Measuring the Payoff from Federally-Funded Research and Development:   
Exploring Approaches to Improving Construct Validity in Studies of Technology Transfer

Malcolm S. Townes

Saint Louis University

Abstract

{ENTER TEXT}

Keywords: technology transfer, technology commercialization, research and development, public policy, patent citations

**Introduction**

Improving the transfer of technologies derived from federally-funded research and development (R&D), so called technology transfer, is arguably one of the highest public policy priorities of the United States of America (U.S. or USA). It was listed as a top priority in the President’s Management Agendas (PMAs) of both the George W. Bush and Donald J. Trump Administrations (Office of Management and Budget [OMB], 2002; OMB, 2018). While the administration of President Barack H. Obama did not issue PMAs, President Obama did issue a presidential memorandum on October 28, 2011 that explicitly focused on technology transfer and commercialization of federal research. In the policy section of this memorandum, he referenced the Startup America initiative which had as one of its objectives “increasing the rate of technology transfer and the economic and societal impact from Federal research and development (R&D) investments” (Daily Comp. Pres. Doc., 2011-October-28). In fact, technology transfer has been a topic of interest to the U.S. federal government since at least the end of World War II (Bush, 1945).

Technology transfer public policy is also a top public policy priority because of the link between national economic prosperity and technological innovation. Solow (1957) estimated that roughly 88 percent of the total increase in real Gross National Product (GNP) was attributable to technological progress. Consequently, it’s important for the nation to maintain its technological prowess to continue the way of life that citizens and residents of the country have come to expect.

From a more pragmatic standpoint, the efficient use of scarce national resources makes technology transfer public policy an important issue for examination. In fiscal year 2018, the U.S. federal budget for total R&D was greater than $142.9 billion (American Association for the Advancement of Science [AAAS], 2018a). Although this represented less than 3.7 percent of the governments $3.9 trillion in total federal outlays (Congressional Budget Office [CBO], 2018), it is not a triviality considering that the amount is greater than the gross domestic product (GDP) of at least 110 countries (United Nations [UN], 2017). Moreover, the U.S. budget deficit for fiscal 2019 was more than $100 billion (U.S. Department of the Treasury, 2018b) and the U.S. total public debt as of October 31, 2018 was more than $21.7 trillion (U.S. Department of the Treasury, 2018a). In this context, making every dollar count is imperative. There are other important problems of national interest to which the government could direct monies currently being spent on R&D such as road repairs, alleviating hunger, and addressing issues with inequity in the court system. As Figure 1 shows, federal R&D expenditures is equivalent to roughly 20 percent of the federal budget deficit and exceeds federal spending on transportation, the Supplemental Nutrition Assistance Program (SNAP), and law courts (U.S. Spending, n.d.). As such, it’s important to ensure that technology transfer public policy is as optimized as possible.

How we conceive and operationalize the construct of technology transfer significantly influences how we measure the payoff from federally-funded R&D and formulate public policy regarding federal R&D initiatives and technology transfer. It therefore stands to reason that improving construct validity in studies of technology transfer will provide new and useful insights that can illuminate opportunities to increase the benefits from federally-funded R&D that accrue to society. Consequently, the purpose of this study is to explore an alternative approach to measuring the payoff from federally-funded R&D, investigate how such an alternative measurement approach modifies our understanding of the technology transfer process, and consider the public policy implications of this new insight about technology transfer.

**Literature Review**

Technology transfer is a concept for which there is no official definition. While most studies of the topic don’t explicitly define technology transfer, they generally seem to operationalize it as a financially-based exchange (Gonzalez-Perni, Kuechle & Pena-Legzkue, 2013; Hallam, Wurth & Mancha, 2014; Markman, Gianiodis & Phan, 2009). However, the operationalization of the construct in these studies generally seems to conflate the concept of technology transfer with the mechanisms for achieving it. Licensing, new venture formation, research collaboration, and faculty consulting are generally used as indicators of technology transfer.

The difficulties encountered with defining and operationalizing the construct of technology transfer are exacerbated by challenges defining what constitutes a technology. There is no universally accepted definition of technology. Again, it seems that most studies don’t bother to define technology. Generally, academic research related to technology transfer seems to conceive of a technology as a patent right to a government recognized invention (Markman, Gianiodis & Phan, 2005). However, this fails to recognize that patentable subject matter is defined by law, which varies from country to country, and is not a universal phenomenon. As such, not all technology is patentable. Moreover, what is patentable may not necessarily constitute a technology. Some studies seem to broaden the idea of technology to include academic knowledge (Gonzalez-Pernia, Kuechle & Pena-Legazkue, 2013). This seems to recognize that technology is not the only benefit that is derived from research and development. However, an argument can be made that academic knowledge and technology are not necessarily synonymous.

