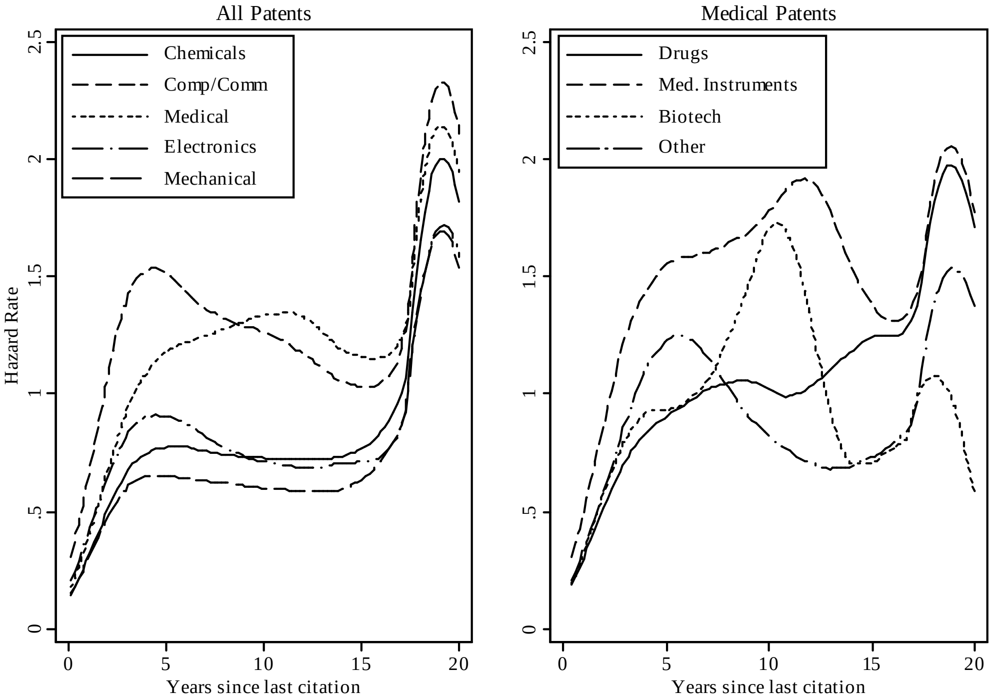


Figure 3: Non-parametric Hazard Estimates


All Patents
Medical Patents
Hazard Rate
Chemicals
Comp/Comm
Medical
Electronics
Mechanical
Drugs
Med. Instruments
Biotech
Other
Years since last citation
Years since last citation