Defining success in the context of technology transfer has been problematic for scholarly studies of the subject. Most research studies seem to select indicators and measures more for convenience rather than to maximize construct validity. Executed patent licenses, established new business entities, and executed sponsored research agreements have all been used as proxies for technology transfer (Gonzalez-Perni, Kuechle & Pena-Legzkue, 2013; Hallam, Wurth & Mancha, 2014; Markman, Gianiodis & Phan, 2009). As previously discussed, these are all financially-based definitions of success. Additionally, there is the risk of mis-categorizing or double counting activities depending on how these measures are used. For example, a patent license is often associated with the formation of a university spinout company (i.e., new business venture to commercialize technology developed at a university). In such situations, using both measures would essentially double count a single instance of technology transfer. Sponsored research may not be related to technology previously developed at the university from federally-fund research. As such, it may be misleading to consider all sponsored research as instances of successful technology transfer. Additionally, these measures of technology transfer don’t accommodate instances that are not financially-based exchanges. Theoretically, technology transfer can occur in the absence of a financial transaction.

Some studies have used allowed patents as a measure of technology transfer success. However, just because a patent is allowed doesn’t mean that it is used to benefit society.

Yoshikane (2013) specifically studied the citation frequency of patents. Yoshikane found that the number of classifications tied to a patent was positively associated with citation frequency. This seems to suggest that the more general a patent the more likely that it will be cited by other patents.

Various studies have used regression analysis in their investigations of technology transfer. According to Licht (1995), the two primary uses of multiple regression analysis in studies are to either predict phenomenon for decision-making purposes or understand and explain the nature of phenomenon to develop or test theories. Studies of technology transfer have used various regression analysis methods to understand and explain the process. These studies demonstrate that regression analysis is a useful method for gaining insight into the factors associated with technology transfer success. Yoshikane (2013) used multiple linear, logistic, and binomial regression analyses to study patent citation data. Appio, Martini & Fantoni (2017) used a series of logistic regression models to explore the role of scientific and technological diversity in developing impactful bioinformatics inventions as measured by forward citation distribution. They found that different degrees of knowledge diversity were associated with different degrees of impact but combinations of scientific and technological knowledge diversity did not always lead to impactful inventions as defined in the study. Kirkman (2013) used multinomial logistic regression to understand how universities use technology transfer to disseminate research discoveries to biotechnology firms. The study found that the innovativeness, proactiveness, and risk taking propensity of biotechnology firms influenced their selection of technology transfer modes. Kirkman specifically limited the modes of technology transfer in the study to licensing, sponsored research, and consulting agreements, which are all financially-based exchanges.

**Research Questions and Expected Results**

**Research Questions**

Assuming the generation of a technology is not the only benefit derived from research and development, it’s reasonable to conclude that the objective of public policy regarding federally-funded R&D should be to maximize all types of benefits derived from R&D efforts. In this context, the transfer of technologies derived from federally-funded R&D to the private sector (i.e., technology transfer) is only one type of benefit. As the literature reveals, there can be other types of benefits such as the transfer of knowledge derived from federally-funded R&D. This is context in which this study is conducted.

The purpose of this study is to investigate important issues regarding federally-funded R&D and technology transfer that remain unanswered or underexplored including how success should be defined, how outcomes should be measured, and what drives desirable outcomes. Specifically, this study investigates the following questions:

1. Aside from technology, what are the benefits derived from federally-funded R&D?
2. How do we measure non-technology benefits derived from federally-funded R&D?
3. Can patent citations be used to evaluate the benefits derived from federally-funded R&D?
4. What insights can be derived about the factors that drive benefits derived from federally-funded R&D based on non-financially-based measures of success?

Most studies in this area assume that the generation of technologies is the primary output of R&D and focus only on the transfer of technologies derived from R&D to the private sector (i.e., technology transfer). Moreover, most studies of technology transfer seem to define it in terms of financially-based exchanges such as the executing license agreements, securing sponsored research, or forming new business ventures. This study defines technology transfer more broadly to include other R&D outputs, specifically new knowledge. Patents allowed by the United States Patent and Trademark Office (USPTO) and the claims of those patents are used as a proxy for new knowledge derived from research and development. Citations of patents are used as a proxy for the transfer of that knowledge.

This study makes several hypotheses about the factors that are associated with the transfer of new knowledge derived from research and development. One hypothesis is that the number of claims made by a patent and the generality of the patent will both have a positive association with the probability that a patent is cited by other patents (i.e., citations received) and the number of citations received. It is theorized that the more claims a patent contains creates significantly more opportunities for that patent to be cited by other patents. It is also theorized that the more general a patent, the more opportunities there are for that patent to be cited by other patents across multiple fields. The originality of the patent and the year a patent was granted are expected to be negatively associated with the probability that the patent is cited by other patents and the number of citations received by the patent. It is theorized that the more original a patent the more difficult it is for other inventors and innovators to conceive applications of the technology in their fields. It is also theorized that the older a patent is the less relevant it becomes because of the half-life of knowledge. Machlup (1962) is credited with proposing the concept of half-life of knowledge, which can be thought of as the time it takes for half of the knowledge in a field to be rendered irrelevant. Machlup proposed that although patents provided several years of protection and exclusivity (17 years at the time) obsolescence reduced the practical duration of this protection to no more than a few years.

**Data and Methods**

**Data Sources**

This study uses a subset of 2,000 observations taken from patent data obtained from the National Bureau of Economic Research (NBER) website. The source data contains both original and constructed variables. The data file included all utility patents granted in the U.S. from January 1, 1963 to December 30, 1999 listed in the Technology Assessment and Forecast (TAF) database of the USPTO. The source file contained data on 2,923,922 patents across 23 variables.

Table 1 details the original USPTO variables of the source data and explanations of their meanings. PATENT indicates the number assigned by the USPTO to the allowed patent. GYEAR is the year the USPTO allowed the patent. APPYEAR is the year the patent application was submitted to the USPTO. GDATE is the number of weeks elapsed since January 1, 1960 to the date the USPTO allowed the patent. COUNTRY is the country of citizenship for the first inventor listed on the patent application. POSTATE is the state of residency for the first inventor listed on the patent application. ASSIGNEE indicates to whom the patent is assigned and is unique to each assignee. ASSCODE indicates the type of assignee. CLAIMS is the number of independent and dependent claims listed on the patent. NCLASS indicates the broad classification for the patent.

Table 2 provides information about the source data constructed variables and explanations of their meanings. CAT is a higher-level classification of the main patent class. SUBCAT is a sub-category of the main patent class. CMADE indicates the number of citations made by the patent. CRECEIVE indicates the number of citations in other patents that reference the patent. RATIOCIT is the ratio of the number of citations made by all patents granted since 1963 to the total number of citations made by the patent. GENERAL is a measure of how broad the influence of a patent spans across fields. ORIGINAL is a measure of the originality of the patent. FWDAPLAG measures forward citations lag. BCKGTLAG measures backward citations lag. SELFCTUB is the upper bound for the share of citations the patent makes to other patents assigned to the same assignee (i.e., self-citations made). SELFCTLB is the lower bound for the share of citations the patent makes to other patents assigned to the same assignee. SECUPBD is the upper bound for the share of citations the patent receives from other patents assigned to the same assignee (i.e., self-citations received). SECDLWBD is the lower bound for the share of citations the patent receives from other patents assigned to the same assignee.

**Data Selection and Modification**

This study used the GYEAR, CLAIMS, CMADE, CRECEIVE, GENERAL, ORIGINAL, FWDAPLAG, and BCKGTLAG variables in the analysis. The APPYEAR was not used because patent applications remain unpublished for a certain period during which time they are not available to other researchers and inventors. As such the knowledge contained in them cannot be transferred. The SUBCAT variable were eliminated from the data set because including it would significantly increase the number of cases needed for certain types of regression analyses and very likely make the model more complicated than necessary. The SELFCTLB, SELFCTUB, SECDLWBD and SECUPBD variables were not used because they were unlikely to have any explanatory value. These variables are undefined when the CRECEIVE or CMADE variables have values of zero. Moreover, these variables only obtain a value after a patent citation is received or made and thus violate the temporal condition necessary for causality. All other variables were eliminated from the data because they were unnecessary for the intended analyses.

Several modifications needed to be made to the source data for the study. In the source data, values for the GENERAL variable were not calculated when the value of the CRECEIVE variable was zero. For the purposes of the study, the GENERAL variable for these cases was imputed with a value of zero. Likewise, values for the ORIGINAL variable were not calculated in the source data when the value of the CMADE variable was zero. The ORIGINAL variable was imputed with a value of 1 in these cases for the purposes of the study. The FWDAPLAG and BCKGTLAG variables in the source data were undefined for cases in which the value of the CRECEIVE and CMADE variables were zero, respectively. In these cases, the FWDAPLAG and BCKGTLAG variables were imputed with the maximum value of the variable found in the data sample.

Table 3 lists the additional variables that were created for the analysis. A variable named CRECbinary was created, which takes on a value of 0 if the value of the CRECEIVE variable is zero and 1 if the value of the CRECEIVE variable is greater than zero. Another dichotomous variable called CRECmdnSplit was created based on the CRECEIVE variable using a median split of the data. Cases were coded as 1 when the value of the CRECEIVE variable was less than or equal to the median. Cases were coded as 1 when the value of the CRECEIVE variable was greater than the median. A variable named CRECordinal was also created, which takes on the same value as the CRECEIVE variable expect that all cases in which the value of the CRECEIVE variable were equal to or greater than 15 citations were coded as 15 to limit the number of ordinal outcomes levels. Several dummy variables (i.e., indicator variables) designated CAT01 through CAT06 were created to capture cases associated each of the nominal categories of the CAT variable. Finally, a variable named CRECsqrt was created using a square root transformation of the CRECEIVE variable.

**Analysis and Results**

The R programming language was used to analyze the data for this study. Appendix B shows the full R Notebook and output. To develop a basic familiarity with the data, histograms of each variable were created to visually inspect each variable’s distribution. Quantile-Quantile (QQ) plots were also created to better understand the distribution of each variable. Scatter plots of the CRECEIVE variable against each of the other primary variables were then created to visually inspect for evidence of linear relationships. Measures of central tendency were then calculated for each variable.

Table 1 shows a correlation matrix for the variables. None of the variables chosen for the analysis appear to be strongly correlated with one another. Observation counts for each outcome level of each categorical and nominal variable were calculated. For logistic regression analysis, the rule of thumb is that there should be at least 10 observations for the least frequent outcome level of each variable (CITATION NEEDED). Observations counts confirmed that this condition was satisfied.

A binomial logistic regression analysis was performed using CRECbinary as the dependent variable and GYEAR, CAT02, CAT03, CAT04, CAT05, CAT06, CLAIMS, CMADE, ORIGINAL, GENERAL, FWDAPLAG, and BCKGTLAG as independent variables. The CAT01 variable was used as the reference category for the indicator variables. The hypothesis for this analysis was that at least one independent variable would be associated with the odds of a patent receiving any citations from other patents. In mathematical terms this can be expressed as follows:

= Logit-1() = Logit-1[)] =

where the null hypothesis was that HO: β0 = β1 …=β12 = 0 and the alternative hypothesis was that HA: β ≠ 0 for at least one independent variable. Moreover, it was expected that β9 0 indicating a negative association with the ORIGINAL variable and that β10 0 indicating a positive association with the GENERAL variable.

Table 1 shows the results of the binomial regression analysis. The log likelihood was improved by 882.6462 from -887.2297 to -4.5835. The Hosmer-Lemeshow goodness of fit test produced value of 1.149(10-9) with a p value of 1, which indicates that there was not sufficient evidence to reject the null hypothesis that the model fit the data well. However, none of the coefficients were significant. Taken in whole, these results indicate that the model did NOT fit the data well and is inefficient in predicting whether a patent receives at least one citation (i.e., whether the knowledge contained in the patent is transferred).

A second binomial logistic regression analysis was performed using CRECmdnSplt as the dependent variable and GYEAR, CAT02, CAT03, CAT04, CAT05, CAT06, CLAIMS, CMADE, ORIGINAL, GENERAL, FWDAPLAG, and BCKGTLAG as independent variables. As before, the CAT01 variable was used as the reference category for the indicator variables. Again, the hypothesis for this analysis was that at least one of the independent variables would be associated with the odds of a patent receiving more than the median number of citations from other patents as expressed by the following:

= Logit-1() = Logit-1[)] =

where the alternative hypothesis was that HA: β ≠ 0 for at least one of the independent variables. As before, it was expected that β9 0 indicating a negative association with the ORIGINAL variable and β10 0 indicating a positive association with the GENERAL variable.

Table 1 shows the results of the second binomial regression analysis. In this case, coefficients for GYEAR, CAT02, CAT03, ORIGINAL, GENERAL, and FWDAPLAG were significant at the 0.001 level. Coefficients for CMADE and BCKGTLAG were significant at the 0.01 level. Coefficients for CAT04 and CLAIMS were significant at the 0.05 level. The coefficients for CAT05 and CAT06 were not significant. The McFadden pseudo-R2 value was 0.295 and the log likelihood was improved by 403.6454 from a value of -1,367.4155 to -963.7701. The Hosmer-Lemeshow goodness of fit test produced a value of 30.313 with a p value less than 0.001, which indicated a lack of fit that is significant. Taken in whole, these results suggest that the model fits the data well and is efficient in predicting whether a patent received more than the median number of citations (i.e., whether the knowledge contained in the patent is transferred at greater than the median amount). It’s interesting to note that a patent with one unit increase in the value of the GENERAL variable was roughly 72 times more likely to receive more than the median number of citations. While this is consistent with the expected results, the magnitude of the association was unexpected. Patents classified as CAT02 and CAT03 were respectively 2.72 and 2.06 times more likely to receive more than the median number of citations than patents classified as CAT01.

An ordinal logistic regression analysis was performed using CRECordinal as the dependent variable and GYEAR, CAT02, CAT03, CAT04, CAT05, CAT06, CLAIMS, CMADE, ORIGINAL, GENERAL, FWDAPLAG, and BCKGTLAG as independent variables. The CAT01 variable was used as the reference category for the indicator variables. The hypothesis for this analysis was that at least one independent variable would be associated with the odds of a patent receiving a given level of citations from other patents as expressed by the following:

= Logit-1() = Logit-1[] =

where the alternative hypothesis was HA: β ≠ 0 for at least one of the independent variables and *j* takes on ordinal values from 1 to 15. It was specifically expected that β9 0 indicating that there would be a negative association with the ORIGINAL variable and that β10 0 indicating that there would be a positive association with the GENERAL variable.

Table 1 shows the results of the second binomial regression analysis. The coefficients for CAT02 were not significant, which indicates that the odds of a patents in this category receiving more citations was no greater than patents classified as CAT01. The coefficients for BCKGTLAG were significant at the 0.05 level while all remaining independent variables were significant at the 0.001 level. The McFadden pseudo-R2 value for the model was 0.189 and the log likelihood for the model was improved by 919.961 from a value of -4,872.063 to -3,952.102. The Hosmer-Lemeshow goodness of fit test produced a value of 1,626,200 with a p value less than 0.001, which indicated a lack of fit that is significant. Taken in whole, these results suggest that the model fits the data well and is efficient in predicting the probability associated with each level of citations by other patents received by a patent (i.e., the probability associated with each level of transfer of knowledge contained in the patent). The model indicated that patents classified as CAT01 had a higher probability of receiving a given level of citations from other patents than all other classifications, except for patents classified as CAT05. As expected, there was a negative association between the ORIGINAL variable and the level of patent citations received. The GENERAL variable had a positive association with the level of patent citations received as theorized. Like the results of the binomial logistic regression, the magnitude of the association was much larger than expected. The odds of a patent achieving a given level of citations received from other patents increased by 104 times for a one unit increase in the value of the GENERAL variable.

A multiple regression analysis was performed using CRECEIVE as the dependent variable and GYEAR, CAT02, CAT03, CAT04, CLAIMS, ORIGINAL, GENERAL, and FWDAPLAG as independent variables. The CAT01 variable was used as the reference category for the indicator variables. These variables were selected based on a review of the adjusted R2 for various regression subsets, which is shown in Figure 2. For the purposes of the multiple regression analysis, the CRECEIVE variable was treated as continuous because the number of integer values that the variable could take on was theoretically infinite. It was theorized that a positive linear relationship existed between the number of citations a patent received from other patents and the various independent variables selected for the model as represented by the following equation:

*CRECEIVE = β0 + β1(GYEAR) + β2(CAT02) + β3(CAT03) + β4(CAT04) +*

*β5(CLAIMS) + β6(ORIGINAL) + β7(GENERAL) + β8(FWDAPLAG) + ε*

where the alternative hypothesis was HA: β ≠ 0 for at least one independent variable. It was expected that β6 0 and β7 0, that is there would be a negative association with the ORIGINAL variable and a positive association with the GENERAL variable.

Table 1 shows the results of the multiple regression analysis. All independent variables were significant at the 0.001 level except for the ORIGINAL variable, which was significant at the 0.01 level. The adjusted R2 value was 0.2342 indicating that the model explained 23.42 percent of the variation in the value of the CRECEIVE variable. Once again, the magnitude of association with the GENERAL variable was considerably larger than all other variables. The ORIGINAL variable had a negative association as expected.

A check of the assumptions of linear regression suggests that the multiple regression model above can be significantly improved. As Figure 2 shows, there appears to be some heteroscedasticity in the data. Moreover, outliers appear to be influencing the results. Additionally, all the variables seemed to exhibit a high level of correlation with the model residuals. The mean of the regression residuals was zero for all intents and purposes. The variability in the values of the independent variables was positive. The variance inflation factors (VIF) for the variables were all low indicating a low level of multicollinearity.

To improve the efficiency of the model, another multiple regression was performed using CRECsqrt as the dependent variable, which is a square root transformation of the CRECEIVE variable. Additionally, cases in which the value of the CRECEIVE variable was equal to or greater than 15 were removed as outliers. Table 1 shows the results of the multiple regression analysis. All independent variables were significant at the 0.001 level except for the CAT02, CAT03, and CLAIMS variables, which were significant at the 0.01 level and the CAT04 variable, which was significant at the 0.05 level. The adjusted R2 value was 0.6401 indicating that the model explained 64.01 percent of the variation in the value of the CRECsqrt variable. The magnitude of association with the GENERAL variable remained considerably larger than all other variables. The ORIGINAL variable had a negative association as before.

A check of the assumptions of linear regression indicate that applying the transformation to the dependent variable improved the multiple regression model. As Figure 4 shows, the heteroscedasticity in the data appears to have been eliminated. Correlation between the model residuals and many of the independent variables appears to have been removed. The mean of the regression residuals was essentially zero. The variability in the values of the independent variables was positive. The VIF for all the independent variables remained low indicating a low level of multicollinearity.

**Discussion**

**Policy Implications**

The analysis provides insight into a topic that is of considerable interest to policymakers. It provides information to help both industry professionals and policymakers better understand the drivers of the technology transfer outcomes and identify possible factors that should be considered when forming public policy regarding technology transfer. The analysis suggests that considering non-transactional measures of knowledge transfer may be feasible. This could significantly affect the objectives of policymakers with regard to technology transfer. As such, this study may influence how policymakers think about technology transfer and how they formulate public policy to increase the transfer of federally-funded research to the private sector.

**Limitations and Future Analysis**

As with any research project or study, this analysis has limitations. Since this analysis was focused on patent data for a five year period from 1995 to 1999, findings based on the data may not be relevant to time frames before or after this period. Additionally, there is a truncation effect in the data. Patents issued in the earliest part of the study period have the potential of receiving citations from patents over a longer period than patents issued in the latter part of the study period.

There are several opportunities to improve upon and extend the analysis presented in this paper. To begin, it might prove useful to secure more recent data and to examine a subset of data buffered by at least 5 years of data on both sides of the period of study to minimize truncation effects. Removing outliers from the data may improve the goodness-of-fit of the model. Measuring only the number of dependent claims in a patent rather than all claims may help to better isolate the association between the number of claims and the number of citations received. Introducing classifications as an indication of the diversity of a technology as well as the category and subcategory of patents into the analysis to determine if the type of technology is associated with technology transfer outcomes might also be useful.

**Conclusion**

In this study, I have continued to explore an alternative conceptualization of technology transfer and an approach to measuring technology transfer based on patent citations received, which represents a non-transactional based modality of technology transfer. Using patent data, I conducted a binary logistic regression analysis to estimate the probability that a patent will receive more than 2 citations from other patents based on the year the patent was granted, the number of claims contained in the patent, and measures of the patent’s originality and generality. The resulting model indicated that the generality of a patent had the strongest association with whether or not the patent received more than 2 citations from other patents. The study results were also consistent previous analyses that indicated an inverse relationship between the year a patent was granted and the originality of a patent with the probability that the patent received more than 2 citations from other patents.

References

Appio, F. P., Martini, A., & Fantoni, G. (2017). The light and shade of knowledge recombination: Insights from a general-purpose technology. *Technological Forecasting & Social Change*, *125*, 154–165. https://doi.org/10.1016/j.techfore.2017.07.018

Bush, V. (1945). *Science, the endless frontier*. A report to the President: Washington, U.S. Government printing office, 1945.

Congressional Budget Office [CBO]. (2018). Historical Budget Data [Data file]. *The Budget and Economic Outlook: 2018 to 2028*. Retrieved from https://www.cbo.gov/about/products/budget-economic-data#2

*Daily Compilation of Presidential Documents*. (2011, October 28). Retrieved from https://www.govinfo.gov/app/collection/CPD/

Hall, B. H., Jaffe, A. B. and Trajtenberg, M. (2001). "The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools." *NBER Working Paper 8498*. Retrieved from http://www.nber.org/patents/

Kirkman, D. M. (2013). Selecting University Technology Transfer Modes: An Examination of Biotechnology Firms’ Entrepreneurial Orientation. *Journal of Technology Management & Innovation, Vol 8, Iss 2, Pp 189-208 (2013)*, (2), 189. https://doi.org/10.4067/S0718-27242013000200016

Licht, M. H. (1995). Multiple regression and correlation. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (pp. 19-64). Washington, D.C.: American Psychological Association.

Machlup, F. (1962). *The production and distribution of knowledge in the United States*. Princeton, NJ: Princeton University Press.

National Bureau of Economic Research. (2018). Patent data, including constructed variables [data file]. Retrieved from http://www.nber.org/patents/

United Nations. (2017). *GDP and its breakdown at current prices in U.S. dollars* [Data file]. Retrieved from https://unstats.un.org/unsd/snaama/dnllist.asp

Office of Management and Budget [OMB]. (2002). *The President's Management Agenda*. Retrieved from http://www.dtic.mil/dtic/tr/fulltext/u2/a394421.pdf

Office of Management and Budget [OMB]. (2018). *The President's Management Agenda*. Retrieved from https://www.whitehouse.gov/wp-content/uploads/2018/03/Presidents-Management-Agenda.pdf

Yoshikane, F. (2013). Multiple regression analysis of a patent’s citation frequency and quantitative characteristics: the case of Japanese patents. *Scientometrics*, *96*(1), 365–379. https://doi.org/10.1007/s11192-013-0953-4

Appendix A. Figures and Tables

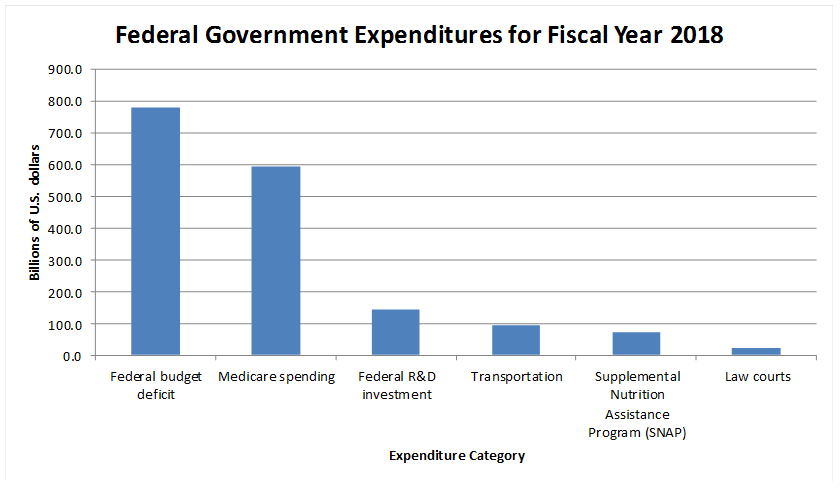


Figure 1. Federal Government Expenditures for Fiscal Year 2018

{INSERT FIGURE}

Figure 2. Regression Subsets

{INSERT FIGURE}

Figure 3. Multiple Regression Model Residuals Plot

{INSERT FIGURE}

Figure 4. Residuals Plot for Multiple Regression Model Using CRECsqrt

Table 1

Original USPTO Variables of Source Data

| Variable | Variable Type | Extended Name | Description |
| --- | --- | --- | --- |
| PATENT | Numeric  Nominal | Patent Number | The number assigned to the allowed patent by the USPTO.  Takes on values integer values between 3070801 and 6009554. |
| GYEAR | Numeric  Interval | Grant Year | The year the USPTO allowed the patent.  Takes on integer values between 1963 – 1999. |
| GDATE | Numeric  Interval | Grant Date | The date the USPTO allowed the patent expressed in terms of the number of weeks elapsed since  January 1, 1960.  Takes on integer values between 156 and 2,028. |
| APPYEAR | Numeric  Interval | Application Year | The year the patent application was submitted to the USPTO.  Takes on integer values between 1963 – 1999. |
| COUNTRY | Character  Nominal | Country of First Inventor | The country of citizenship for the first inventor listed on the patent application.  Takes on values of two character string data. |
| POSTATE | Character  Nominal | State of First Inventor (US) | The state of residency for the first inventor listed on the patent application if the country of citizenship is the United States of America.  Takes on values of two character string data. |
| ASSIGNEE | Numeric  Nominal | Assignee Identifier | Unique identifier for the assignee of the patent.  Takes on values from 10950 to 99550. |
| ASSCODE | Numeric  Nominal | Assignee Code | A one character code categorizing the type of assignee.  Takes on values from 1 to 7. |
| CLAIMS | Numeric  Interval | Number of Claims | Number of independent and dependent claims on the patent.  Takes on integer values from 1 to . |
| NCLASS | Numeric  Nominal | Main Patent Class | A code that categorizes the patent into one of several broad classifications.  Takes on integer values from 1 to 800. |

Table 2

Source Data Constructed Variables

| Variable | Variable Type | Extended Name | Description |
| --- | --- | --- | --- |
| CAT | Numeric  Nominal | Technological Category | A higher-level classification of the Main Patent Class.  Takes on integer values from 1 to 6. |
| SUBCAT | Numeric  Nominal | Technological Sub-category | The sub-category of the primary technological category to which the patent is assigned.  Takes on integer values from 1 to 69. |
| CMADE | Numeric  Interval | Number of Citations Made | The number of citations made by the patent.  Takes on integer values from 1 to . |
| CRECEIVE | Numeric  Interval | No. of Citations Received | The number of citations in other patents that reference the patent.  Takes on integer values from 1 to . |
| RATIOCIT | Numeric  Ratio | Percent of Citations Made to Patents Granted Since 1963 | The ratio of the number of citations made by all patents granted since 1963 to the total number of citations made by the particular patent.  Takes on continuous values between 0 and 1. |
| GENERAL | Numeric  Ratio | Measure of Generality | A measure of how broad the influence of a patent spans across fields as determined by the number of different fields of all patents that cite the patent of interest.  Calculated as the following:  Generalityi = 1 - , where *sij* denotes the percentage of citations received by patent *i* that belong to patent class *j*, out of *ni* patent classes.  Takes on continuous values between 0 and 1. |
| ORIGINAL | Numeric  Ratio | Measure of Originality | A measure of the originality of a patent as determined by the number of different fields for all patents cited by the patent of interest.  Calculated as the following:  Originalityi = 1 - , where *sij* denotes the percentage of citations made by patent *i* that belong to patent class *j*, out of *ni* patent classes.  Takes on continuous values between 0 and 1. |
| FWDAPLAG | Numeric  Ratio | Mean Forward Citation Lag | The mean time difference between the application or grant date of the patent and that of the other patents citing this patent.  Takes on continuous values between 0 and 1. |
| BCKGTLAG | Numeric  Ratio | Mean Backward Citation Lag | The mean time difference between the application or grant date of the patent and that of the patents it cites.  Takes on continuous values between 0 and 1. |
| SELFCTUB | Numeric  Ratio | Share of Self-Citations Made – Upper Bound | The number of citations made by the patent to other patents with the same assignee divided by the total number of citations made by all patents with assignee codes.  Takes on continuous values between 0 and 1. |
| SELFCTLB | Numeric  Ratio | Share of Self-Citations Made – Lower Bound | The number of citations made by the patent to other patents with the same assignee divided by the total number of citations made by all patents.  Takes on continuous values between 0 and 1. |
| SECUPBD | Numeric  Ratio | Share of Self-Citations Received – Upper Bound | The number of citations received by the patent from other patents with the same assignee divided by the total number of citations received by all patents with assignee codes.  Takes on continuous values between 0 and 1. |
| SECDLWBD | Numeric  Ratio | Share of Self-Citations Received – Lower Bound | The number of citations received by the patent from other patents with the same assignee divided by the total number of citations received by all patents.  Takes on continuous values between 0 and 1. |

Table 3

Variables Used in Analysis

| Variable | Variable Type | Extended Name | Description |
| --- | --- | --- | --- |
| GYEAR | Numeric  Interval | Grant Year | The year the USPTO allowed the patent.  Takes on integer values between 1963 – 1999. |
| CLAIMS | Numeric  Interval | Number of Claims | Number of independent and dependent claims on the patent.  Takes on integer values from 1 to . |
| CRECBINARY | Numeric  Nominal | No. of Citations Received | 1 indicates 0-2 citations  2 indicates 3 or more  Takes on an integer value of 1 or 2. |
| GENERAL | Numeric  Ratio | Measure of Generality | A measure of how broad the influence of a patent spans across fields as determined by the number of different fields of all patents that cite the patent of interest.  Takes on continuous values between 0 and 1. |
| ORIGINAL | Numeric  Ratio | Measure of Originality | A measure of the originality of a patent as determined by the number of different fields for all patents cited by the patent of interest.  Takes on continuous values between 0 and 1. |

Table 4

Correlation Matrix

Table 5

Binomial Logistic Regression for CRECbinary

Table 6

Binomial Logistic Regression for CRECmdnSplt

Table 7

Ordinal Logistic Regression

Table 8

Multiple Regression

Table 9

Autocorrelation Analysis